Modélisation, simulation numérique et problèmes inverses. Contributions en physique des plasmas de Tokamak, en écologie marine et autres travaux

Blaise Faugeras

UNIVERSITÉ DE NICE SOPHIA ANTIPOLIS

Mémoire présenté pour l'obtention d'une

HABILITATION À DIRIGER DES RECHERCHES

Spécialité : Mathématiques Appliquées

soutenue le 12 Octobre 2015

par

Blaise Faugeras

Modélisation, simulation numérique et problèmes inverses. Contributions en physique des plasmas de tokamak, en écologie marine et autres travaux

---

Devant le jury composé de :

Mark Asch	Université de Picardie	*Rapporteur*
Pierre Auger	IRD Bondy	*Rapporteur*
Jacques Blum	Université Nice Sophia Antipolis	*Examinateur*
Hervé Guillard	INRIA Sophia Antipolis	*Examinateur*
Enzo Lazzaro	IFP CNR Milan	*Rapporteur*
François Saint-Laurent	IRFM CEA Cadarache	*Examinateur*

Laboratoire J.A. Dieudonné, UMR CNRS 7351

Remerciements

Ce document synthétise mon travail de ces dernières années. Ce travail est également celui des co-auteurs des articles présentés ici et je tiens à les saluer amicalement. Au cours de la rédaction j'ai souvent repensé à tous ces bons moments partagés avec beaucoup d'entre eux à écrire des petites équations, des petits modèles, des petits codes de calcul et finalement ces petits articles scientifiques.

J'aimerai sincèrement remercier les membres du jury d'avoir accepté cette charge. J'ai été très sensibles aux rapports sur mon travail de spécialistes tels que Mark Asch, Pierre Auger et Enzo Lazzaro chacun dans leur domaine. Hervé Guillard et François Saint-Laurent ont accepté sans détour les rôles d'examinateur et de président du jury. J'en suis très heureux. Enfin *last but not least*, un grand merci à Jacques Blum à qui je dois beaucoup professionnellement et depuis longtemps maintenant.

Pour finir je voudrais saluer les collègues du Laboratoire J.A. Dieudonné, de l'équipe CASTOR de l'INRIA et de l'IRFM au CEA Cadarache et remercier l'équipe administrative pour son aide dans la préparation de la soutenance.

Avant-propos

Dans ce document je décris mon activité scientifique depuis la fin de ma thèse soutenue en octobre 2002. Il est constitué de deux parties. La première est une synthèse de mon travail et la seconde un recueil de mes articles les plus représentatifs en rapport avec les travaux décrits en première partie [1].

La partie synthèse est constituée de deux chapitres. Le premier regroupe mes travaux les plus récents, depuis fin 2007, sur la thématique de la simulation numérique et des problèmes inverses pour les plasmas de Tokamak. Le second regroupe des travaux plus anciens concernant la modélisation et l'assimilation variationnelle de données en écologie marine ainsi que deux autres travaux isolés. Les liens avec les articles contenus dans la seconde partie du document sont faits au fil du texte. Les thématiques abordées dans ces deux chapitres étant relativement variées ils ne contiennent pas de bibliographie exhaustive ni trop de détails techniques. Pour une information plus détaillée je réfère le lecteur aux articles collectés dans la deuxième partie. Enfin cette partie synthèse se termine par une liste de mes publications numérotées de [1] à [35], suivie de la bibliographie générale du document commençant donc à la référence [36].

1. Dans la version courte du document ce recueil d'articles est simplement constitué des liens de téléchargement de ces articles. La version longue inclut les articles eux-mêmes

Table des matières

Curriculum vitae **1**

Synthèse des travaux de recherche **7**

1 Simulation numérique et problèmes inverses pour les plasmas de Tokamak **7**
1.1 Modélisation de l'équilibre quasi-statique du plasma 7
1.2 Reconstruction de l'équilibre à partir de mesures expérimentales 10
1.2.1 Identification de la densité de courant 11
1.2.2 Reconstruction de la frontière du plasma seule 16
1.3 Calculs d'équilibres et couplage avec la diffusion résistive dans le plasma 22
1.4 Conclusion 28

2 Modélisation, assimilation de données en écologie marine et autres travaux isolés **29**
2.1 Généralités sur l'identification de paramètres par assimilation variationnelle de données 29
2.2 Modélisation 32
2.2.1 Modélisation pour les pêcheries de thons tropicaux 33
2.2.2 Modélisation du flux d'énergie dans les écosystèmes marins 35
2.3 Identification de paramètres par assimilation variationnelle de données de pêche . . 38
2.4 Travaux isolés 44
2.4.1 Un schéma numérique précis pour l'intégration en temps d'un système d'équations de diffusion - dissolution / précipitation 44
2.4.2 Analyse asymptotique d'un modèle élastique 3D pour la segmentation d'images du coeur 45
2.5 Conclusion 47

Liste de publications **49**

Bibliographie **52**

Recueil d'articles **59**

Article A : [4] J. Blum, C. Boulbe et B. Faugeras. Reconstruction of the equilibrium of the plasma in a Tokamak and identification of the current density profile in real time. *J. Comp. Phys.* 231 (2012), p. 960–980 59

Article B : [28] D. Mazon, P. Lotte, B. Faugeras, C. Boulbe, J. Blum, F. Saint-Laurent, S. Bremond, P. Moreau, A. Murari et P. Blanchard. "Validation of the new real-time equilibrium code EQUINOX on JET and ToreSupra". In : *Proceedings of the 39th EPS Conference and 16th Int. Congress on Plasma Physics.* Stockholm, Sweden, juil. 2012 81
Article C : [7] B. Faugeras, J. Blum, C. Boulbe, P. Moreau et E. Nardon. 2D interpolation and extrapolation of discrete magnetic measurements with toroidal harmonics for equilibrium reconstruction in a Tokamak. *Plasma Phys. Control Fusion* 56 (2014), p. 114010 87
Article D : [6] B. Faugeras, A. Ben Abda, J. Blum et C. Boulbe. Minimization of an energy error functional to solve a Cauchy problem arising in plasma physics : the reconstruction of the magnetic flux in the vacuum surrounding the plasma in a Tokamak. *ARIMA* 15 (2012), p. 37–60 99
Article E : [9] H. Heumann, J. Blum, C. Boulbe, B. Faugeras, G. Selig, J.-M. Ané, S. Brémond, V. Grangirard, P. Hertout et E. Nardon. Quasi-static free-boundary equilibrium of toroidal plasma with CEDRES++ : computational methods and applications. *J. Plasma. Phys.* (2015). doi : 10.1017/S0022377814001251 . . 125
Article F : [5] G. L. Falchetto, D. Coster, R. Coelho, B.D. Scott, L. Figini, D. Kalupin, E. Nardon, L.L. Alves, J.F. Artaud, V. Basiuk, J. Bizarro, C. Boulbe, A. Dinklage, D. Farina, B. Faugeras, J. Ferreira, A. Figueiredo, P. Huynh, F. Imbeaux, I. Ivanova-Stanik, T. Jonsson, H.-J. Klingshirn, C. Konz, A. Kus, N.B. Marushchenko, E. Nardon, S. Nowak, G. Pereverzev, M. Owsiak, E. Poli, Y. Peysson, R. Reimer, J. Signoret, O. Sauter, R. Stankiewicz, P. Strand, I. Voitsekhovitch, E. Westerhof, T. Zok, W. Zwingmann, ITM-TF contributors, the ASDEX Upgrade Team et JET-EFDA Contributors. The European Integrated Tokamak Modelling (ITM) effort : achievements and first physics results. *Nucl. Fusion* 54.4 (2014), p. 043018. doi : 10.1088/0029-5515/54/4/043018 161
Article G : [15] B. Faugeras et O. Maury. An advection-diffusion-reaction size-structured fish population dynamics model combined with a statistical parameter estimation procedure : Application to the Indian Ocean skipjack tuna fishery. *Math. Biosciences and Engineering* 2.4 (2005), p. 719–741 183
Article H : [16] B. Faugeras et O. Maury. Modelling fish population movements : from an individual-based representation to an advection-diffusion equation. *J. Theor. Biol.* 247 (2007), p. 837–848 207
Article I : [11] S. Dueri, B. Faugeras et O. Maury. Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E : Part 1. Model formulation. *Ecological Modelling* 245 (2012), p. 41–54 221
Article J : [12] S. Dueri, B. Faugeras et O. Maury. Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E : Part 2. Parameter estimation and sensitivity analysis. *Ecological Modelling* 245 (2012), p. 55 –64 243
Article K : [18] O. Maury, B. Faugeras, Y-J. Shin, J.C. Poggiale, T. Ben Ari et F. Marsac. Modeling environmental effects on the size-structured energy flow through marine ecosystems. Part 1 : the model. *Progress in Oceanography* 74 (2007), p. 479–499255
Article L : [21] B. Faugeras, J. Pousin et F. Fontvieille. An efficient numerical scheme for precise time integration of a diffusion - dissolution / precipitation chemical system. *Math. of Computation* 75.253 (2006), p. 209–222 277
Article M : [20] B. Faugeras et J. Pousin. Variational asymptotic derivation of an elastic model arising from the problem of 3D automatic segmentation of cardiac images. *Analysis and Applications (AA)* 2.4 (2004), p. 275–307 293

Curriculum vitae

Blaise Faugeras

Etat civil

Né le 17 Avril 1975 à Salt Lake City (USA), nationalité française, pacsé, deux enfants

Adresse

Laboratoire J.-A. Dieudonné (UMR 7351),
Université de Nice Sophia-Antipolis,
Faculté des Sciences, Parc Valrose,
06108 Nice Cedex 02
tel : 06 50 37 41 54
e-mail : Blaise.Faugeras@unice.fr
url : `http://math.unice.fr/~faugeras/`

Parcours professionnel

Depuis 09/2007 Ingénieur de Recherche CNRS (IR2 puis IR1) au Laboratoire Dieudonné, Nice
Membre de l'équipe EDP et Analyse numérique
Membre de l'équipe-projet INRIA CASTOR depuis sa création en 2012

10/2006 - 08/2007 Chargé de Recherche IRD (CR2) au Centre de Recherche Halieutique (CRH), Sète. Membre de l'UR 109 Thetis

03/2005 - 09/2006 Ingénieur de Recherche CNRS (IR2) au Laboratoire I3S, Sophia Antipolis

11/2003 - 02/2005 Postdoctorant IRD au CRH, Sète

09/2003 - 10/2003 Postdoctorant INRIA Sophia Antipolis dans le projet COMORE

09/2002 - 08/2003 ATER à l'INSA de Lyon

09/1999 - 08/2002 Doctorant à l'IMAG, Grenoble. Moniteur à l'Université de Savoie, Chambéry

Formation

2002 Doctorat de Mathématiques appliquées, Université Joseph Fourier, Grenoble.
"Assimilation variationnelle de données dans un modèle couplé océan - biogéochimie" sous la direction de J. Blum et la codirection de J. Verron

1999 DEA "Analyse numérique, équations aux dérivées partielles et calcul scientifique", Université Claude Bernard, Lyon.

1999 Ingénieur civil de l'Ecole Nationale Supérieure des Mines de St Etienne

Activité scientifique

Mes principaux sujets de recherche s'articulent autour de la modélisation, de la simulation numérique et de problèmes inverses d'identification pour des systèmes physiques et biologiques régis par des équations aux dérivées partielles.

Ces travaux m'ont amené à participer au développement de plusieurs codes de calcul et ont donné lieu à plus d'une trentaine d'articles et actes de conférences sur les thèmes suivants entre autres :

- A l'Université de Nice en collaboration avec l'Institut de Recherche sur la Fusion Magnétique au CEA Cadarache (IRFM Tore Supra - WEST) et avec le Joint European Torus (JET) en Angleterre :
  - identification de la frontière libre du plasma et de la densité de courant à partir de données expérimentales dans un Tokamak. Développement du code de calcul VacTH qui permet de reconstruire en temps réel la frontière plasma et est utilisé au CEA [7]. Principal développeur du code de calcul EQUINOX qui résout en temps réel le problème d'identification de la densité de courant [4] et est implémenté au JET et à ToreSupra.
  - Simulation de l'évolution de l'équilibre du plasma à l'échelle de temps de la diffusion résistive. Participation au développement du code d'équilibre CEDRES++ [9].
- A l'IRD : modélisation de la dynamique de populations structurées en taille et en espace. Identification de paramètres par assimilation de données de pêche du thon Listao dans l'Océan Indien. Développement du code de calcul APECOSM-E [11, 12].
- Pendant ma thèse : problème inverse d'identification de paramètres dans un modèle couplé océan-biogéochimie par assimilation de données type "couleur de l'océan".

Enseignement et encadrement

Cours et TD

2012 - 2013 40h par an de TD d'optimisation en MASTER IMEA et Mathmods à l'Université de Nice Sophia-Antipolis

2010 - 2014 40h par an de TD de mathématiques en 2ième année à l'Ecole Polytechnique Universitaire de Nice Sophia-Antipolis

2006 - 2012 Animation d'un atelier "Assimilation de données" en MASTER 2 MASS puis IMEA à l'Université de Nice Sophia-Antipolis 30h eq. TD par an

2005 3h de cours "Introduction à l'assimilation variationnelle de données" pour 20 étudiants du MASTER Océanographie et Environnements Marins de l'Université Pierre et Marie Curie

2002 - 2003 ATER en mathématiques à l'INSA de Lyon 96h de TD d'analyse et algèbre en premier cycle 2ième année

1999 - 2001 Moniteur en mathématiques à l'Université de Savoie, Chambéry 64h par an de TD d'analyse et algèbre en DEUG MIAS

1997 - 1999 Interrogations orales hebdomadaires de mathématiques en classe de Mathsup MPSI au Lycée Claude Fauriel, St-Etienne

Encadrement

PFE 2014 - 2015 Co-encadrement avec Sid Touati et Cedric Boulbe du projet de fin d'étude (cycle ingénieur SI5 de l'Ecole Polytechnique Universitaire de Nice Sophia-Antipolis) de 4 étudiants, Pierre Bouillet, Damien Viano, Hennani Jamal et William Tassoux sur l'optimisation du temps de calcul d'un code de reconstruction de la frontière du plasma dans un Tokamak.

PFE 2008 - 2009 Co-encadrement avec Jacques Blum et Cedric Boulbe du projet de fin d'étude (cycle ingénieur MAM5 de l'Ecole Polytechnique Universitaire de Nice Sophia-Antipolis) d'Antoine Jarrier sur la reconstruction d'équilibre dans un Tokamak

Travail avec des doctorants et post-doctorants

2011-2012 Participation au travail de postdoc de Holger Heumann à Nice. Responsable Jacques Blum.

2009-2012 Participation au travail de thèse de Gael Sélig à Nice en particulier sur la partie couplage équilibre-transport pour un plasma de Tokamak. Directeur de thèse Jacques Blum.

2007-2009 Participation au travail de postdoc de Cédric Boulbe à Nice. Responsable Jacques Blum.

2007-2008 Participation au travail de postdoc de Sibylle Dueri à l'IRD. Responsable Olivier Maury.

Tâches d'animation scientifique

- Organisateur du minisymposium "Simulation, Identification and Control in Tokamak Plasma Physics" au congrès ICIAM 2011 à Vancouver, Canada
- Co-organisateur avec Cédric Boulbe d'un "Code Camp" d'une semaine à Nice en 2011 dans le cadre du projet européen EFDA-ITM (Integrated Tokamak Modeling)
- Membre du comité scientifique du colloque "Problèmes Inverses : des Plasmas à l'Océanographie. PIPO'2011" en l'honneur de Jacques Blum pour ses 60 ans.
- Relecture d'articles (environ 2 par an)
  - de modélisation mathématique en écologie marine dans "Journal of Mathematical Biology", "Applied Mathematical Modelling", ...
  - de simulation numérique en physique des plasmas de Tokamak dans "Journal of Computer Physics", "Fusion Engineering and Design", ...

Synthèse des travaux de recherche

Chapitre 1

Simulation numérique et problèmes inverses en physique des plasmas de Tokamak

Ce chapitre synthétise mon travail depuis mon arrivée en septembre 2007 au Laboratoire J.A. Dieudonné comme ingénieur de recherche CNRS. Cette activité est centrée sur le calcul scientifique pour la fusion par confinement magnétique. Mes collaborateurs niçois proches sont J. Blum, C. Boulbe et H. Heumann. A travers le LRC (Laboratoire de Recherche Conventionné entre le CEA, l'Université de Nice Sophia Antipolis et le CNRS) dirigé par J. Blum et la thématique modélisation et contrôle des plasmas de fusion dans le cadre du projet ITER (International Thermonuclear Experimental Reactor) nous sommes membres de la Fédération de Recherche sur la Fusion par Confinement Magnétique - ITER (FR-FCM) et participons à des projets européens du programme EUROFUSION et en particulier dans le Work Package Code Development (WPCD). Nous sommes aussi membres de l'équipe INRIA-LJAD CASTOR créée en 2012 et également dirigée par J. Blum.

Le chapitre est organisé de la manière suivante. Dans la section 1.1 on rappelle succintement la problématique de la modélisation de l'équilibre quasi-statique du plasma dans un Tokamak. La section 1.2 traite du problème inverse de la reconstruction de l'équilibre à partir de mesures expérimentales. Enfin la section 1.3 aborde le problème du couplage entre équilibre et diffusion résistive dans le plasma.

1.1 Modélisation de l'équilibre quasi-statique du plasma

Au cours d'une expérience de fusion dans un tokamak un champ magnétique est utilisé pour confiner le plasma dans une chambre à vide toroidale. Ce champ magnétique est généré par des bobines externes entourant la chambre à vide et par un courant circulant dans le plasma lui même. Les équations utilisées pour la modélisation de l'évolution de l'équilibre quasi-statique du plasma dans un tokamak sont celles de la magnétohydrodynamique (MHD). Cette modélisation est rappelée ci-dessous et les références pour cette section sont nombreuses. On pourra noter en particuler parmis les articles de références [80, 57, 54] et parmis les ouvrages plus récents [43, 85, 63].

On note $\mathbf{j}$ la densité de courant, $\mathbf{B}$ le champ magnétique et p la pression. L'équation du mouvement dans le plasma s'écrit

$$\rho(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\nabla\mathbf{u}) + \nabla p = \mathbf{j} \times \mathbf{B} \tag{1.1}$$

où $\mathbf{u}$ est la vitesse du fluide et ρ la densité de masse. Pour un plasma de fusion magnétique

dans un tokamak le nombre de Lundquist $S = \frac{\tau_R}{\tau_A}$, où τ_R est l'échelle de temps caractéristique de la diffusion résistive du courant et de la chaleur dans le plasma et τ_A est l'échelle de temps d'Alfven caractérisant les instabilités de déplacement du plasma, est de l'ordre de 10^6 à 10^{12}. Après adimentionalisation de (1.1) le terme d'inertie apparait comme un terme d'ordre S^{-2} et les autres comme des termes d'ordre S^0. A l'échelle de temps de la diffusion ce terme d'inertie est donc négligeable et l'équation du mouvement devient la relation d'équilibre

$$\nabla p = \mathbf{j} \times \mathbf{B} \tag{1.2}$$

A chaque instant les forces de Laplace et de pression se compensent et le plasma est à l'équilibre.

On ajoute les équations de Maxwell (théorème d'Ampère et conservation du champ magnétique) valables dans tout l'espace

$$\mathbf{j} = \nabla \times (\frac{\mathbf{B}}{\mu}), \quad \text{et } \nabla \cdot \mathbf{B} = 0 \tag{1.3}$$

où μ représente la perméabilité magnétique.

Sous l'hypothèse d'axisymétrie, dans un système de coordonnées cylindriques $(\mathbf{e}_r, \mathbf{e}_\phi, \mathbf{e}_z)$, on introduit le flux poloidal

$$\psi(r,z) = \frac{1}{2\pi} \int_D \mathbf{B}\mathbf{ds} = \int_0^r B_z r dr$$

où D est le disque dont la circonférence est donnée par le cercle centré sur l'axe Oz et passant par le point (r,z). On introduit également la fonction diamagnétique définie par $f = rB_\phi$. Les équations (1.3) permettent de décomposer le champ magnétique et la densité de courant en une composante poloidale dans le plan (r,z) et une composante toroidale

$$\begin{cases} \mathbf{B} = \mathbf{B}_p + \mathbf{B}_\phi \\ \mathbf{B}_p = \frac{1}{r}[\nabla\psi \times \mathbf{e}_\phi] \\ \mathbf{B}_\phi = \frac{f}{r}\mathbf{e}_\phi \end{cases} \quad \text{et} \quad \begin{cases} \mathbf{j} = \mathbf{j}_p + \mathbf{j}_\phi \\ \mathbf{j}_p = \frac{1}{r}[\nabla(\frac{f}{\mu}) \times \mathbf{e}_\phi] \\ \mathbf{j}_\phi = -\Delta^*_\mu \psi \mathbf{e}_\phi \end{cases} \tag{1.4}$$

où

$$\Delta^*_\mu . = \frac{\partial}{\partial r}(\frac{1}{\mu r}\frac{\partial .}{\partial r}) + \frac{\partial}{\partial z}(\frac{1}{\mu r}\frac{\partial .}{\partial z}).$$

La perméabilité magnétique est partout celle, constante, du vide μ_0 et l'opérateur $\Delta^*_{\mu_0}$ est linéaire, excepté dans les structures ferromagnétiques présentes dans certains tokamaks (comme le JET, Joint European Torus à Culham en Angleterre, ou ToreSupra et sa future extension WEST au CEA Cadarache) où elle est fonction du champ magnétique, $\mu = \mu(\mathbf{B}_p^2)$, et l'opérateur Δ^*_μ devient non-linéaire.

Dans le plasma l'équation d'équilibre (1.2) montre que les lignes de champ magnétique et de courant sont portées par les surfaces isobares. On les appelle surfaces magnétiques. Elles forment une famille de tores emboités qui dégénère au centre du plasma en une courbe que l'on appelle axe magnétique. D'autre part la décomposition (1.4) permet de voir que ces surfaces sont également des iso-ψ et des iso-f. On peut donc considérer p et f comme des fonctions de ψ. Enfin (1.2) conduit à l'équation de Grad-Shafranov

$$-\Delta^*_{\mu_0}\psi = r\frac{\partial p}{\partial \psi}(\psi) + \frac{1}{2\mu_0 r}\frac{\partial f^2}{\partial \psi}(\psi) \tag{1.5}$$

Les équations (1.2) et (1.3) se réduisent ainsi à une équation posée en 2 dimension d'espace dans le plan poloidal $(r,z) \in \Omega_\infty = (0,\infty) \times (-\infty,\infty)$ pour le flux poloidal ψ :

$$-\Delta^*_\mu \psi = j_\phi \tag{1.6}$$

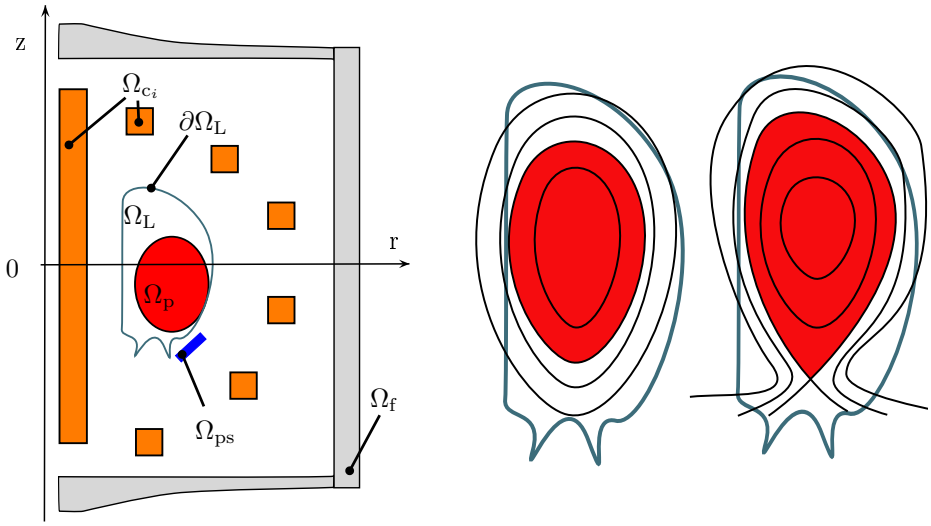

FIGURE 1.1 – Gauche : représentation schématique du plan poloidal d'un tokamak. Ω_p est le domaine du plasma, Ω_L est le domaine du limiteur accessible au plasma, Ω_{c_i} représentent les bobines de champ poloidal, Ω_{ps} les structures passives et Ω_f les structures ferromagnétiques. Droite : exemple de plasma dont la frontière est définie par le contact avec le limiteur ou par la présence d'un point X.

La composante toroidale de la densité de courant j_ϕ est nulle partout en dehors du domaine du plasma et des bobines. Les différents sous-domaines du plan poloidal d'un tokamak (voir la figure 1.1 gauche) sont rappelés ci-dessous :

- Ω_L est le domaine accessible au plasma. Sa frontière est le limiteur $\partial\Omega_L$.
- Ω_p est le domaine du plasma. C'est une inconnue, $\Omega_p = \Omega_p(\psi)$, le problème est à frontière libre. Ce domaine est défini par sa frontière qui est le plus grand iso-contour de ψ fermé et entièrement contenu dans le limiteur Ω_L. Le plasma peut être soit limité si cet iso-contour est tangent au limiteur $\partial\Omega_L$ soit défini par la présence d'un point selle appelé point-X (voir la figure 1.1 droite)
  Dans le domaine plasma Ω_p, ψ satisfait l'équation de Grad-Shafranov (1.5).
- Ω_f représente les structures ferromagnétiques. Elles ne portent pas de courant, $j_\phi = 0$ mais la perméabilité magnétique μ n'y est pas constante
- Les domaines Ω_{c_i} représentent les bobines de champ poloidal. Elles sont parcourues par des courants I_i. Il y a deux possibilités de modélisation :
  — On peut considérer que ces courants sont des données du problème (ils sont par exemple mesurés) auquel cas
  $$j_\phi = I_i/S_i$$
  où S_i est la section de la bobine. Le problème est alors statique au sens où aucune dérivée en temps n'apparait
  — On peut considérer que la donnée est la tension V_i appliquée aux bornes des circuits, supposés ici indépendants, de chaque bobine de champ poloidal auquel cas en utilisant en plus les lois de Faraday et d'Ohm on obtient
  $$j_\phi = \frac{n_i V_i}{R_i S_i} - \frac{2\pi n_i^2}{R_i S_i^2} \int_{\Omega_{C_i}} \dot{\psi} ds$$
  où n_i est le nombre de tours de conducteur dans la bobine, R_i la résistance et $\dot{\psi}$ est la dérivée temporelle de ψ au point (r, z). Le problème est alors dynamique.

- Ω_{ps} représente les structures passives. De même ici dans le cas d'une modélisation statique on a

$$j_\phi = 0$$

alors que dans le cas dynamique on prend en compte les courants induits et

$$j_\phi = -\frac{\sigma}{r}\dot{\psi}$$

où σ est la conductivité.

Enfin pour terminer de poser le problème, les conditions limites naturelles sont

$$\psi(0,z) = 0 \quad \text{et} \quad \lim_{\|(r,z)\|\to+\infty} \psi(r,z) = 0$$

Avec cette modélisation les fonctions $\frac{\partial p}{\partial \psi}(\psi)$ et $\frac{\partial f^2}{\partial \psi}(\psi)$ intervenant dans le second membre de l'équation de Grad-Shafranov ne sont ni connues, ni des inconnues du modèle d'équilibre quasi-statique. Le modèle est incomplet. Au cours d'une décharge dans un Tokamak l'évolution du plasma se fait par une succession d'équilibres reliés entre eux par la dynamique électromagnétique externe des bobines qui est modélisée ici dans le cas du problème dynamique, mais également par celle interne non modélisée ici. Cette dynamique électromagnétique interne est décrite par les équations dites de transport dans le plasma (diffusion résistive du flux magnétique, transport des particules et de la chaleur) qui permettent de calculer l'évolution de la densité de courant. La simulation d'une décharge complète nécessite de coupler de manière consistante le modèle de l'évolution de l'équilibre quasi-statique du plasma avec le modèle de transport. Réussir une telle simulation est un des challenges actuels que nous abordons succintement à la section 1.3.

Néanmoins la modélisation de l'équilibre est déjà une fin en soi. Elle permet de traiter la question de la reconstruction de la frontière plasma et de la densité de courant au cours de la décharge à partir de mesures expérimentales. La section 1.2 synthétise les travaux sur ce sujet.

La modélisation de l'équilibre est également essentielle pour les études de scénario, de dimensionnement ou pour tester les algorithmes de contrôle du plasma. Dans la section 1.3 sont rapidement présentées les méthodes de simulation numérique directe d'équilibres et de résolution des problèmes inverses associés.

1.2 Reconstruction de l'équilibre à partir de mesures expérimentales

J'ai commencé à travailler sur cette thématique à mon arrivée à Nice en septembre 2007 avec Jacques Blum et Cédric Boulbe. Cédric était tout d'abord ATER (2007-2008) puis post-doc (2008-2010) et est maintenant Maitre de conférence à l'Université de Nice Sophia Antipolis depuis 2010.

Nous avons en particulier développé le code de calcul EQUINOX qui permet de reconstruire l'équilibre en temps réel au cours d'une décharge. Grâce à notre très bonne collaboration avec des collègues du CEA le code a pu être testé et validé avec des mesures des Tokamaks ToreSupra (Cadarache) et JET (Culham, UK). Avec Cédric nous avons effectué plusieurs séjours de travail d'une à deux semaines au JET. Le travail sur ce thème est synthétisé dans la section 1.2.1.

J'ai également travaillé sur le problème de la reconstruction de la frontière plasma seule i.e sans identification de la densité de courant. Ceci est synthétisé dans la section 1.2.2. Avec Amel Ben Abda (Prof. Ecole d'ingénieur de Tunis) nous avons étudié une méthode basée sur la minimisation d'un fonctionnelle de type "Kohn-Vogelius". Plus tard je me suis intéressé à l'utilisation des harmoniques toroidales pour l'interpolation 2D des mesures magnétiques. Ce problème d'interpolation est fortement lié à celui de la reconstruction de la frontière plasma. La motivation pour ce travail est double. D'une part le nouveau tokamak WEST au CEA Cadarache nécessite le développement

d'une méthode de reconstruction de la frontière pour le contrôle de la position du plasma. Et d'autre part l'interpolation des mesures magnétiques permet de fournir des conditions limites au code EQUINOX de manière générique et donc de l'intégrer à la plateforme européenne ITM [1] pour pouvoir être utilisé sur n'importe quel Tokamak. Dans le cadre de l'approche élargie ITER et de la collaboration avec la Chine, le code EQUINOX a récemment été choisi pour être utilisé sur le Tokamak JT60.

1.2.1 Identification de la densité de courant

- ▷ **Articles disponibles dans la partie recueil d'articles** : Article A [4], Article B [28]
- ▷ **Autres publications associées** : [10, 24, 27, 29, 25, 23]
- ▷ **Collaborateurs** : Jacques Blum (Université de Nice), Cédric Boulbe (Université de Nice), Sylvain Brémond (CEA), Didier Mazon (CEA), Philippe Moreau (CEA), Eric Nardon (CEA), Francois Saint-Laurent (CEA)

L'objectif de la reconstruction d'équilibre est double. Il s'agit d'une part de trouver la frontière plasma et d'autre part d'identifier la densité de courant, à savoir le second membre de l'équation de Grad-Shafranov, les fonctions $\frac{\partial p}{\partial \psi}$ et $\frac{\partial f^2}{\partial \psi}$ ne pouvant pas être mesurées directement. Dans les tokamaks actuels seule la frontière plasma (ou même uniquement quelques points de cette frontière) est utilisée pour le contrôle du plasma en temps réel au cours d'une décharge. Le profil de courant pourrait à terme être contrôlé après avoir été identifié avec une méthode du type de celle présentée ci-dessous.

Mesures expérimentales

- Les mesures expérimentales essentielles permettant la reconstruction d'équilibre sont les mesures magnétiques externes. Des bobines mesurent le champ magnétique poloidal et des boucles de flux mesurent le flux poloidal en différents points autour de la chambre à vide. Considérons ici que l'on est capable d'obtenir après un prétraitement des données magnétiques des données de Cauchy à savoir, la valeur du flux $\psi = g$ et de sa dérivée normale $\frac{1}{r}\frac{\partial \psi}{\partial n} = h$ en tout point d'un contour Γ_V définissant le domaine de calcul Ω ($\partial\Omega = \Gamma_V$ et Ω est la chambre à vide par exemple). Ce prétraitement peut être une simple interpolation linéaire des mesures si cela est possible, ou bien le résultat d'un code de reconstruction de frontière qui calcule ψ à l'extérieur du plasma satisfaisant $\Delta^*\psi = 0$ sous la contrainte des données magnétiques. Nous revenons sur ce point à la section 1.2.2. Les mesures magnétiques permettent également d'obtenir une mesure du courant plasma $I_p = \int_{\Omega_p} j_\phi dx$
- D'autres mesures, internes, apportent une information importante pour l'identification du profil de courant. Il s'agit en premier lieu des mesures d'interférométrie et de polarimétrie. Les premières fournissent les valeurs des intégrales le long de cordes C_i traversant le plasma de la densité électronique $n_e(\psi)$ considérée comme constante sur les isoflux :

$$\gamma_i = \int_{C_i} n_e(\psi)\, dl$$

1. Le projet ITM, Integrated Tokamak Modeling [62], fait maintenant partie du programme EUROFUSION. Il a pour ambition de regrouper les codes de calculs européens liés à la fusion dans un environnement informatique unifié afin qu'ils puissent être utilisés par différents laboratoires, comparés entre-eux et couplés dans le but de créer un véritable simulateur pour ITER

Les secondes fournissent les intégrales

$$\alpha_i = \int_{C_i} \frac{n_e(\psi)}{r} \frac{\partial \psi}{\partial n} dl$$

sur les mêmes cordes. $\frac{\partial \psi}{\partial n}$ est la dérivée normale de ψ le long de C_i. Des mesures de MSE (motional Stark effect) peuvent également être utilisées. Elles apportent une information sur la valeur du champ magnétique en certains points du plasma. La possibilité d'utiliser ces mesures est implémentée dans le code de calcul EQUINOX mais elles ne sont pas utilisées de manière routinière.

Problème direct

Comme déja dit, le problème de l'équilibre du plasma est un problème à frontière libre. Le domaine du plasma est défini par

$$\Omega_{\rm p}(\psi) = \{(r,z) \in \Omega_{\rm L} \ , \psi(r,z) > \psi_b\} \, .$$

où la valeur du flux à la frontière du plasma est

$$\psi_b = \max \left(\max_{(r,z)\in\partial\Omega_{\rm L}} \psi(r,z), \max_{(r_{\rm X},z_{\rm X})\in\Omega_{\rm L}} \psi(r_{\rm X},z_{\rm X}) \right)$$

et $(r_{\rm X}, z_{\rm X})$ est un éventuel point-X de $\psi(r,z)$. On note également ψ_a la valeur de ψ à l'axe magnétique i.e le maximum de ψ dans $\Omega_{\rm p}$. On introduit alors le flux normalisé $\psi_N = \frac{\psi - \psi_a}{\psi_b - \psi_a} \in [0,1]$ pour tout point de $\Omega_{\rm p}$, et les fonctions sans dimension définies sur $[0,1]$

$$A(\psi_N) = \frac{r_0}{\lambda} \frac{\partial p}{\partial \psi}(\psi) \quad \text{et} \quad B(\psi_N) = \frac{1}{\lambda \mu_0 2 r_0} \frac{\partial f^2}{\partial \psi}(\psi)$$

où λ est un facteur de normalisation et r_0 une constante.

En imposant des conditions de Dirichlet au bord le problème direct s'écrit

$$\left\{ \begin{array}{rcl} -\Delta^* \psi & = & \lambda [\frac{r}{r_0} A(\psi_N) + \frac{r_0}{r} B(\psi_N)] \chi_{\Omega_p(\psi)} \quad \text{dans } \Omega \\ \psi & = & g \quad \text{sur } \Gamma_V \end{array} \right. \tag{1.7}$$

L'aspect frontière libre du problème apparait comme une non-linéarité particulière à travers la fonction caractéristique χ_{Ω_p}. Le paramètre de normalisation λ permet de s'assurer que la valeur donnée du courant plasma I_p est bien vérifiée dans le modèle.

Problème inverse

Le problème inverse consiste en l'identification des fonctions A and B à partir des mesures disponibles. Il est formulé comme un problème aux moindres carrés dans lequel on cherche à minimiser une fonction coût J définie comme

$$J(A,B,n_e) = J_0 + J_1 + J_2 + J_\varepsilon$$

J_0 mesure l'écart aux données de la composante tangentielle de $\mathbf{B_p}$

$$J_0 = \frac{1}{2} \sum_{k=1}^{N} (w_k)^2 (\frac{1}{r} \frac{\partial \psi}{\partial n}(M_k) - h(M_k))^2$$

où N est le nombre de points M_k de la frontière $\partial\Omega$ où les mesures sont données.

$$J_1 = \frac{1}{2}\sum_{k=1}^{N_c}(w_k^{polar})^2(\int_{C_k}\frac{n_e(\psi_N)}{r}\frac{\partial\psi}{\partial n}dl - \alpha_k)^2$$

et

$$J_2 = \frac{1}{2}\sum_{k=1}^{N_c}(w_k^{inter})^2(\int_{C_k} n_e(\psi_N)dl - \gamma_k)^2$$

N_c est le nombre de cordes pour lesquelles les mesures d'interférométrie et polarimétrie sont données. Les poids w donnent l'importance relative des différentes mesures. Enfin le problème inverse d'identification étant mal-posé un terme de régularisation de Tikhonov J_ε est introduit

$$J_\varepsilon = \frac{\varepsilon}{2}\int_0^1 [A''(x)]^2dx + \frac{\varepsilon}{2}\int_0^1 [B''(x)]^2dx + \frac{\varepsilon_{n_e}}{2}\int_0^1 [n_e''(x)]^2dx$$

où ε et ε_{n_e} sont les paramètres de régularisation.

Si les mesures magnétiques sont utilisées seules alors seulement A et B apparaissent comme variables de contrôle et les termes J_1 et J_2 ne sont pas utiles. Si les mesures de polarimétrie et d'interférométrie sont utilisées alors la densité $n_e(\psi_N)$ doit également être identifiée même si elle n'apparait pas dans le modèle direct (1.7).

Méthodes numériques

Un des objectifs de ce travail étant la reconstruction d'équilibres en temps réel au cours d'une décharge les méthodes numériques que nous proposons sont simples mais efficaces. Le problème direct est discrétisé par éléments finis P1 et les non-linéarités traitées par une méthode de type point fixe. Les fonctions A, B et éventuellement n_e à identifier sont décomposées dans une base $(\Phi_i)_{i=1,...,m}$ de fonctions définies sur $[0,1]$ (fonctions linéaires par morceaux, splines cubiques, Bsplines, ...)

$$A(x) = \sum_i^m a_i\Phi_i(x) \text{ et } B(x) = \sum_i^m b_i\Phi_i(x).$$

Notons $u = (a_1, ..., a_m, b_1, ..., b_m) \in \mathbb{R}^{2m}$ le vecteur des composantes des fonctions A and B dans la base (Φ_i). Après discrétisation le problème direct (1.7) peut s'écrire

$$K\psi = Y(\psi^*)u + g \tag{1.8}$$

où ψ représente ici les valeurs du flux aux noeuds du maillage et ψ^* représente les valeurs connues de l'itération précédente, K est la matrice de masse, $Y(\psi^*)u$ est le second membre linéaire en u, non-linéaire en ψ, et g représente les conditions limites de Dirichlet. La fonction coût discrète s'écrit

$$J(u) = \frac{1}{2}\|C(\psi)\psi - d\|_D^2 + \frac{\varepsilon}{2}u^T\Lambda u$$

On peut faire apparaitre explicitement la dépendance en u en approchant ψ par ψ^* dans la matrice C et en injectant (1.8) pour remplacer le ψ restant. On obtient alors une fonctionnelle quadratique en u

$$J(u) = \frac{1}{2}\|Eu - f\|_D^2 + \frac{\varepsilon}{2}u^T\Lambda u$$

avec $E = C(\psi^*)K^{-1}Y(\psi^*)$ and $f = -C(\psi^*)K^{-1}g + d$, que l'on minimise directement en résolvant l'équation normale.

Ainsi les problèmes direct et inverse sont résolus simultanément par une méthode de type point fixe dans laquelle le vecteur u est mis à jour à chaque itération.

Cette méthode est particulièrement bien adaptée aux applications temps réel. Toutes les 100 ms environ on reconstruit l'équilibre en prenant comme état initial l'équilibre reconstruit au pas de temps précédant. L'algorithme converge en quelques itérations.

Quelques résultats numériques

La figure 1.2 montre deux exemples de reconstuction d'équilibre avec le code EQUINOX. Dans l'article [4] différentes expériences numériques sont menées avec des données simulées afin de tester et valider l'algorithme. On y étudie notamment l'influence du paramètre de régularisation sur la reconstruction. Il apparait que si la qualité de l'identification des fonctions A et B dépend beaucoup de la valeur de ce paramètre, en revanche la qualité de la reconstruction du profil de courant moyenné[2] $\frac{<j_\phi/r>}{<1/r>}$ et du facteur de sécurité[3] q en est beaucoup moins dépendante. Pour une large gamme de valeurs du paramètre de régularisation ε, même si A et B ne sont pas parfaitement reconstruits, le profil de courant moyenné lui est très bon. Ceci explique que la méthode L-curve [60] pour le choix de la valeur du paramètre de régularisation fonctionne mal.

Le code a été testé, règlé et validé sur plus d'une centaine de décharges du JET et de ToreSupra. Pour ces deux tokamaks les données de Cauchy au bord du domaine de calcul sont fournies par les codes de reconstruction du flux dans le vide XLOC pour JET [73, 78] et APOLO pour ToreSupra [77]. L'article [28] relate en partie ce travail de validation. Un résultat intéressant est le bon accord trouvé entre les sorties du code et des données indépendantes issues d'études MHD donnant la position des surface $q = 3/2$ (voir fig. 1.3).

2. Moyenne sur les surfaces magnétiques : la moyenne $< A >$ d'une quantité quelconque A sur la surface magnétique S_ψ correspondant à une valeur ψ du flux poloidal dans le plasma [55, 56, 43] est définie comme $< A >= \frac{\partial}{\partial V}\int_V AdV$ où V est le volume contenu dans S_ψ. Cette moyenne a la propriété suivante utilisée dans les calculs $< A >= \int_{C_\psi} \frac{Adl}{B_p} / \int_{C_\psi} \frac{dl}{B_p}$ où C_ψ est le contour isoflux fermé et $B_p = \frac{1}{r}||\nabla\psi||$

3. Le facteur de sécurité q tire son nom du rôle qu'il joue dans les études stabilité MHD [85]. On peut le voir comme la variation d'angle toroidal obtenue lorsqu'une ligne de champ effectue un tour poloidal complet, $q = \frac{\Delta\phi}{2\pi}$. Le facteur de sécurité est constant sur les surfaces de flux et est calculé de la manière suivante $q(\psi) = \frac{1}{2\pi}\int_{C_\psi} \frac{B_\phi}{rB_p} dl$ où $B_\phi = \frac{f}{r}$ est la composante toroidale du champ magnétique

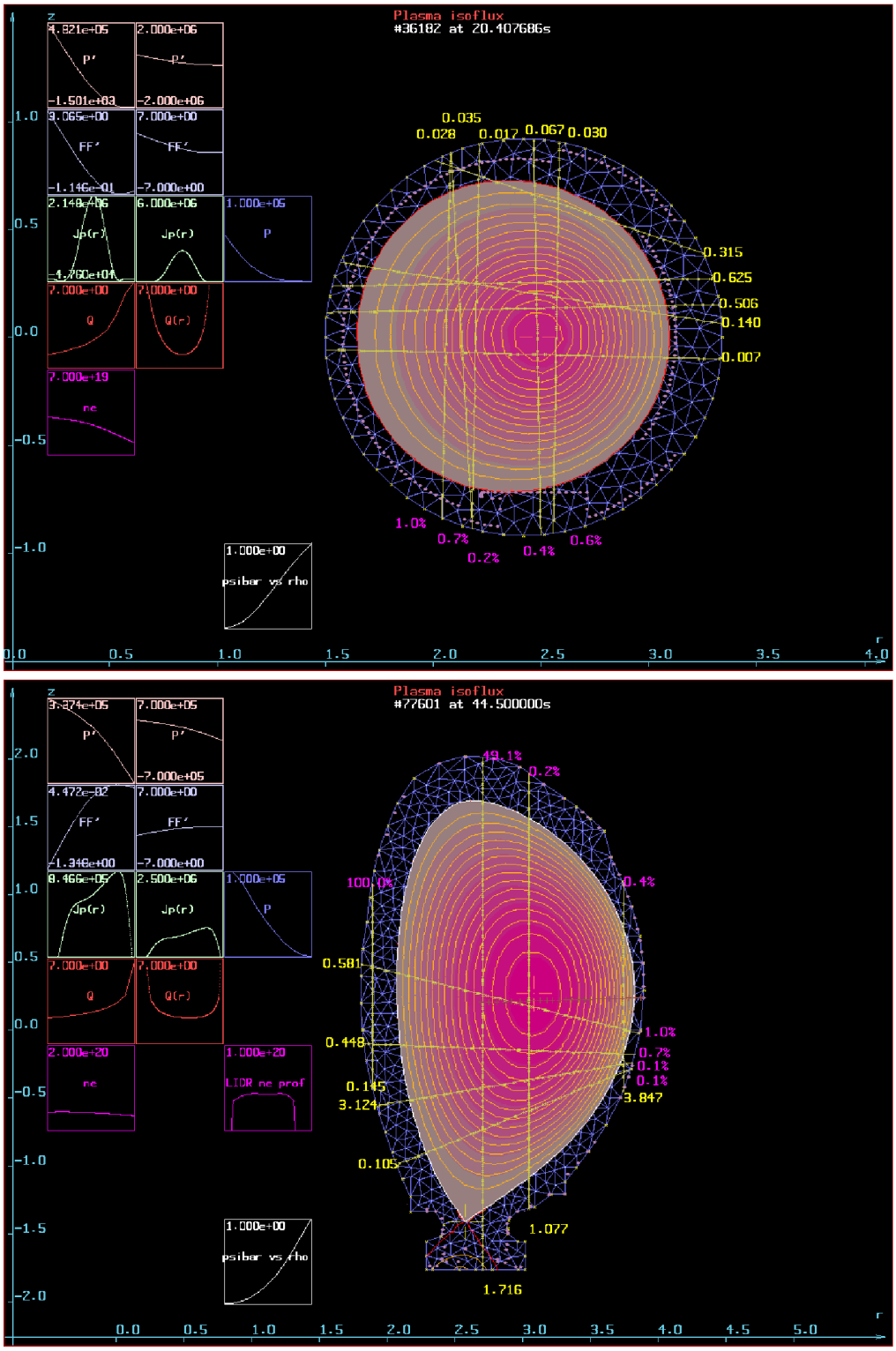

FIGURE 1.2 – Equilibres reconstruits. ToreSupra en haut et JET en bas. Le maillage du domaine apparait là où le plasma en couleur n'est pas présent. Dans le plasma les isoflux sont représentées depuis l'axe magnétique jusqu'à la frontière. Les cordes des mesures d'interférométrie et polarimétrie sont représentées en jaune. Les graphes sur la gauche permettent de visualiser différents profils comme les fonctions p' et ff' identifiées, la densité de courant toroidal fonction de r sur l'axe magnétique $z = z_a$, la pression, le facteur de sécurité q et la densité électronique identifiée.

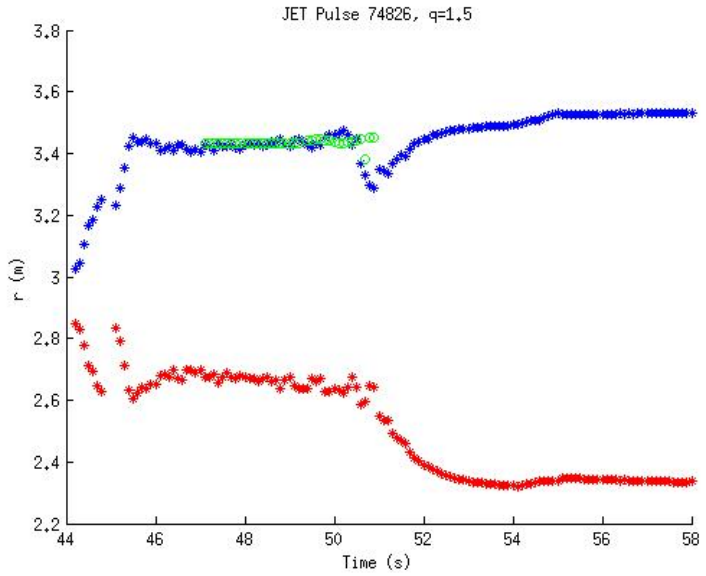

FIGURE 1.3 – Comparaison de la position des surfaces magnétiques $q = 3/2$ au cours d'une décharge JET. En rouge et bleu sorties d'EQUINOX côté fort et faible champ. En vert position trouvée par analyse MHD indépendente basée sur des mesures ECE (electron cyclotron emission).

1.2.2 Reconstruction de la frontière du plasma seule

- ▷ **Articles disponibles dans la partie recueil d'articles** : Article C [7], Article D [6]
- ▷ **Autres publications associées** : [26, 35]
- ▷ **Collaborateurs** : Jacques Blum (Université de Nice), Cédric Boulbe (Université de Nice), Sylvain Brémond (CEA), Didier Mazon (CEA), Philippe Moreau (CEA), Eric Nardon (CEA), Francois Saint-Laurent (CEA), Amel Ben Abda (Ecole d'ingénieur de Tunis)

Dans cette section on traite uniquement de la reconstruction du flux poloidal dans le vide entourant le plasma et de la frontière du plasma. C'est-à-dire que l'on ne cherche pas à résoudre l'équation de Grad-Shafranov dans le plasma comme précédemment.

Minimisation d'une fonctionnelle de Kohn-Vogelius

Comme dans la section précédente on considère ici que l'on dispose d'un jeu complet de données de Cauchy $g = \psi$ et $h = \dfrac{1}{r}\dfrac{\partial \psi}{\partial n}$ sur Γ_V. Le problème de l'identification de la frontière plasma se présente alors comme un problème de Cauchy que l'on résout par minimisation d'une fonctionnelle de type "'Kohn-Vogelius".

Notons Ω_X le sous domaine de la chambre à vide Ω_V où le plasma n'est pas présent (voir Fig. 1.4). Le flux poloidal satisfait

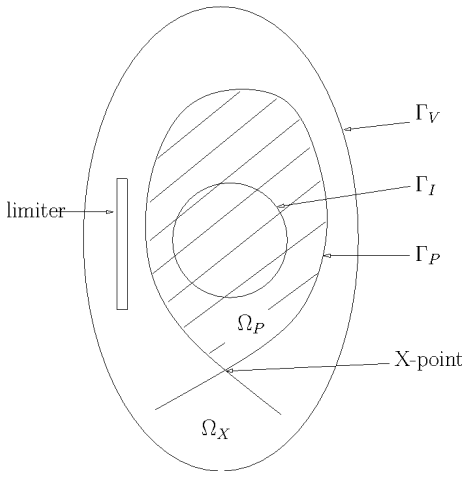

FIGURE 1.4 – Le domaine Ω_V de fontière Γ_V est la réunion du domaine Ω_X et du domaine plasma Ω_p dont la frontière Γ_p est ici définie par un point-X. Γ_I est le contour fictif.

$$\begin{cases} L\psi = 0 \quad \text{dans } \Omega_X \\ \psi = g \quad \text{sur } \Gamma_V \\ \dfrac{1}{r}\dfrac{\partial \psi}{\partial n} = h \quad \text{sur } \Gamma_V \\ \psi = \psi_b \quad \text{sur } \Gamma_P \end{cases} \tag{1.9}$$

où $L = \mu_0 \Delta^*$. Ici le domaine $\Omega_X = \Omega_X(\psi)$ est inconnu car la frontière plasma Γ_p l'est. De plus ce problème est mal posé en raison des deux conditions limites données sur Γ_V.

Afin de calculer le flux dans le vide et de trouver la frontière plasma, on commence par définir un problème approché, posé sur un domaine annulaire fixe Ω de fontière extérieure Γ_V et de frontière intérieure un contour fictif Γ_I fixe contenu dans le plasma (voir Fig. 1.4).

La seconde étape est alors de découper le problème en deux sous-problèmes bien posés. Dans le premier on ne considère que la donnée de Dirichlet f sur Γ_V et une donnée de Dirichlet v sur Γ_I

$$\begin{cases} L\psi_D = 0 \quad \text{dans } \Omega \\ \psi_D = g \quad \text{sur } \Gamma_V \\ \psi_D = v \quad \text{sur } \Gamma_I \end{cases} \tag{1.10}$$

et dans le deuxième on ne retient que la condition de Neumann

$$\begin{cases} L\psi_N = 0 \quad \text{dans } \Omega \\ \dfrac{1}{r}\dfrac{\partial \psi_N}{\partial n} = h \quad \text{sur } \Gamma_V \\ \psi_N = v \quad \text{sur } \Gamma_I \end{cases} \tag{1.11}$$

En notant $\psi_D(v, f)$ et $\psi_N(v, g)$ les solutions des deux problèmes (1.10) et (1.11), on cherche à trouver la condition limite $u_\varepsilon \in \mathcal{U} = H^{1/2}(\Gamma_I)$ qui minimise la fonctionnelle

$$J_\varepsilon(u) = \frac{1}{2}\int_\Omega \frac{1}{r}||\nabla\psi_D(u,g) - \nabla\psi_N(u,h)||^2 dx + \frac{\varepsilon}{2}\int_\Omega \frac{1}{r}||\nabla\psi_D(u,g)||^2 dx \tag{1.12}$$

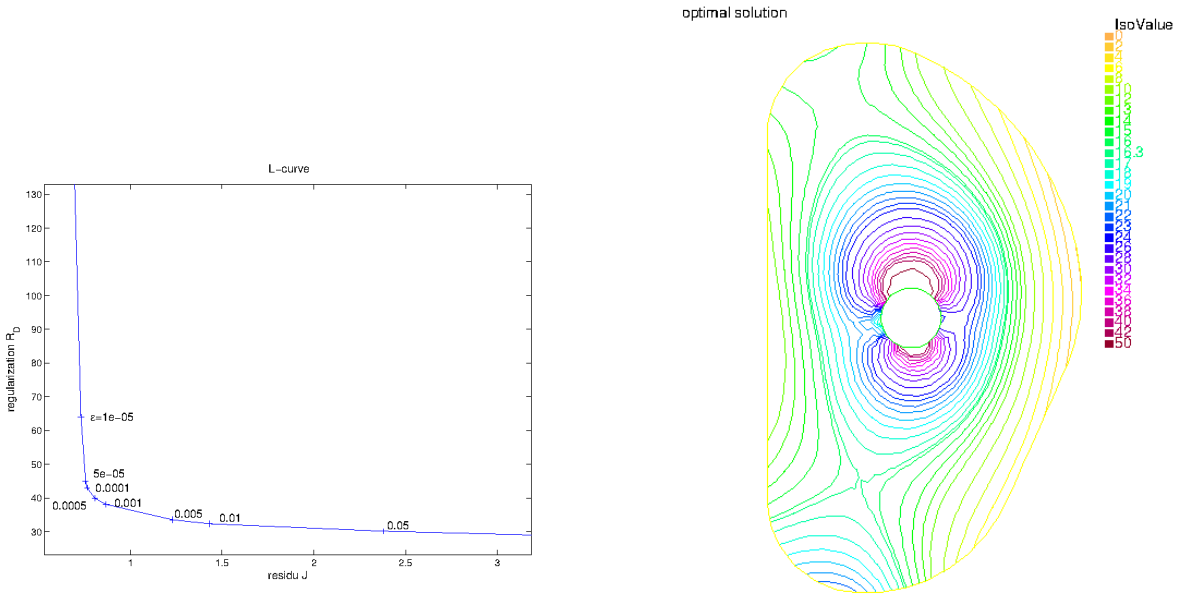

FIGURE 1.5 – Reconstruction du flux poloidal dans le vide pour la géométrie ITER. Gauche : L-curve donnant $\varepsilon = 5 \times 10^{-4}$. Droite : Reconstruction de ψ dans le domaine Ω

Le premier terme mesure l'écart entre les solutions des deux problèmes de Dirichlet et Neumann et le second est un terme de régularisation. On montre dans [6] que ce problème de minimisation admet une unique solution. La condition d'optimalité du premier ordre la caractérisant est la simple égalité variationnelle suivante

$$(J'_\varepsilon(u_\varepsilon), v) = \varepsilon s_D(u_\varepsilon, v) + s_D(u_\varepsilon, v) - s_N(u_\varepsilon, v) - l(v) = 0 \quad \forall v \in \mathcal{U} \tag{1.13}$$

où s_D et s_N sont des formes bilinéaires et l une forme linéaire. De plus la solution est stable par rapport aux données g et h.

Cette méthode donne de bons résultats illustrés par les expériences numériques de l'article [7] et par la Fig. 1.5 pour la géométrie ITER avec données de Cauchy simulées. On utilise une discrétisation en éléments finis P1. La méthode est très rapide, la condition d'optimalité revenant à résoudre un système linéaire $\mathbf{Su} = \mathbf{l}$ de taille N_{Γ_I} le nombre de noeuds du contour Γ_I. La matrice $\mathbf{S}$ ne dépend que de la géométrie fixe du problème et n'a pas à être recalculée pour chaque nouveau jeu de données. Le paramètre de régularisation peut être choisi par la méthode de la L-curve comme montré sur la Fig. 1.5.

Cette première méthode de reconstruction du flux poloidal dans le vide et donc d'identification de la frontière plasma est élégante et efficace. Néanmoins, comme c'est le cas pour la méthode d'identification de la densité de courant avec le code EQUINOX présenté à la section précédente, elle présente un défaut lié à l'hypothèse selon laquelle on dispose de données de Cauchy sur Γ_V. L'obtention de données de Cauchy sur un contour à partir des véritables mesures discrètes n'est généralement pas aisé. Comme déjà dit si les capteurs sont placés sur un contour et suffisamment proches les uns des autres, une simple interpolation peut être envisagée. Cette méthode n'est néanmoins pas robuste en cas de capteur défaillant. De plus dans les Tokamaks actuels comme le JET les capteurs sont répartis dans une région annulaire autour de la chambre à vide et pas forcément le long d'un unique contour. Pour ces raisons j'ai travaillé sur une méthode d'interpolation 2D des données magnétiques présentée dans le paragraphe suivant.

Méthode des harmoniques toroidales

Les harmoniques toroidales sont des solutions explicites de l'équation $\Delta^*\psi = 0$ [46]. Elles sont obtenues en utilisant une méthode de quasi-séparation de variable pour la recherche de solutions de cette équation sur un domaine annulaire D dans un système de coordonnées toroidales (voir par exemple [50] pour le détails des calculs). Ces coordonnées $(\zeta, \eta) \in \mathbb{R}_*^+ \times [0, 2\pi]$, aussi appelées coordonnées bipolaires lorsque comme ici on omet l'angle toroidal sont reliées aux coordonnées cylindriques (r, z) par

$$r = \frac{r_0 \sinh\zeta}{\cosh\zeta - \cos\eta} \quad \text{et } z - z_0 = \frac{r_0 \sin\eta}{\cosh\zeta - \cos\eta}$$

où $F_0 = (r_0, z_0)$ avec $r_0 > 0$ est le pôle du système de coordonnées. En supposant que ce pôle se trouve dans le domaine entouré par la couronne D la solution du problème

$$\left\{ \begin{array}{rcl} \Delta^*\psi & = & 0 \quad \text{dans} \quad D \\ \psi & = & g \quad \text{sur} \quad \partial D \end{array} \right. \tag{1.14}$$

peut être décomposé de manière unique sous la forme

$$\left\{ \begin{array}{rcl} \psi & = & \psi_{ext} + \psi_{int} \\ \psi_{ext} & = & \dfrac{r_0 \sinh\zeta}{\sqrt{\cosh\zeta - \cos\eta}} \times \\ & & [\displaystyle\sum_{n=0}^{\infty} a_n^e Q^1_{n-1/2}(\cosh\zeta)\cos(n\eta) + \sum_{n=1}^{\infty} b_n^e Q^1_{n-1/2}(\cosh\zeta)\sin(n\eta)] \\ \psi_{int} & = & \dfrac{r_0 \sinh\zeta}{\sqrt{\cosh\zeta - \cos\eta}} \times \\ & & [\displaystyle\sum_{n=0}^{\infty} a_n^i P^1_{n-1/2}(\cosh\zeta)\cos(n\eta) + \sum_{n=1}^{\infty} b_n^i P^1_{n-1/2}(\cosh\zeta)\sin(n\eta)] \end{array} \right. \tag{1.15}$$

où $P^1_{n-1/2}$ et $Q^1_{n-1/2}$ sont les fonctions de Legendre associées de première et deuxième espèce, de degré 1 et d'ordre demi entier [36]. Elles sont aussi appelées harmoniques toroidales lorsqu'elles sont évaluées au point $\cosh\zeta$. Les fonctions $P^1_{n-1/2}$ présentent une singularité lorsque $\zeta \to \infty$ i.e au point F_0 et ψ_{int} represente le flux généré par des courants qui circuleraient dans le domaine entouré par la couronne D. Ce dernier doit donc contenir le pôle F_0 car ψ est régulière dans D. Au contraire les fonctions $Q^1_{n-1/2}$ sont singulières en $\zeta \to 0$ i.e sur l'axe $r = 0$ et ψ_{ext} represente le flux généré par des courants qui circuleraient à l'extérieur de D.

Dans l'article [7] on utilise un développement tronqué en harmoniques toroidales pour représenter le flux dans un domaine annulaire fictif D incluant tous les capteurs magnétiques. Le problème d'interpolation 2D des données magnétiques est formulé comme le problème de minimisation d'une fonction coût $J(u)$ dépendant des coefficients du développement

$$u = (a_0^e, \cdots a_{n_e}^e, b_1^e, \cdots b_{n_e}^e, a_0^i, \cdots a_{n_i}^i, b_1^i, \cdots b_{n_i}^i)$$

et mesurant l'écart au sens des moindres carrés entre les mesures et les valeurs données par le développement. Cette fonction coût est quadratique et est minimisée directement en résolvant l'équation normale. On montre que l'on peut utiliser en plus, même en présence de structures ferromagnétiques, une modélisation à l'aide de fonctions de Green du flux généré par les courants circulant dans les bobines de champ poloidal. Ceci peut être important si ces bobines sont très proches des points de mesures, comme les bobines de divertor permettant de créer le point-X. Cela consiste simplement en une soustraction de la contribution de ces bobines aux mesures et permet de réduire le nombre d'harmoniques extérieures utilisées dans le développement.

On obtient au final une représentation explicite du flux dans le domaine D que l'on peut évaluer (ainsi que sa dérivée) sur n'importe quelle contour Γ de D afin d'obtenir des conditions de Cauchy

sur ce contour et fournir ainsi des conditions limites au code de reconstruction EQUINOX déjà présenté à la section 1.2.1.

On obtient en fait plus que cela. En effet si le contour interne du domaine fictif D ne peut pas être dans le plasma où $\Delta^*\psi = 0$ n'est plus vérifiée, il peut par contre être donné par la frontière plasma. Le développement en harmoniques toroidales est valable jusqu'à la frontière plasma et peut donc permettre de la reconstruire. Ainsi le problème de l'interpolation 2D des mesures magnétiques est intimement lié à celui de la reconstruction de la frontière plasma. Le caractère mal posé de ce problème de reconstruction se manisfeste par la singularité au pôle F_0 des harmoniques $P^1_{n-1/2}$ et donc de la solution intérieure ψ_{int}. Cette solution intérieure ne dépend que du choix du pôle et du nombre d'harmoniques utilisées.

Différentes expériences numériques sont menées dans l'article [7] pour la configuration du tokamak WEST (voir Fig. 1.6). Le code d'équilibre CEDRES++, présenté à la section suivante 1.3, est utilisé pour créer des mesures expérimentales synthétiques pour différentes configurations de frontière plasma. Les résultats numériques montrent qu'un bon choix pour le pôle est celui de l'axe magnétique du plasma qui peut être facilement approché par des moments de la densité de courant dans le plasma. Ils montrent également qu'un faible nombre d'harmoniques permet d'obtenir une excellente approximation du flux tout en permettant de reconstruire une frontière plasma régulière. Prendre des harmoniques d'ordre 4 (i.e 9 fonctions de base pour ψ_{int} et 9 pour ψ_{ext}) est optimal au sens où en prendre plus ne permet pas de diminuer significativement l'écart aux données et prendre plus d'harmoniques intérieures peut par contre dégrader la qualité de la reconstruction de la frontière car la "zone d'explosion" de la solution intérieure peut s'étendre jusqu'à la frontière du plasma (Fig. 1.7 gauche). Ce phénomène disparait lorsque l'équilibre est reconstruit en utilisant le code EQUINOX et l'interpolation des mesures par la méthode des harmoniques toroidales (Fig. 1.7 droite). Ceci n'est pas surprenant étant donné que le caractère mal-posé du problème inverse traité par EQUINOX porte sur l'identification des termes sources de la densité de courant, le caractère frontière libre du problème étant lui réduit à une non-linéarité particulière du modèle.

Les résultats numériques démontrent également une très bonne stabilité de la méthode notamment pour l'identification du point-X. Elle est peu sensible à une valeur erronnée d'un capteur. Elle est également générique au sens où elle peut être utilisée facilement pour n'importe quel Tokamak contrairement aux codes XLOC spécifique au JET et APOLO spécifique à ToreSupra. Enfin elle est peu coûteuse en temps de calcul et il est actuellement envisagé de l'utiliser dans le système temps réel de contrôle pour le tokamak WEST.

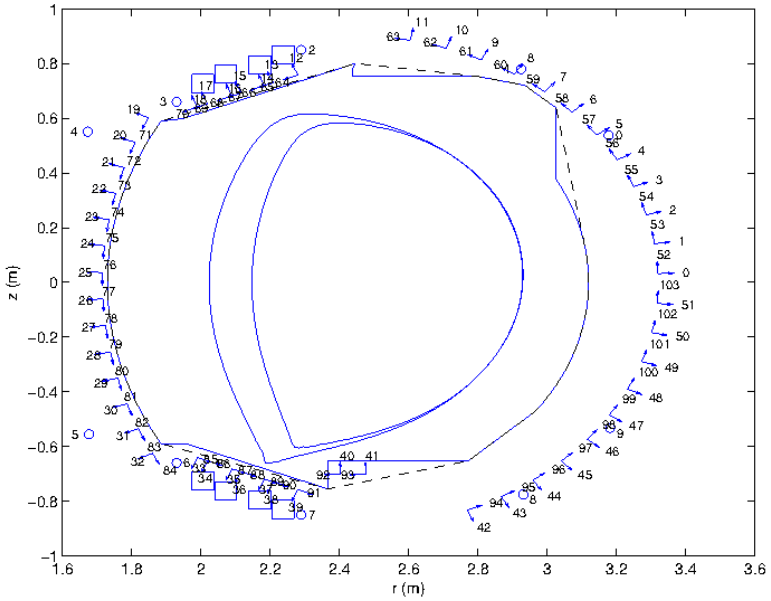

FIGURE 1.6 – Section poloidale du tokamak WEST. Deux frontières plasma différentes calculées à partir du code CEDRES++ sont représentées. Les bobines de mesures du champ poloidal sont représentées par des flèches et numérotées de 0 à 103. Les boucles de flux sont représentées par des cercles et numérotées de 0 à 9. Les quatres bobines basses du divertor et les quatres bobines hautes sont représentées. Le contour du limiteur est également représenté ainsi que son enveloppe convexe (en tirets) qui est utilisé comme contour Γ_V i.e la frontière du domaine de calcul pour le code EQUINOX.

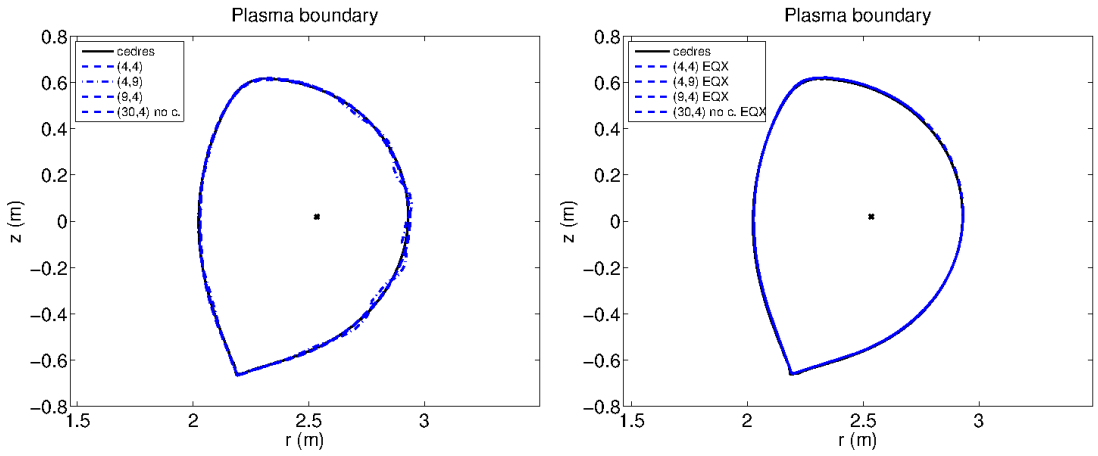

FIGURE 1.7 – Frontières plasma reconstruites. Gauche : à partir des harmoniques toroidales seules. Les frontières calculées avec $(n^e = 4, n^i = 4)$ ou $(n^e = 9, n^i = 4)$ et $(n^e = 30, n^i = 4)$ sans prendre en compte les bobines de champ poloidal (no c.) sont presque superposées avec la frontière de référence calculée avec le code CEDRES++. La frontière calculée avec $(n^e = 4, n^i = 9)$ présente des irrégularités. Droite : mêmes frontières calculées avec EQUINOX et interpolation des données sur le contour Γ_V de la Fig. 1.6

1.3 Calculs d'équilibres et couplage avec la diffusion résistive dans le plasma

- ▷ **Article disponibles dans la partie recueil d'articles** : Article E [9], Article F [5]
- ▷ **Autre publication associée** : [8]
- ▷ **Collaborateurs** : Jacques Blum (Université de Nice), Cédric Boulbe (Université de Nice), Holger Heumann (Université de Nice - INRIA), Gael Selig (Université de Nice), Jean-Marc Ané (CEA), Jean-Francois Artaud (CEA), Vincent Basiuk (CEA), Sylvain Brémond (CEA), Patrick Hertout (CEA), Philippe Huynh (CEA), Philippe Moreau (CEA), Eric Nardon (CEA)

Un des challenges actuels dans la communauté de la fusion et notamment du programme européen EUROFUSION (WPCD ITM) est de réussir à simuler une décharge complète dans un Tokamak à l'échelle de temps de la diffusion résistive. L'objectif est la mise au point de scénarios de décharge qui pourraient ainsi être testés numériquement pour la machine ITER.

La première brique fondamentale pour ce type de simulation est un code permettant le calcul de l'évolution de l'équilibre quasi-statique du plasma. Pour cette raison je commence par présenter le code CEDRES++ [9] ci-dessous avant d'aborder la question de son couplage avec l'équation de diffusion résistive dans le plasma.

Calculs directs et inverses d'équilibres avec le code CEDRES++

On s'intéresse ici en premier lieu à la simulation numérique directe de l'équilibre quasi-statique du plasma i.e avec les équations de la section 1.1 rappelées ci-dessous (pour le cas du modèle statique avec un courant plasma I_p non fixé).

On considère que les fonctions $\dfrac{\partial p}{\partial \psi}$ et $\dfrac{1}{2}\dfrac{\partial f^2}{\partial \psi}$ sont données sous la forme de deux fonctions définies pour $\psi_N \in [0,1]$, $S_{p'}(\psi_N)$ et $S_{ff'}(\psi_N)$. Les variables d'entrée du modèle direct ou variables de contrôle pour le problème inverse que nous allons décrire ensuite sont les courants dans les bobines de champs poloidal $I = (I_1, ...I_N)$. Le problème direct est alors le suivant. Trouver ψ dans Ω_∞ tel que

$$\begin{cases} -\Delta^*_\mu \psi = \begin{cases} r S_{p'}(\psi_N) + \dfrac{1}{\mu_0 r} S_{ff'}(\psi_N) & \text{dans } \Omega_{\rm p}(\psi)\,; \\ \dfrac{I_i}{S_i} & \text{dans } \Omega_{{\rm c}_i}\,; \\ 0 & \text{ailleurs}\,, \end{cases} \\ \psi(0,z) = 0\,; \quad \lim\limits_{\|(r,z)\| \to +\infty} \psi(r,z) = 0\,; \end{cases} \tag{1.16}$$

Le problème inverse associé consiste à définir une frontière plasma désirée Γ_{desi} et à chercher les courants dans les bobines qui permettent de l'obtenir. Ceci est formulé comme un problème de contrôle optimal dans lequel on cherche à minimiser la fonction coût définie ci-dessous. On considère $N_{\rm desi}+1$ points : $(r_{\rm desi}, z_{\rm desi}) \in \Gamma_{\rm desi}$ et $(r_1, z_1), \dots (r_{N_{\rm desi}}, z_{N_{\rm desi}}) \in \Gamma_{\rm desi}$ et la fonction coût

$$J(I) = \frac{1}{2}\sum_{i=1}^{N_{\rm desi}} \big(\psi(r_i, z_i) - \psi(r_{\rm desi}, z_{\rm desi})\big)^2 + \frac{1}{2}\sum_{i=1}^{N} w_i I_i^2$$

où ψ est relié à I par (1.16). Le premier terme s'annule si $\Gamma_{\rm desi}$ est une iso-ψ et le second est un terme de régularisation.

On peut définir trois autres variantes de ces problèmes directe et inverse. Une première concerne toujours le cas statique mais on cherche également à imposer le courant plasma I_p. On suppose

que $S_{p'}$ et $S_{ff'}$ ne sont connues qu'à un facteur scalaire λ près. On ajoute alors une inconnue, le facteur λ vérifiant

$$I_P = \lambda \int_{\Omega_{\rm p}(\psi)} \left(r S_{p'}(\psi_{\rm N}(r,z)) + \frac{1}{\mu_0 r} S_{ff'}(\psi_{\rm N}(r,z)) \right) dr dz. \tag{1.17}$$

Les deux autres variantes correspondent à la version dynamique de la modélisation directe dans laquelle les fonctions $S_{p'}(\psi_N, t)$ et $S_{ff'}(\psi_N, t)$ sont données pour tout t et les variables de contrôle sont les tensions dans les bobines. Pour le problème inverse on cherche les tensions fonctions du temps permettant d'obtenir une évolution de la frontière plasma désirée [45]

Les difficultées pour traiter ces problèmes sont dues d'une part au domaine infini Ω_∞ et d'autre part aux non-linéarités qui apparaissent dans la description de la frontière plasma $\Omega_{\rm p}(\psi)$, dans la densité de courant plasma avec $S_{p'}(\psi_N)$ et $S_{ff'}(\psi_N)$ et dans les structures ferromagnétiques avec $\mu(\psi)$.

Le code de calcul CEDRES++ au développement duquel nous participons permet aujourd'hui de traiter ces quatre problèmes directes et inverses et l'article [9] se veut être la référence sur les méthodes numériques utilisées et en contient une description précise. La première version de CEDRES++ date de 1999 [58]. C'est un code écrit en C++ qui reprend les méthodes développées dans les codes d'équilibre SCED [44] et PROTEUS [37, 38]. Cette version d'origine traite le cas d'une modélisation statique de l'équilibre avec I_p fixé en utilisant une discrétisation spatiale par éléments finis P1, une méthode de Newton pour résoudre les non-linéarités et une méthode d'éléments frontières pour ramener le domaine infini à un demi disque. Le problème inverse est traité par une méthode séquentielle quadratique.

En collaboration avec nos collègues du CEA à Cadarache nous avons récupéré et développé ce code à partir de 2009 et aujourd'hui CEDRES++ est un code performant et mature. Il est utilisé au niveau européen grâce à la plateforme de l'ITM [5] et c'est un outil de modélisation numérique important utilisé au CEA pour la préparation de WEST (fig. 1.8 et 1.9) comme le montrent différents exemples de simulation dans [9] ainsi que [72].

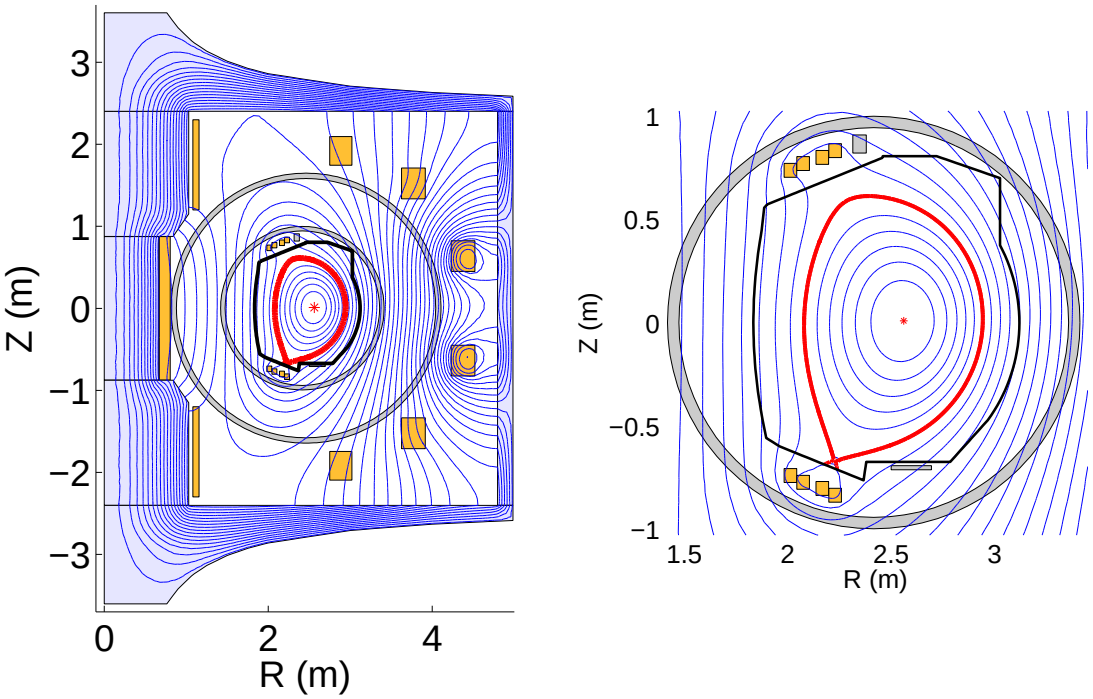

FIGURE 1.8 – Exemple de calcul d'équilibre pour le tokamak WEST représentant les isoflux dans le plan poloidal. Gauche : vue d'ensemble. Droite : zoom sur la chambre à vide. Le fer est représenté en bleu clair, les bobines de champ poloidal en orange et les structures passives en gris clair (chambre à ide et éléments de stabilisation verticale). La courbe noire est le limiteur. La courbe rouge est la frontière du plasma calculée. L'axe magnétique est également représenté en rouge.

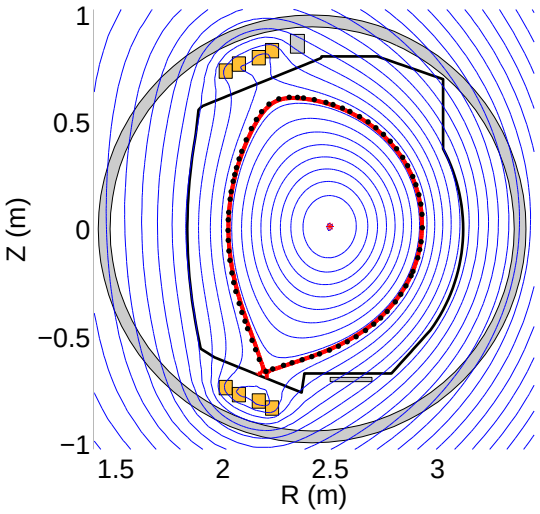

FIGURE 1.9 – Iso-ψ calculées par CEDRES++ en mode inverse statique. Les points noirs représentent la frontière désirée et la courbe rouge la frontière calculée.

Couplage entre équilibre et diffusion résistive dans le plasma

Au paragraphe précédant j'ai introduit le code de calcul d'équilibre CEDRES++ qui constitue le premier élément du problème de couplage. Il s'agit donc pour cette première brique de résoudre le système évolutif

$$\begin{cases} -\Delta^*_\mu \psi = \begin{cases} rS_{p'}(\psi_N,t) + \dfrac{1}{\mu_0 r} S_{ff'}(\psi_N,t) & \text{dans } \Omega_{\rm p}(\psi)\,; \\ \dfrac{n_i V_i(t)}{R_i S_i} - 2\pi \dfrac{2\pi n_i^2}{R_i S_i^2} \displaystyle\int_{\Omega_{{\rm c}_i}} \dot{\psi}\, drdz & \text{dans } \Omega_{{\rm c}_i}\,; \\ -\dfrac{\sigma}{r}\dot{\psi} & \text{dans } \Omega_{\rm ps}\,; \\ 0 & \text{ailleurs}\,, \end{cases} \\ \psi(0,z,t) = 0\,; \quad \displaystyle\lim_{\|(r,z)\|\to+\infty} \psi(r,z,t) = 0\,; \\ \psi(r,z,0) = \psi^0(r,z)\,, \end{cases} \tag{1.18}$$

Les fonctions

$$S_{p'}(\psi_N,t) = \frac{\partial p}{\partial \psi}(\psi(t)) \text{ et } S_{ff'}(\psi_N,t) = \frac{1}{2}\frac{\partial f^2}{\partial \psi}(\psi(t))$$

ne sont pas connues ici et contrairement aux calculs du paragraphe précédant on ne les impose pas mais on cherche à les calculer.

Le terme de pression ne représente pas le noeud du problème. La pression est calculée par les codes de transport comme CRONOS [39] au CEA ou ETS [47] dans l'ITM. Ces codes résolvent les équations d'évolution des densités et températures dans le plasma. A l'échelle de temps à laquelle on se place ces quantités sont constantes sur les surfaces de flux et les équations sont 1D. La dimension d'espace est un label des surfaces de flux. Différents choix sont possibles et celui fait dans ces codes est celui du label $\rho = \sqrt{\dfrac{\phi}{\pi B_0}}$ défini à partir du flux toroidal $\phi = \displaystyle\int_{D_\psi} B_\phi ds$ où D_ψ est le domaine intérieur à l'iso-contour C_ψ. La dérivation des équations 1D de transport fait appel à la technique des moyennes sur les surfaces magnétiques [55, 56]. Dans la suite de cette section nous utilisons les quantités géométriques moyennes [43] :

$$C_2 = V' < \frac{|\nabla\rho|^2}{r^2} > \text{ et } C_3 = V' < \frac{1}{r^2} >$$

avec $V' = \dfrac{\partial V}{\partial \rho}$, V étant le volume du plasma. Ces quantités sont calculées à chaque pas de temps par le code d'équilibre résolvant (1.18). Les coefficients géométriques étant donnés le code de transport fournit la pression $p(\rho,t)$. Nous la considérons ici comme connue.

La difficulté provient du calcul de l'évolution du terme diamagnétique $\dfrac{1}{2}\dfrac{\partial f^2}{\partial \psi}(\psi(t))$. Celui-ci doit être obtenu à partir des équations qui n'ont pas encore été utilisées à savoir la loi de Faraday

$$-\dot{\mathbf{B}} = \nabla \times \mathbf{E}$$

et la loi d'Ohm dans le plasma

$$\mathbf{E} + \mathbf{u} \times \mathbf{B} = \eta \mathbf{j}$$

dans laquelle η est le tenseur de résistivité du plasma et l'on omet les termes sources de courants non-inductifs. Lorsque le label indexant les surfaces de flux est choisi comme étant ρ, il est montré dans [43] que la composante toroidale de la loi de Faraday combinée à la loi d'Ohm projetée sur $\mathbf{B}$

et moyennée sur les surfaces de flux donne une équation de diffusion 1D pour le flux poloidal vu comme une fonction de ρ

$$\frac{\partial \psi}{\partial t} = \frac{\eta^{\parallel}\rho}{\mu_0 C_3^2}\frac{\partial}{\partial \rho}(\frac{C_2 C_3}{\rho}\frac{\partial \psi}{\partial \rho}) \tag{1.19}$$

où $\frac{\partial}{\partial t}$ désigne la dérivée temporelle à ρ fixé et la quantité scalaire $\eta^{\parallel}$ que l'on considère donnée également est la composante du tenseur de résistivité qui est parallèle aux lignes de champs.

L'équation (1.19) donne l'évolution du profil $\psi(\rho, t)$. On peut en déduire l'évolution du terme $\frac{\partial p}{\partial \psi}$ et enfin grâce à l'équation de Grad-Shafranov moyennée

$$-\frac{\partial}{\partial \rho}(C_2\frac{\partial \psi}{\partial \rho}) = \mu_0 V'\frac{\partial p}{\partial \psi} + \frac{C_3}{2}\frac{\partial f^2}{\partial \psi} \tag{1.20}$$

il est possible de calculer l'évolution du terme $\frac{1}{2}\frac{\partial f^2}{\partial \psi}$.

Les difficultés du couplage équilibre à frontière libre - diffusion résistive dans le plasma viennent en partie du fait qu'il faut garantir au cours de l'intégration numérique du système (1.18),(1.19) et (1.20) la consistance entre le flux poloidal vu comme une fonctions de l'espace $\psi(r, z)$ dans le code d'équilibre et vu comme comme une fonction $\psi(\rho)$ dans l'équation de diffusion résistive. A ceci s'ajoutent différentes difficultés numériques comme par exemple le calcul des coefficients géométriques en sortie de (1.18) et leur interpolation pour pouvoir être utilisés dans (1.19) et (1.20), le calculs de dérivées et les valeurs aux bords non connues pour les 3 termes de l'équation de Grad-Shafranov moyennée (1.20).

Les conditions aux limites pour (1.19) sont $\frac{\partial \psi}{\partial \rho}(0, t) = 0$ sur l'axe magnétique et au bord du plasma

$$\frac{\partial \psi}{\partial \rho}(\rho_{max}, t) = -\frac{2\pi\mu_0 I_p(t)}{C_2(\rho_{max}, t)} \tag{1.21}$$

où I_p est le courant plasma total.

Au cours du calcul d'un pas de temps $[t^n, t^{n+1}]$, les coefficients géométriques au temps t^n étant donnés, on commence par avancer (1.19) et (1.20) et calculer $\frac{\partial p}{\partial \psi}$ et $\frac{1}{2}\frac{\partial f^2}{\partial \psi}$ au temps t^{n+1} avant de calculer un nouvel équilibre avec (1.18). Le courant I_p doit être pris implicite au temps t^{n+1} dans la condition limite (1.21) mais évidemment cette valeur n'est pas connue étant donné quelle est calculée par le code d'équilibre (1.18) qui intervient dans un second temps. On peut par un processus itératif ou par une méthode de prédictieur-correcteur ([43] que nous avions reprise dans la thèse de Gael Selig [79]) assurer que le I_p^{n+1} fourni dans la condition limite (1.21) en entrée de l'équation de diffusion soit très proche de la valeur calculée par le code d'équilibre à la fin du pas de temps. Malheureusement les différentes expériences numériques menées avec les collègues du CEA et différentes méthodes et codes ont montré que ceci n'était pas suffisant pour assurer la consistance du couplage. Une divergence numérique peut apparaitre entre la valeur du flux au bord du plasma ψ_b calculée par le code d'équilibre d'une part et l'équation de diffusion d'autre part. Pour assurer la consistance du flux au bord nous utilisons dans [5] (section 4.3 et idée originale de Jean Francois Artaud (CEA)) une condition au bord du type

$$I_{p,D}^{n+1} = I_{p,E}^{n}(1 + \frac{\psi_{b,D}^n - \psi_{b,E}^n}{\psi_{a,E}^n - \psi_{b,E}^n})$$

où les indices D et E renvoient à la diffusion (1.19) et à l'équilibre (1.18) respectivement. Condition sur laquelle il est éventuellement possible d'itérer en faisant à nouveaux les calculs pour le même pas de temps. Cette méthode fonctionne relativement bien et permet d'obtenir des simulations

dans lesquelles les erreurs $e_I = |I_{p,D} - I_{p,E}|$ et $e_{\psi_b} = |\psi_{b,D} - \psi_{b,E}|$ sont petites. On peut également obtenir le même type de résultats, mais pour un coût de calcul plus élevé, en cherchant à chaque pas de temps $I_{p,D}$ minimisant la fonctionnelle

$$J(I) = (I - I_{p,E})^2 + \frac{w}{2}(\psi_{b,D} - \psi_{b,E})^2$$

Dans [5] (sec. 4.3) la simulation d'un VDE pour le tokamak ITER est donnée comme premier exemple de ce type de simulation. Beaucoup de travail reste néanmoins à faire avant de pouvoir obtenir de vraies simulations de scenarios de décharges. De nombreux problèmes numériques demeurent comme notamment le fait que les erreurs e_I et e_{ψ_b} ont tendance à croître au cours du temps jusqu'à l'arrêt de la simulation. Ajouté à cela il est vraissemblablement nécessaire d'utiliser dans ces simulations directes un algorithme de contrôle en boucle fermée pour calculer les tensions dans les bobines de champ poloidal permettant de stabiliser la position du plasma et d'éviter le VDE. Nous travaillons actuellement à ceci dans le cadre du WPCD EUROFUSION.

1.4 Conclusion

Dans ce chapitre j'ai donné un tour d'horizon de mon activité scientifique actuelle concernant la simulation numérique et les problèmes inverses en physique des plasmas de tokamak.

En ce qui concerne la reconstruction du flux poloidal dans le vide à partir des mesures magnétiques, la méthode des harmoniques toroidales apporte une bonne solution. Le code VacTH qui implémente cette méthode pour la reconstruction de la frontière plasma va bientôt entrer dans une période de tests intensifs au CEA avant d'être éventuellement un jour utilisé dans le système de contrôle temps réel du tokamak WEST. Dans le futur j'aimerai regarder la méthode des équations intégrales de frontière qui pourrait également apporter une réponse complémentaire à ce problème [65]. La présence de fer dans WEST risque néanmoins d'être problématique.

Concernant la reconstruction complète de l'équilibre, j'ai présenté la méthode implémentée dans le code EQUINOX. Ce dernier est un code mature qui a aujourd'hui été testé sur différents tokamaks avec succès notamment depuis son implantation dans la structure de l'ITM grâce à la méthode des harmoniques toroidales qui permet de calculer des conditions de Cauchy sur le bord du domaine de calcul [32]. Il est possible aujourd'hui avec EQUINOX, en plus des données magnétiques, d'utiliser des données d'interférométrie, polarimétrie et MSE. Il sera intéressant dans le futur de regarder la possibilité d'utiliser des données complémentaires liées à l'effet Cotton-Mouton sur la polarimétrie [74]. En effet ceci risque d'être important dans le tokamak ITER.

Enfin le challenge à relever est celui de la simulation de scénario à l'aide d'un modèle couplé équilibre à frontière libre - diffusion résistive dans le plasma.

Chapitre 2

Modélisation, assimilation variationnelle de données en écologie marine et autres travaux isolés.

Ce chapitre synthétise mon activité de recherche principalement pendant la période allant de la fin de ma thèse en octobre 2002 jusqu'à mon arrivée à Nice fin 2007.

J'ai effectué deux périodes de travail à l'IRD au sein de l'UR THETIS (Thons tropicaux et Ecosystèmes pélagiques : Taxies, Interactions et Stratégie d'exploitation). Une première de 16 mois comme post-doc de novembre 2003 à octobre 2005 et la seconde d'un an comme chargé de recherche d'octobre 2006 à septembre 2007. Pendant ces deux périodes j'ai travaillé sur des sujets connexes à celui de mon travail de thèse "Assimilation variationnelle de données dans un modèle couplé océan-biogéochimie" [1, 3, 2]. Mon principal collaborateur a été Olivier Maury, modélisateur spécialiste d'écologie halieutique à l'IRD. Nous avons travaillé ensemble et en parallèle, lui sur un modèle de flux d'énergie dans les écosystèmes marins qui est aujourd'hui devenu APECOSM [18, 19, 71] et moi sur un modèle plus destiné à être confronté directement aux données de pêche pour l'estimation de ses paramètres. Ce modèle est devenu le modèle APECOSM-E [11, 12] sur lequel nous avons également travaillé avec Sybille Dueri qui était post-doc à l'IRD. Cette aventure IRDienne a donné lieu aux articles [15, 14, 16, 11, 12, 18, 19] brièvement synthétisés dans les trois premières sections de ce chapitre.

Grâce à un court séjour post-doctoral dans l'équipe INRIA COMORE fin 2003, j'ai pu prolonger directement mon travail de thèse en collaboration avec Olivier Bernard (INRIA), Antoine Sciandra (CNRS) et Marina Lévy (CNRS). Ceci a donné lieu à l'article [13].

Pendant l'année scolaire 2002-2003 j'étais ATER à l'INSA de Lyon où j'ai travaillé avec Jérôme Pousin. Nous avons écrit deux articles [21, 20] présentés dans la dernière section du chapitre. Un troisième article isolé [22] que je ne présente pas ici a été écrit en collaboration avec Clément Faugeras (CNRS).

2.1 Généralités sur l'identification de paramètres par assimilation variationnelle de données

Construire un modèle exige de synthétiser la somme des connaissances accessibles à un moment donné sur un système, de sélectionner les processus important pour décrire un phénomène ainsi que les paramétrisations adéquates pour le quantifier. En écologie marine les modèles ne reposent sur aucune loi exacte et les erreurs de prévision peuvent s'expliquer en grande partie par une mauvaise

paramétrisation des processus ou par un mauvais choix des valeurs des paramètres. Ces derniers sont généralement nombreux et leurs valeurs sont mal connues. En effet ils ne représentent souvent rien de mesurable. Tout modèle de simulation en écologie est destiné à évoluer assez rapidement. De nouvelles expériences ou données vont mettre en évidence certains défauts et conduire à proposer un nouveau modèle. Ceci motive fortement la mise en place de méthodes numériques permettant d'ajuster les valeurs de ces paramètres pour que les modèles rendent compte le mieux possible des observations que l'on peut avoir sur le système modélisé. Ceci est typiquement un problème inverse d'estimation de paramètres et les méthodes d'assimilation variationnelle de données permettent de le résoudre. Le problème inverse est formulé comme un problème de minimisation d'une fonction coût mesurant l'écart du modèle aux données et dépendant des paramètres du modèle. La minimisation est effectuée par un algorithme de descente type gradient et une difficultée est le calcul de ce gradient à chaque pas de descente. La méthode adjointe venant de la théorie du contrôle optimal des équations aux dérivées partielles [68] et les techniques de différentiation automatique permettent de la surmonter.

Modèle direct et problème inverse

Afin d'introduire formellement le principe de l'estimation de paramètres par assimilation variationnelle de données, prenons un modèle générique. Celui-ci est pris comme c'est généralement l'usage pour simplifier comme un système différentiel non linéaire en dimension N provenant de la discrétisation en espace d'un modèle en équations aux dérivées partielles. Le modèle direct s'écrit :

$$\begin{cases} \dfrac{dx}{dt} = F(x,a), \quad 0 \leq t \leq T \\ x(0) = u \end{cases} \tag{2.1}$$

L'application F est non-linéaire et dépend d'un vecteur $a \in \mathbb{R}^{N_a}$ représentant les paramètres du modèle. Le vecteur u représente les conditions initiales. La construction de modèles qui a constitué une partie importante de mon travail à l'IRD fait l'objet de la section 2.2 suivante.

Supposons que l'on dispose d'observations $y_i = H_i(x(t_i)) + \varepsilon_i$ de l'état. H_i est l'opérateur non-linéaire d'observation à des instants t_i, $i = 1, \cdots m$ et ε_i représente l'erreur d'observation.

Le problème inverse est alors formulé comme un problème de minimisation pour la fonction coût

$$J(u,a) = \frac{1}{2}\sum_{i=1}^{m} ||H_i(x(t_i)) - y_i||^2_{W_i} + \frac{1}{2}||u - u_0||^2_{W_u} + \frac{1}{2}||a - a_0||^2_{W_a} \tag{2.2}$$

où $||x||^2_W = (Wx, x)$ et les matrices W_i, W_u et W_a sont des matrices symétriques, définies positives de pondération. Le premier terme est le terme d'écart aux données et les deux seconds des termes de pénalisation (ou d'ébauche) mesurant l'écart à des valeurs de référence u_0 et a_0. Ces termes permettent d'une part d'inclure l'information a priori des ces valeurs de références dans la fonction coût et d'autre part de régulariser le problème inverse étant donné son caractère mal-posé si on considère uniquement le terme d'écart aux observations.

Calcul du gradient, modèles tangent linéaire et adjoint

La minimisation se fait par un algorithme de descente qui nécessite le calcul du gradient de J. Ce calcul présente une difficulté technique qui est levée par l'introduction de l'état adjoint venant de la théorie du contrôle optimal [68, 66].

La dérivée de J au point (u,a) dans la direction (h,k) s'exprime en fonction de celle de x (on note ces dérivées $\hat{J}$ et $\hat{x}$)

$$\hat{J} = \sum_{i=1}^{m} (W_i(H_i(x(t_i)) - y_i), H_i'(x(t_i))\hat{x}(t_i)) + (W_u(u - u_0), h) + (W_a(a - a_0), k) \tag{2.3}$$

La quantité $\hat{x}$ est solution du modèle tangent linéaire

$$\left\{ \begin{array}{l} \dfrac{d\hat{x}}{dt} = \dfrac{\partial F}{\partial x}(x,a)\hat{x} + \dfrac{\partial F}{\partial a}(x,a)k, \quad 0 \leq t \leq T \\ \hat{x}(0) = h \end{array} \right. \tag{2.4}$$

On est alors amené à introduire la variable adjointe q définie comme solution de l'équation rétrograde du modèle adjoint :

$$\left\{ \begin{array}{l} -\dfrac{dq}{dt} - \left[\dfrac{\partial F}{\partial x}(x,a)\right]^T q = \displaystyle\sum_{i=1}^{m} H_i'^T(x(ti))W_i(H_i(x(t_i)) - y_i)\delta_{t_i}, \quad T \geq t \geq 0 \\ q(T) = 0 \end{array} \right. \tag{2.5}$$

Ceci permet d'exhiber la linéarité en (h,k) de $\hat{J}$ et d'exprimer le gradient de J en fonction de q et x :

$$\nabla J(u,a) = \left(\begin{array}{c} \nabla_u J(u,a) \\ \nabla_a J(u,a) \end{array} \right) = \left(\begin{array}{c} q(0) + W_u(u - u_0) \\ \displaystyle\int_0^T \left[\frac{\partial F}{\partial a}(x,a)\right]^T q dt + W_a(a - a_0) \end{array} \right) \tag{2.6}$$

A chaque itération de l'algorithme de descente le calcul du gradient peut donc s'effectuer après une intégration du modèle direct et une intégration rétrograde du modèle adjoint.

Cette méthodologie est bien établie mais sa mise en pratique reste délicate. Numériquement les calculs de gradient se font en utilisant les codes dérivés tangent linéaire et adjoint du code direct. Ceci permet de s'assurer que le gradient obtenu est bien celui de la fonction coût numérique discrète. Ces codes dérivés peuvent s'obtenir par des outils de différentiation automatique comme le logiciel TAPENADE [61] développé à l'INRIA. Malheureusement les codes directs sont souvent extrêmement complexes et n'ont pas été pensés pour être différentiés ce qui rend souvent la tâche relativement ardue et ingrate.

Contrairement à la météorologie ou à l'océanographie physique, en écologie marine généralement l'identification de la condition initiale n'est pas d'un grand intérêt, les erreurs s'expliquant plus par de mauvaises valeurs des paramètres du modèle. On considère alors une fonction coût $J(a)$ dépendant uniquement des paramètres. Si le nombre N_a de paramètres n'est pas trop élevé, un calcul de gradient effectué à partir du modèle tangent linéaire peut être envisagé car ce calcul est facilement parallélisable, les appels au code tangent étant tous indépendants les uns des autres.

Analyse de sensibilité locale apriori, matrice hessienne et identifiabilité des paramètres

Dans le cadre d'un problème d'estimation de paramètres l'objectif de l'analyse de sensibilité est de déterminer les paramètres dont les variations ont le plus d'impact sur la valeur de la fonction coût. En d'autres termes il s'agit de déterminer les paramètres qui vont ou non pouvoir être déterminés précisément en résolvant un problème inverse numériquement bien posé.

Considérons que le modèle direct est représenté par une application ϕ qui a un jeu de paramètres a associe l'équivalent numérique des observations $\phi(a)$ et que la fonction coût, uniquement constituée du terme d'écart aux observations, s'écrit

$$J(a) = \frac{1}{2}||\phi(a) - y||_W^2$$

Une première approche consiste simplement à calculer et comparer entre elles les composantes normalisées du gradient au point de référence a_0,

$$\frac{\partial J}{\partial a_i}(a_0)\frac{a_{0i}}{J(a_0)}$$

La valeur de ces sensibilités est d'autant plus importante que la fonction coût est sensible aux paramètres correspondants.

Une autre manière de faire est de se placer dans le cas de données simulées à partir du modèle en utilisant le jeu de paramètres de références a_0. Le développement de la fonction coût à l'ordre 2 peut s'écrire

$$J(a_0+h) \approx J(a_0) + (\nabla J(a_0), h) + \frac{1}{2}(W\phi''(a_0)(h,h), (\phi(a_0)-y)) + \frac{1}{2}(\phi'(a_0)^T W \phi'(a_0)h, h)$$

et a_0 étant ici le minimum les 3 premiers termes du membre de droite sont nuls. On a donc l'expression exacte dans ce cas de la matrice hessienne au minimum

$$H = \phi'(a_0)^T W \phi'(a_0)$$

qui peut être facilement calculée à partir du modèle tangent linéaire. La matrice hessienne permet d'obtenir des indications sur la convergence et les incertitudes du problème d'optimisation [82]. Le conditionnement de la matrice (rapport de la plus grande valeur propre sur la plus petite) caractérise le degré de singularité du problème et détermine le taux de convergence de l'algorithme de minimisation. Aux plus petites valeurs propres correspondent une grande incertitude dans l'identification des paramètres correspondants aux composantes principales des vecteurs propres associés. Ainsi l'étude de la matrice hessienne permet de détecter les paramètres les plus difficilement identifiables.

2.2 Modélisation

Dans la section précédente j'ai rappelé la méthodologie de résolution d'un problème d'estimation de paramètres par assimilation variationnelle de données. Au cours de ma thèse j'avais déjà mis au point ce type de méthode pour un modèle de biogéochimie océanique. Il s'agissait d'un modèle NNPZD-DOM (pour nitrate $NO3$, ammonium $NH4$, phytoplancton P, zooplancton Z, détritus D, matière organique dissoute DOM) maintenant devenu le modèle LOBSTER [67] qui nous avait été fourni par Marina Lévy (CNRS) et Laurent Mémery (CNRS). C'était un système couplé d'équations d'advection-diffusion-réaction. L'estimation de paramètres avait entre autre permis de montrer que si l'on donnait à l'un des paramètres du modèle, le rapport Carbone-Chlorophylle, la possibilité de varier avec la profondeur on réduisait fortement l'écart aux données. Par la suite avec Olivier Bernard nous avions prolongé ce travail en complexifiant le modèle pour lui donner une représentation mécaniste de l'évolution du rapport Carbone-Chlorophylle. Cela a donné lieu à l'article [13] et c'était pour moi le premier travail d'écriture d'un modèle destiné à être confronté à des données réelles. Par la suite à l'IRD j'ai participé à plusieurs travaux de modélisation pour les écosystèmes marins résumés ci-dessous.

Les écosystèmes marins sont soumis aujourd'hui à deux perturbations importantes : une pêche croissante d'une part et des variations climatiques d'autre part. Dans ce contexte la mise au point d'outils numériques d'évaluation et de prévision fiables des stocks halieutiques revêt une importance particulière. Les logiciels de référence pour l'évaluation des stocks, comme MULTIFAN-CL [51], sont basés sur un modèle discret de population structurée en classes d'âge que l'on appelle les équations de capture (catch equations).

A partir de ce modèle et des données classiques en halieutique :

— données de capture (masse cumulée de poissons pêchés)

— données de fréquences de taille (taille des poissons échantillonnées dans les captures)

il est possible d'estimer un certain nombre de quantités importantes pour la gestion des pêches comme la mortalité, le taux de croissance ou la biomasse totale.

Pour ce qui concerne les pêcheries de thons tropicaux, espèces cibles de l'UR THETIS à l'IRD, ce type de modèle présente un certain nombre d'insuffisances. Un premier point est que ces pêcheries

sont très hétérogènes en espace et en temps et d'importantes migrations d'individus ont lieu à différentes échelles. De plus la croissance des individus est également potentiellement variable en espace et temps. Des poissons du même âge peuvent avoir des tailles sensiblement différentes selon leur histoire. Ainsi il parait important de représenter explicitement les mouvements et la variabilité de la croissance à l'aide de modèles spatialisés. Un second point est que les données de fréquences de taille ne sont pas immédiatement utilisable avec des modèles structurés en âge. Une relation âge-taille doit être utilisée afin de pouvoir prendre en compte ces données. Malheureusement cela peut engendrer des biais dans les estimations des taux de croissance et de mortalité. Il semble donc nécessaire de proposer des modèles structurés en taille. C'est ce que nous avons fait en proposant une modélisation mécaniste de la dynamique de populations de thons incorporant un maximum de connaissances écologiques et physiologiques.

Le travail de modélisation effectué s'est fait en plusieurs étapes retracées ci-dessous.

2.2.1 Modélisation pour les pêcheries de thons tropicaux

- ▷ **Articles disponibles dans la partie recueil d'articles** : Article G [15], Article H [16], Article I [11]
- ▷ **Autres publications associées** : [14, 17]
- ▷ **Collaborateurs** : Olivier Maury (IRD), Sibylle Dueri (IRD)

Un premier modèle [17, 14]

Le projet originel lorsque je suis arrivé à l'IRD était de mettre au point une méthode d'assimilation variationnelle de données pour identifier les paramètres d'un modèle qu'Olivier Maury (IRD) avait commencé à développer [17]. Dans ce modèle l'océan (l'espace 2D) est découpé en grandes régions et les équations décrivent la dynamique d'une population structurée en âge et en taille et pour laquelle les mouvements d'une région de l'espace à une autre sont représentés par des taux de migration. La population est représentée par des densités de nombre d'individus $p_i(t, a, s)$ mesurant pour chaque région i le nombre d'individus par classe d'âge a et de taille s. Ces densités sont solutions d'un système d'équations d'advection-diffusion-réaction. Un terme de diffusion en taille permet de rendre compte de la variabilité des tailles dans une même classe d'âge. Une non-linéarité non-locale apparait dans la condition limite qui repésente les entrées d'individus dans le modèle et qui est formulée à l'aide d'une relation stock-recrutement de Beverton et Holt [42]. La particularité de ce système réside dans cette non-linéarité et dans le fait que les équations sont constituées d'une partie hyperbolique en âge et parabolique en taille. Dans [14] nous introduisons une formulation variationnelle pour ce modèle et prouvons l'existence, l'unicité et la positivité d'une solution faible. La non-linéarité est traitée par une méthode de point fixe. Nous prouvons également un résultat de comparaison : si la mortalité par pêche augmente dans une région alors la population décroit globalement.

Le travail sur l'assimilation de données de pêche dans ce modèle n'a jamais été finalisé pour deux raisons principales. La première est que ce modèle permet difficilement de rendre compte du forçage océanique sur la dynamique de la population et assez rapidement j'ai été tenté par une approche continue de la représentation de l'espace. La deuxième, moins noble, est que le code avait commencé à être développé avec le logiciel Automatic Differentiation Model Builder (ADMB [52]). Ce logiciel est une surcouche de C++ permettant de faire de la différentiation automatique de code de manière transparente par surcharge d'opérateurs. Ce genre d'approche est très performante pour des modèles de petite dimension. Malheureusement, elle ne supporte que difficilement les grandes dimensions, particulièrement pour le mode adjoint. Le problème s'est posé pour ce modèle, et faire de l'assimilation variationnelle de données avec ADMB paraissait difficile. Ainsi nous nous sommes lancés dans l'écriture d'un nouveau modèle, dans lequel l'espace est représenté de manière continue. La stratégie informatique suivie dans un second temps a été d'écrire le code en Fortran et de bien penser l'organisation du programme de manière à pouvoir utiliser directement le logiciel

de différentiation automatique "source to source" Tapenade [61]. Cette approche "source to source" permet plus facilement de retravailler le code différentié en mode adjoint afin de gérer les problèmes de mémoire.

Un deuxième modèle : APECOSM-E - version 1 [15]

La dynamique de population de poissons est maintenant décrite au travers d'une fonction densité du nombre d'individus $p(x, y, s, t)$, où la position $(x, y) \in \Omega$ le domaine representant l'océan, s représente la taille et t le temps. La densité de population obéit à une équation d'advection-diffusion-réaction en espace et en taille

Sans rentrer dans les détails donnés dans [15] la paramètrisation des mouvements repose sur une séparation en composante physique (les courants calculés par un modèle de circulation océanique) et composante biologique de la vitesse et de la diffusion. Les composantes biologiques dépendent d'une fonction donnée $h(x, y, t)$ (pour habitat suitability index) mesurant la qualité de l'habitat et calculée à partir de la température $T(x, y, t)$ provenant du même modèle de circulation océanique et d'une fonction de fourrage $F(x, y, t)$ (représentant les proies) qui peut soit être extrapolée à partir de la variable zooplancton d'un modèle de biogéochimie marine soit provenir du modèle de flux d'énergie présenté à la section suivante. Le terme de recrutement présente le même type de non-linéarité que dans le modèle précédant. D'autre part avec cette modélisation le taux de croissance est une simple fonction de la taille et ne présente pas de variabilité en espace et en temps. Ainsi comme dans le premier modèle nous introduisons un terme de diffusion en taille. Du point de vue de mathématique dans [15] on introduit une formulation variationnelle et on prouve l'existence d'une unique solution positive.

Ce modèle, pas encore tout à fait abouti au niveau de la représentation des différents processus n'a jamais été confronté à des données réelles contrairement à la version du paragraphe suivant [11, 12]. Il a par contre servi de base pour le développement du code et permis de tester la résolution numérique du problème inverse d'estimation de paramètres grâce à des expériences d'assimilation de données synthétiques qui sont également présentées dans [15].

La dernière version du modèle : APECOSM-E - version 2 [11]

Dans cette dernière version la modélisation est enrichie sur plusieurs aspects :

Les procéssus physiologiques Tous les processus, en particulier la croissance, la mortalité naturelle et la reproduction, varient en temps, en espace, et en taille. Ils sont fonction de l'environnement (courants, température, oxygène dissous et fourrage). Cela permet de s'affranchir du terme de diffusion en taille et de la non-linéarité dans le terme de recrutement qui n'a plus à être saturé par une relation du type Berverton et Holt. La variable de taille structurant la population n'est plus la longueur mais le volume structural des individus. Cela permet d'utiliser plus facilement qu'avec la longueur la théorie DEB (Dynamic Energy Budget [64]) pour la paramétrisation des processus de reproduction, de croissance et de mortalité.

La troisième dimension d'espace z i.e la profondeur a été rajoutée. Cela permet de prendre en compte les mouvements verticaux ainsi que la selectivité sur la profondeur de la mortalité par pêche. Ainsi la population est représentée par une densité $p(x, y, z, v, t)$ où la position $(x, y, z) \in \mathcal{D} = \Omega \times (0, Z)$ est le domaine representant l'océan, la taille ou volume structural $v \in (V_0, V_1)$ avec V_0 le volume structural à la naissance et $t \in (0, T)$.

Comme précédemment la densité de population obéit à un processus d'advection-diffusion en espace. Notons $\mathbf{v}$ le champ de vitesse horizontale, v_z la vitesse verticale, $\mathbf{D}$ la matrice de diffusion horizontale et d_z la diffusion verticale. La croissance des individus est représentée par un processus d'advection en v avec un taux de croissance g. On note m et f les taux de mortalité naturelle et

par pêche. La densité p satisfait :

$$\partial_t p = \mathrm{div}(\mathbf{D}\nabla p - \mathbf{v}p) + \partial_z(d_z\partial_z p - v_z p) - \partial_v(gp) - (m+f)p, \quad \text{sur } \mathcal{D}\times(V_0,V_1)\times(0,T), \quad (2.7)$$

où ∇ et div sont les opérateurs différentiels sur Ω. Cette équation est complétée par des conditions initiales, des conditions aux limites de Neumann homogène en espace et de la condition de Dirichlet non-locale représentant la reproduction

$$gp(x,y,z,V_0,t) = \int_{V_0}^{V_1} bp dv, \quad \forall (x,y,z,t)\in\mathcal{D}\times(0,T), \quad (2.8)$$

Mouvements horizontaux A nouveau la représentation des mouvements repose sur une séparation en composante physique (les courants) et biologique. La paramétrisation de la composante biologique fait l'objet d'une publication [16]. Dans ce travail on commence par considérer une description des mouvements 2D à l'échelle des individus. On formule une modélisation type "random walk" dans laquelle la vitesse de chaque individu a une composante deterministe et une composante stochastique. Toutes deux dépendent d'une fonction de favorabilité de l'habitat h et de son gradient. A l'aide de développements de Taylor en espace et en temps, combinés et tronqués on obtient une équation d'advection-diffusion approchant le modèle originel. Le procédé d'approximation permet d'obtenir explicitement l'expression des coefficients de diffusion et d'advection. Ce sont eux qui sont utilisés dans le modèle APECOSM-E. L'advection et la diffusion sont liées : une forte advection va avec une faible diffusion donnant un mouvement dirigé des individus vers un habitat plus favorable, au contraire une faible advection va avec une forte diffusion ce qui correspond à un comportement de recherche de nourriture. Dans [16] des expériences numériques sont également réalisées et montrent que l'équation aux dérivées partielles est une bonne approximation du modèle individu centré.

Mouvements verticaux et réduction du modèle On peut faire l'hypothèse que les mouvements verticaux sont des processus beaucoup plus rapides que les mouvements horizontaux, la reproduction, la croissance et la mortalité. Ceci permet de réduire la dimension verticale du modèle et de passer d'une équation posée en 4 dimensions d'espace à 3 dimensions. En adimensionnalisant le modèle, des dynamiques rapides et lentes caractérisées par un petit paramètre ε apparaissent. Le modèle réduit est déduit en prenant la limite $\varepsilon = 0$. Dans cette limite le terme d'advection-diffusion en z doit être nul. Ceci donne une équation différentielle en z qui peut s'intégrer analytiquement. On peut alors définir pour chaque processus une moyenne selon un profil vertical et se ramener à un modèle réduit approché posé sur Ω. Ceci est détaillé dans une annexe du papier [11] et c'est ce modèle réduit qui est utilisé dans les simulations. La figure 2.1 donne une représentation schématique du modèle.

2.2.2 Modélisation du flux d'énergie dans les écosystèmes marins

- ▷ **Article disponible dans la partie recueil d'articles** : Article K [18]
- ▷ **Publication associée** : [19]
- ▷ **Collaborateurs** : Olivier Maury (IRD), Yunne Shin (IRD), Francis Marsac (IRD), Tamara Ben Ari (doctorante IRD à l'époque), Jean-Christophe Poggiale (Université de Marseille)

Dans ce travail nous proposons un modèle original du flux d'énergie dans les écosystèmes marins. Le modèle décrit l'évolution d'une variable d'état $\mathcal{E}(w,t)$ représentant la densité d'énergie par classe de taille (w la masse des individus est reliée à la taille par une fonction allométrique $w = as^3$) et par unité de volume d'eau. Les flux d'énergie dans les écosystèmes marins sont contrôlés par la prédation et le modèle se focalise particulièrement sur les organismes dits consommateurs

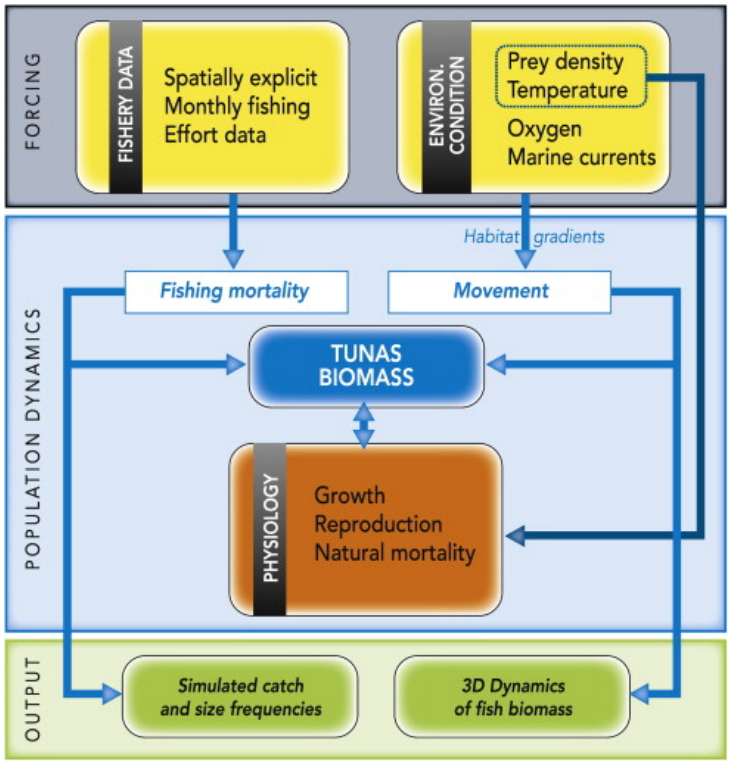

FIGURE 2.1 – Schéma de synthèse présentant les forçages, les processus affectant la dynamique de population et les sorties du modèle APECOSM-E

(organismes hétérotrophes comprenant un grand nombre de groupes de zooplancton et les poissons) qui gagnent de l'énergie uniquement par prédation. Ces organismes se reproduisent et leurs oeufs sont supposés avoir une masse w_0.

Les processus modélisés sont la prédation, la mortalité, l'assimilation et l'utilisation de l'énergie pour la maintenance, la croissance et la reproduction. L'équation utilisée pour modéliser les flux d'énergie à travers les tailles des organismes fait intervenir un terme de transport représentant la croissance et trois termes de mortalité pour la mortalité par prédation, la mortalité due à la famine et la mortalité due à d'autres causes. Le modèle s'écrit alors

$$\begin{cases} \partial_t \mathcal{E} = -\partial_w(g\mathcal{E}) - (\lambda + m + z)\mathcal{E}, \quad \text{sur } (w_0, w_\infty) \\ g\mathcal{E}(t, w_0) = r(\mathcal{E}) \\ \mathcal{E}(0, w) = \mathcal{E}^0(w) \end{cases}$$

A nouveau toute la sophistication du modèle vient des paramétrisations utilisées pour les taux de croissance g, de mortalité λ, m et z, et pour la reproduction r. Je ne les détaille pas ici et réfère à notre papier [18]. Tous ces coefficients sont des fonctions non-linéaires et non-locales de $\mathcal{E}$. Ils sont obtenus en suivant le principe de conservation de l'énergie. La prédation correspond à une perte d'énergie pour les proies et à un gain pour les prédateurs. Elle n'est contrôlée que par une fonction de sélectivité dépendant du rapport entre la taille du prédateur et de la proie. Ainsi tous les organismes peuvent être à la fois proie et prédateur. La prédation est supposée être opportuniste et les proies d'une certaine taille sont mangées proportionnellement au rapport de leur biomasse sur la biomasse de toutes les proies possibles.

Une partie de l'énergie ingérée est utilisée pour la croissance et la maintenance et l'autre pour la reproduction et la maintenance des gonades. La croissance et la reproduction ne peuvent être négatives ainsi quand le coût énergétique de la maintenance est plus important que l'énergie assimilée, la croissance et la reproduction s'annulent et la mortalité due à la famine devient au contraire active.

Enfin les différents taux de croissance, de mortalité et de reproduction sont également modulés par un facteur de correction dépendant de la température.

Les organismes plus petits ($0 < w < w_0$), producteurs primaires (organismes autotrophes composés majoritairement de phytoplancton) qui convertissent l'énergie solaire et les nutriments minéraux en biomasse, ne sont pas modélisés très finement dans ce travail. Néanmoins la densité d'énergie qu'ils représentent intervient dans les calculs de prédation.

De nombreuses simulations numériques ont été effectuées avec ce modèle [19] pour un intervalle de taille allant de $1mm$ à $2m$. Avec des conditions environnementales stables (production primaire et température) la solution évolue vers un état stationnaire correspondant à un spectre de taille log-log linéaire. Ce spectre de taille est peu sensible aux valeurs des paramètres du modèle. Une version spatialisée du modèle existe aujourd'hui [71] et permet de fournir des champs de fourage au modèle APECOSM-E.

2.3 Identification de paramètres par assimilation variationnelle de données de pêche

- ▷ **Articles disponibles dans la partie recueil d'articles** : Article J [12], Article I [11] Article G [15]
- ▷ **Collaborateurs** : Olivier Maury (IRD), Sibylle Dueri (IRD)

Le modèle numérique direct

Nous disposons maintenant d'un modèle pour la dynamique de population de thons dans l'océan décrit à la section 2.2.1.

$$\begin{cases} \partial_t p = \mathrm{div}(\mathbf{D}\nabla p - \mathbf{v}p) - \partial_v(gp) - (m+f)p, \quad \text{sur } \Omega \times (V_0, V_1) \times (0,T), \\ p(x,y,v,0) = p^0(x,y,v), \quad \forall (x,y,v) \in \Omega \times (V_0, V_1), \\ gp(x,y,V_0,t) = \displaystyle\int_{V_0}^{V_1} bp dv, \quad \forall (x,y,t) \in \Omega \times (0,T), \\ \nabla p(x,y,v,t) \cdot n(x,y) = 0, \quad \text{sur } \partial\Omega, \ \forall (v,t) \in (V_0, V_1) \times (0,T), \end{cases} \tag{2.9}$$

dans lequel les coefficients $\mathbf{D}$, $\mathbf{v}$, g, m, f et b sont des fonctions de (x,y,v,t) et d'un vecteur de paramètres $a \in \mathbb{R}^{N_a}$.

Le modèle est discrétisé par une méthode classique de différences finies. Une grille de 1^o par 1^o est utilisée et couvre l'Océan Indien. Les tailles des organismes considérés dans le modèle vont de $1mm$ à $1m$. Les simulations sont faites avec un pas de temps journalier pour la période 1958-2001. L'exploitation des pêcheries industrielles débute en 1984 et les 15 premières années de la simulation représentent une phase de "spin-up" du modèle. Les conditions environnementales déterminant l'habitat des thons sont fournies par des champs 3D de température, oxygène, mésozooplancton et courants marins générés par le modèle couplé physique-biogéochimie NEMO-PISCES [40]. Le modèle nécessite également des données d'effort de pêche. Pour simplifier je ne considère ici qu'une seule flotille mais dans [11, 12] 4 flotilles différentes sont considérées et les données d'effort de pêche sont issues de la base de données de l'IOTC (Indian Ocean Tuna Commission). Les flotilles considérées représentent les pêcheries principales de thons Listao de l'Océan Indien pour lesquelles on dispose de séries temporelles de données de pêche depuis 1984. L'effort de pêche est constant sur des domaines espace-temps $\Omega_k \times [t_d^l, t_f^l]$ pour $k = 1 \dots N_k$ et $l = 1 \dots N_l$. Dans la suite pour simplifier les notations on n'utilise qu'un seul indice pour repérer ces domaines $D_i = \Omega_k \times [t_d^l, t_f^l]$ avec $i = k + (l-1) \times N_k$.

Le modèle contient 48 paramètres : 18 sont associés à la mortalité par pêche et définissent les sélectivités en taille et profondeur du matériel de pêche, la capturabilité et l'accroissement de la puissance de pêche dû aux développements technologiques, 8 sont des paramètres DEB décrivant la croissance, la reproduction et la mortalité naturelle, et enfin 22 sont des paramètres écologiques décrivant les interactions entre environnement et population. Un jeu de référence pour ces paramètres est obtenu à partir de la littérature et d'un réglage manuel du modèle.

Dans [11] un certain nombre de résultats numériques sont présentés permettant de tester la capacité du modèle à représenter la dynamique spatiale de la biomasse de thons avec des conditions environnementales variables, la variabilité de la croissance ainsi que pour tester l'impact de l'exploitation industrielle des pêcheries sur la population. La Fig. 2.2 donne un exemple de sorties du modèle (voir [11] pour plus de détails).

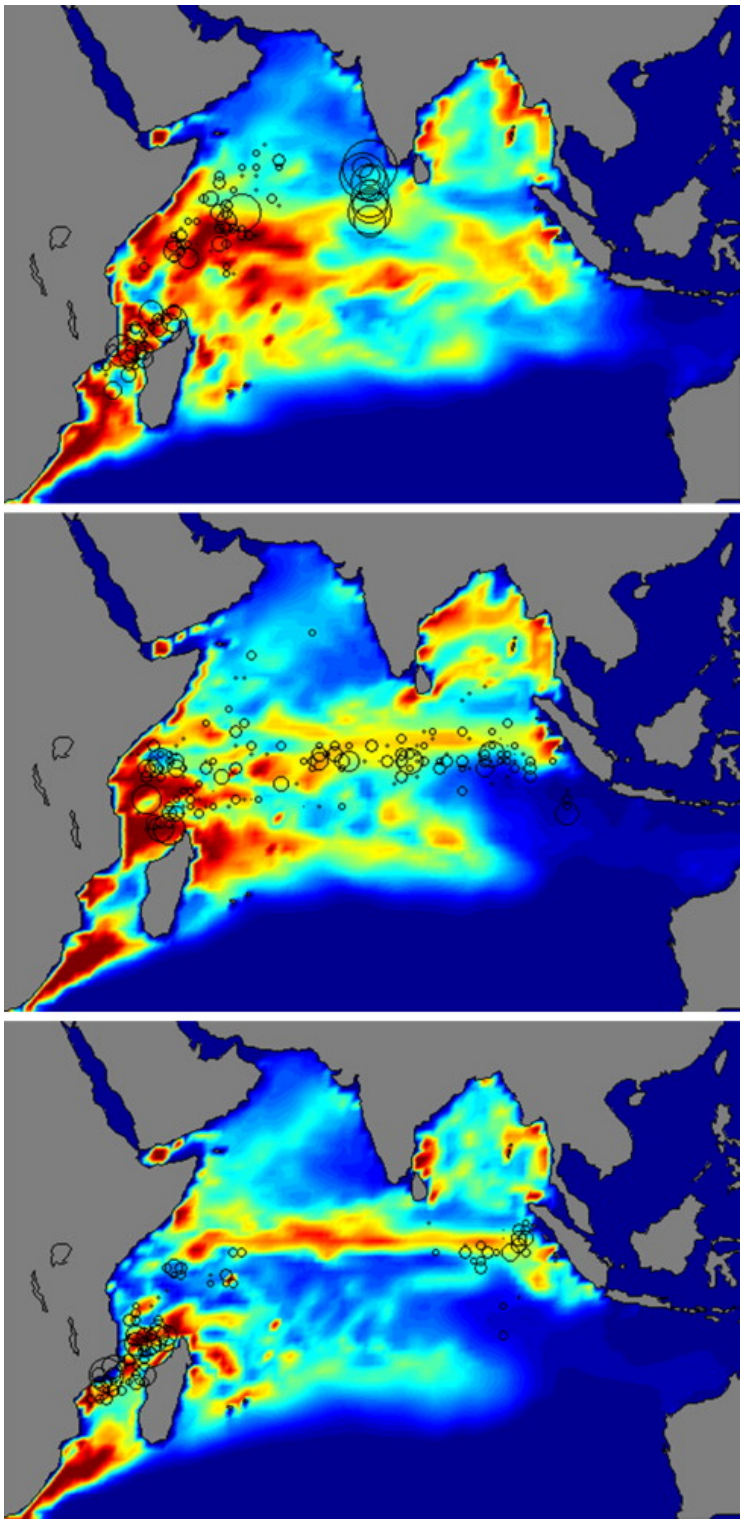

FIGURE 2.2 – Population exploitable de thons Listao calculée versus captures observées (cercles) dans l'océan Indien. Avril 1993, fevrier 1998 et avril 1998

Le problème inverse

Nous disposons de données de pêche (captures et fréquences de tailles). L'équivalent des m données de capture C_i^0 est calculé dans le modèle comme

$$C_i = \int_0^T \int_{V_0}^{V_1} \int_\Omega \mathbf{1}_{D_i} f p w(v) dx dy dv dt$$

où $w(v)$ est la masse des individus de taille v et $\mathbf{1}_{D_i}$ est la fonction caractéristique des domaines espace-temps D_i de sommation des captures.

En définissant

$$C_i(v) = \int_0^T \int_\Omega \mathbf{1}_{D_i} f p w(v) dx dy dt$$

l'équivalent des données de fréquences de taille $Q_i^0(v)$ est calculé comme

$$Q_i(v) = C_i(v)/C_i$$

Le problème inverse d'estimation de paramètre peut être formulé comme un problème de minimisation pour la fonction coût

$$\begin{aligned} J(a) &= \sum_{i=1}^m \frac{1}{2\sigma_C^2}(C_i - C_i^0)^2 + \sum_{i=1}^m \frac{1}{2\sigma_Q^2} \int_{V_0}^{V_1} (Q_i - Q_i^0)^2 dv + \sum_{i=1}^{N_a} \frac{1}{2\sigma_i^2}(a_i - a_i^0)^2 \\ &= J_C(a) + J_Q(a) + J_r(a) \end{aligned}$$

Les termes J_C et J_Q sont les termes d'écart aux observations et J_r est un terme de pénalisation. Son rôle est double, il permet d'une part d'inclure de l'information apriori avec le jeu de paramètres de référence a^0 et les variances σ_i et il joue d'autre part un rôle régularisant.

La minimisation de J se fait par un algorithme de descente nécessitant le calcul du gradient. Ce calcul se fait en introduisant le modèle adjoint. Pour simplifier la présentation on considère que J ne contient qu'un terme d'écart aux données de capture

$$J(a) = \frac{1}{2} \sum_{i=1}^m (C_i - C_i^0)^2$$

Je donne ci-dessous un exemple de calcul du gradient par le modèle adjoint. La dérivée de J au point a dans la direction h s'écrit

$$\hat{J} = \sum_{i=1}^m (C_i - C_i^0)\hat{C}_i = \sum_{i=1}^m (C_i - C_i^0) \int_0^T \int_{V_0}^{V_1} \int_\Omega \mathbf{1}_{D_i} (f\hat{p} + \hat{f}p) w(v) dx dy dv dt$$

La quantité $\hat{p}$ est solution du modèle tangent linéaire :

$$\begin{cases} \partial_t \hat{p} = \mathrm{div}(\mathbf{D}\nabla\hat{p} - \mathbf{v}\hat{p}) - \partial_v(g\hat{p}) - (m+f)\hat{p} + \mathrm{div}(\hat{\mathbf{D}}\nabla p - \hat{\mathbf{v}}p) - \partial_v(\hat{g}p) - (\hat{m}+\hat{f})p, \\ \hat{p}(x,y,v,0) = 0, \\ g\hat{p}(x,y,V_0,t) = \displaystyle\int_{V_0}^{V_1} (\hat{b}p + b\hat{p}) dv - \hat{g}p(x,y,V_0,t), \\ \nabla\hat{p}(x,y,v,t) \cdot n(x,y) = 0 \end{cases} \tag{2.10}$$

On multiplie alors le modèle tangent linéaire par la variable adjointe q. Après une intégration par partie et des manipulations permettant d'exhiber la linéarité en h de $\hat{J}$ on est amené à définir q comme solution de l'équation rétrograde du modèle adjoint :

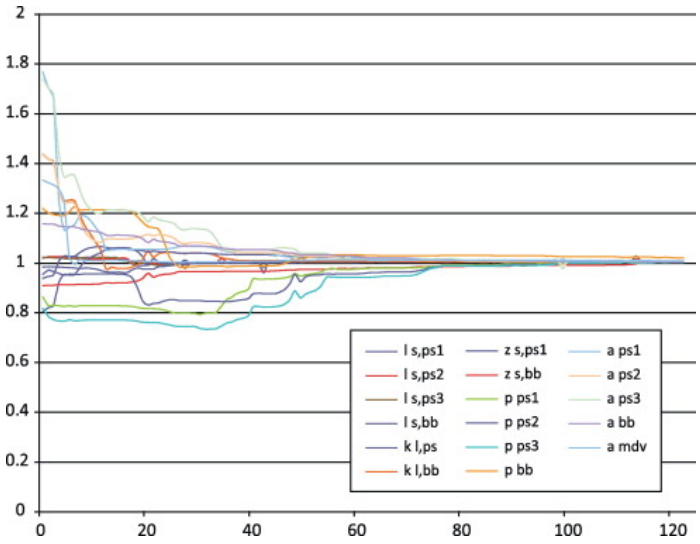

FIGURE 2.3 – Variations des paramètres en fonction des itérations de minimisation. Les valeurs sont normalisées par les valeurs optimales.

$$\begin{cases} -\partial_t q - \mathrm{div}(\mathbf{D}\nabla q) - \mathbf{v}\nabla q - g\partial_v q + (m+f)q - bq(V_0) = \sum_{i=1}^{m}(C_i - C_i^0)\mathbf{1}_{D_i} f w, \\ q(x,y,v,T) = 0, \\ q(x,y,V_1,t) = 0, \\ \nabla q(x,y,v,t)\cdot n(x,y) = 0 \end{cases} \tag{2.11}$$

et l'on peut conclure que le gradient de J s'exprime en fonction de p et q :

$$\begin{aligned} \nabla J(a) \quad &= \int_0^T \int_{V_0}^{V_1} \int_\Omega [\sum_{i=1}^{m}((C_i - C_i^0 - q)p\mathbf{1}_{D_i}(\nabla_a f))w \\ &+(\mathrm{div}((\nabla_a \mathbf{D})\nabla p - (\nabla_a \mathbf{v})p) - \partial_v((\nabla_a g)p) - (\nabla_a m)p)q \\ &+(\nabla_a b - \frac{b}{g(V_0)}\nabla_a g(V_0))pq(V_0)]dxdydvdt \end{aligned} \tag{2.12}$$

où ∇_a indique la dérivation par rapport aux paramètres a. Le calcul du gradient peut donc s'effectuer après une intégration du modèle direct et une intégration rétrograde du modèle adjoint. La formule (2.12) déjà complexe dans ce cas simplifié a peu d'intérêt si ce n'est de mettre en avant le fait que pour un tel modèle il est quasiment impossible de faire tous les calculs à la main sans erreur et donc que les logiciels de différentiation automatique comme TAPENADE sont essentiels dans ce type de problème.

L'optimisation ne faisant intervenir que des données de pêche dans [12] nous n'incluons dans la fonction coût que 19 paramètres directement liés à la mortalité par pêche. L'analyse de la matrice hessienne indique que principalement 2 de ces paramètres sont liés aux 2 valeurs propres les plus petites. Il ne sont donc pas optimisés. La minimisation porte ainsi sur 17 paramètres (Fig. 2.3) et est effectuée à l'aide de l'algorithme de quasi-Newton implémenté dans le code M1QN3 [53] .

Je renvoie à l'article [12] pour une analyse détaillée des résultats (Fig. 2.4 et 2.5) et conclus simplement en exprimant le fait que cet exercice d'assimilation de données réelles de pêche a permis de mettre en évidence que d'une part ce type de modèle, nouveau, donne des résultats

tout à fait satisfaisants et que les données de pêche permettent d'en optimiser certains paramètres mais que d'autre part certains phénomènes en particulier l'aggrégation des thons sous les FAD (fish aggregation devices) sont visibles dans les données mais pour l'instant ne sont pas modélisés dans les équations du modèle. Malheureusement ce phénomène d'attraction n'est pas encore bien compris par les spécialistes.

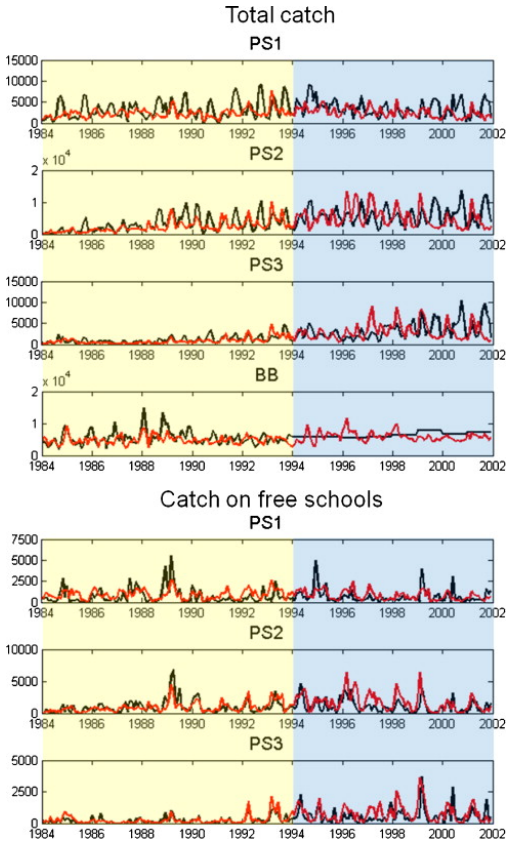

FIGURE 2.4 – Comparaison entre les captures aggrégées par mois simulées (rouge) et observées (noir) pour les différentes flotilles : PS1, PS2 et PS3 les senneurs (purse seiners), et BB les canneurs (bait boats). En haut pour les captures totales et en bas pacture sur les bancs libres uniquement. L'optimisation est réalisée à partir des données de captures totales de 1984 à 1993.

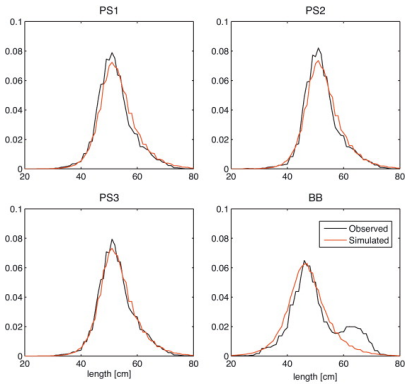

FIGURE 2.5 – Comparaison entre fréquences de tailles simulées et observées pour les 4 flotilles sur la période 1984-1993

2.4 Travaux isolés

2.4.1 Un schéma numérique précis pour l'intégration en temps d'un système d'équations de diffusion - dissolution / précipitation

- ▷ **Article disponible dans la partie recueil d'articles** : Article L [21]
- ▷ **Collaborateurs** : Jérome Pousin (INSA de Lyon), Franck Fontvielle (doctorant à l'INSA de Lyon à l'époque)

L'objectif de ce travail est de mettre au point un schéma numérique précis (c'est-à-dire ici d'ordre 2) et numériquement peu coûteux pour l'intégration temporelle de systèmes d'équations de diffusion-dissolution/précipitation du type :

$$\partial_t C = \Delta C + f(S,C), \tag{2.13}$$
$$\partial_t S = -f(S,C) \tag{2.14}$$

où f est une fonction non linéaire vérifiant,

$$\begin{aligned} f(S,C) &= f_1(S,C) \quad \text{si } g(S) > 0, \\ &= f_2(S,C) \quad \text{si } g(S) \leq 0. \end{aligned}$$

Ces équations modélisent le stockage de déchets dans des matrices en béton [69] et sont dans cette référence intégrées par un schéma peu précis. La résolution numérique de tels systèmes présente deux types de difficultés.
La première est celle de l'intégration de l'équation (2.13). Le schéma numérique le plus simple pour résoudre une équation de réaction-diffusion,

$$\partial_t C = \Delta C + f(C),$$

est le schéma d'Euler explicite, mais avec cette méthode le pas de temps Δt est limité par $O(\Delta x^2)$. Pour s'affranchir de cette condition contraignante on peut utiliser le schéma d'Euler implicite, inconditionnellement stable, mais alors un grand système non linéaire doit être résolu à chaque pas de temps. Une méthode de résolution bien adaptée à ce type d'équations est celle du splitting d'opérateur qui consiste à résoudre séparément et successivement les équations,

$$\begin{aligned} \partial_t C &= \Delta C, \\ \partial_t C &= f(C). \end{aligned}$$

La formule du splitting de Strang [81, 70] permet d'obtenir un schéma d'ordre 2 [41].
La deuxième difficulté, moins classique, tient à la forme de la fonction $f(S,C)$. En effet les instants, t_d, auxquels $g(S)$ change de signe ne sont pas connus par avance et doivent être détectés au cours de l'intégration.
Pour résoudre numériquement ce système , nous proposons un schéma construit à partir du splitting de Strang, et d'un algorithme de détection des instants t_d, auxquels $g(S)$ change de signe, basé sur une formule de "dense output" [59]. En effectuant une analyse des erreurs locales on montre que le schéma proposé est d'ordre 2 en temps.

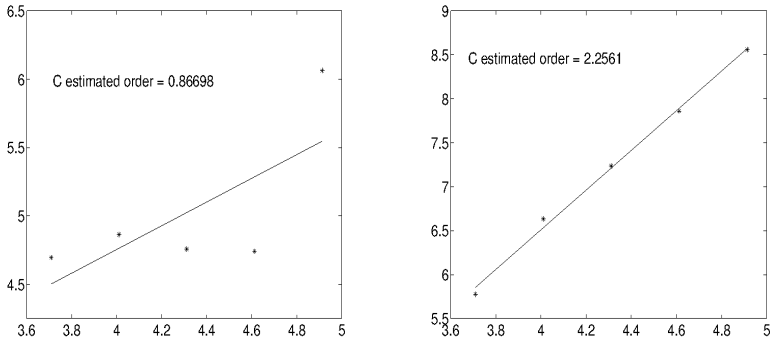

FIGURE 2.6 – $-\log(Erreur)$ versus $-\log(\Delta t)$. Courbes de convergence pour le schéma de splitting classique à gauche, et pour le schéma d'ordre 2 proposé à droite.

Ce résultat, illustré sur un cas test (Figure 2.6), est dû au fait que les instants t_d, auxquels l'expression des termes de réaction $f(S, C)$ change, sont détectés de manière suffisamment précise pour ne pas dégrader l'ordre du schéma de Strang malgré le peu de régularité de f.

2.4.2 Analyse asymptotique d'un modèle élastique 3D pour la segmentation d'images du coeur

- ▷ **Article disponible dans la partie recueil d'articles** : Article M [20]
- ▷ **Collaborateurs** : Jérome Pousin (INSA de Lyon)

Afin d'améliorer les algorithmes de segmentation automatique d'images médicales du coeur (i.e. la détection automatique des contours du coeur dans l'image permettant aux modèles de s'adapter à chaque patient) il a été proposé d'utiliser un modèle élastique du coeur [84, 76, 75]. La stratégie est la suivante : un objet a priori, représentant le coeur, est immergé dans l'image-donnée et est soumis à un champ de forces qui déforme ses frontières vers les contours de l'image.
Notre contribution est la suivante. En partant d'un modèle à trois couches composé d'une couche intérieure homogène et isotrope entourée de deux couches fines de fibres myocardiaques, nous avons obtenu un modèle asymptotique en montrant rigoureusement que lorsque l'épaisseur des fines couches externes tend vers 0, elles peuvent être remplacées par des conditions aux limites particulières sur la couche interne. Ces conditions aux limites ont un effet régularisant et permettent d'améliorer la qualité de la segmentation (Figure 2.7).

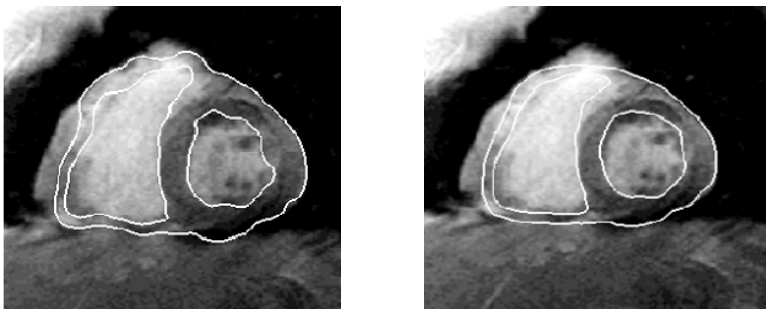

FIGURE 2.7 – Impact de la régularisation sur les résultats de segmentation d'une image transversale du coeur : sans (gauche) et avec (droite) les conditions limites obtenues (d'après [76]).

Pour la démonstration on se place dans un cadre simplifié et l'on considère un modèle de coque mince élastique à deux couches représentant la paroi du coeur (cf. Figure 2.8) : une couche interne, Ω^- , qui suit les lois de l'élasticité linéaire classique pour un matériau homogène et isotrope, et une couche externe, Ω_ε^+, d'épaisseur ε, modélisant un matériau constitué de fibres. Pour ce type de matériau la loi constitutive liant le tenseur des contraintes au tenseur des déplacements fait intervenir explicitement, $\mathbf{d}$, le vecteur 3D d'orientation des fibres.

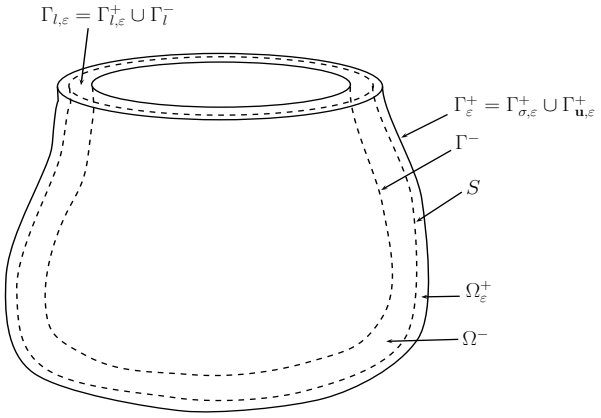

FIGURE 2.8 – Le domaine $\Omega_\varepsilon = \Omega^- \cup \Omega_\varepsilon^+$.

Le champ de déplacement pour chaque point matériel en coordonnées cartésiennes est noté $\mathbf{u}$, $e(\mathbf{u})$ représente le tenseur des déformations de Green-Lagrange linéarisé sous l'hypothèse des petites déformations, et σ représente le tenseur des contraintes. λ et μ, et μ_e sont des constantes, I est le tenseur identité, et le corps élastique est soumis à un champ de forces $\mathbf{f}$.
A l'équilibre on a :

$$\begin{cases} div(\sigma(\mathbf{u})) + \mathbf{f} = 0 & \text{dans } \Omega_\varepsilon, \\ \sigma(\mathbf{u}) = \lambda trace(e(\mathbf{u}))I + 2\mu e(\mathbf{u}) & \text{dans } \Omega^-, \\ \sigma(\mathbf{u}) = (\mathbf{d}.e(\mathbf{u})\mathbf{d}) \mathbf{d} \otimes \mathbf{d} + 2\mu_e \varepsilon e(\mathbf{u}) & \text{dans } \Omega_\varepsilon^+, \\ \mathbf{u} = 0 & \text{sur } \Gamma^- \cup \Gamma_{l,\varepsilon} \cup \Gamma_{\mathbf{u},\varepsilon}^+, \\ \sigma\mathbf{n} = 0 & \text{sur } \Gamma_{\sigma,\varepsilon}^+, \\ \mathbf{u}^- = \mathbf{u}^+ \text{ et } \sigma^-\mathbf{n} = \sigma^+\mathbf{n} & \text{sur } S, \end{cases}$$

où $\mathbf{n}$ est le vecteur normal unitaire.
On montre que lorsque l'épaisseur de la couche externe, ε, tend vers 0 le modèle asymptotique est donné par :

$$\begin{cases} div(\sigma(\mathbf{u})) + \mathbf{f} = 0 & \text{dans } \Omega^-, \\ \sigma(\mathbf{u}) = \lambda trace(e(\mathbf{u}))I + 2\mu e(\mathbf{u}) & \text{dans } \Omega^-, \\ \mathbf{u} = 0 & \text{sur } \Gamma^- \cup \Gamma_l^-, \\ \sigma\mathbf{n} = -2\mu_e u_n \mathbf{n} - \mu_e \mathbf{u}_T & \text{sur } S, \end{cases}$$

où $u_n\mathbf{n}$ est la composante de $\mathbf{u}$ normale à la surface S et $\mathbf{u}_T$ est la composante tangentielle. La condition limite obtenue sur S ne dépend plus de $\mathbf{d}$ le vecteur d'orientation des fibres.
La preuve de ce résultat donnée dans [20] repose sur

- une formulation variationnelle mixte du problème en coordonnées curvilignes généralisées,
- un changement d'échelle permettant de se ramener à un domaine Ω indépendant de ε,
- différentes estimations a priori qui permettent de justifier le passage à la limite lorsque $\varepsilon \to 0$.

2.5 Conclusion

Je ne travaille plus aujourd'hui sur les sujets abordés au chapitre 2. Néanmoins grâce à Google Scholar (et tout de même quelques échanges de mail avec Olivier Maury) je suis de loin le devenir des modèles APECOSM et APECOSM-E que nous avions commencé à développer ensemble. En particulier Olivier et Sibylle Dueri ont continué à utiliser le modèle APECOSM-E dans 2 études récentes. La première concerne l'impact sur la population de thons Listao de l'océan Indien de la création de zones marines protégées [49] et la seconde concerne les conséquences du changement climatique sur cette même population [48]. Un travail sur l'assimilation de données de marquage-recapture qui pourraient permettre de mieux calibrer les paramètres liés aux mouvements et à la croissance est également engagé.

Liste de publications

Publications issues de la thèse

Thèse

[1] B. Faugeras. *Assimilation variationnelle de données dans un modèle couplé océan-biogéochimie.* Thèse de Doctorat. Université Joseph Fourier Grenoble I, 2002.

Articles dans des revues internationales à comité de lecture

[2] B. Faugeras. On the well-posedness of a coupled one-dimensional biological-physical model for the upper ocean. *Mathematical Models and Methods Applied Sciences* 8.13 (2003), p. 1157–1184.

[3] B. Faugeras, M. Lévy, L. Mémery, J. Verron, J. Blum et I. Charpentier. Can biogeochemical fluxes be recovered from nitrate and chlorophyll data? A case study assimilating data in the Northwestern Mediterranean Sea at the JGOFS-DYFAMED station. *J. Mar. Sys.* 40-41 (2003), p. 99–125.

Publications postérieures à la thèse

Articles dans des revues internationales à comité de lecture

[4] J. Blum, C. Boulbe et B. Faugeras. Reconstruction of the equilibrium of the plasma in a Tokamak and identification of the current density profile in real time. *J. Comp. Phys.* 231 (2012), p. 960–980.

[5] G. L. Falchetto, D. Coster, R. Coelho, B.D. Scott, L. Figini, D. Kalupin, E. Nardon, L.L. Alves, J.F. Artaud, V. Basiuk, J. Bizarro, C. Boulbe, A. Dinklage, D. Farina, B. Faugeras, J. Ferreira, A. Figueiredo, P. Huynh, F. Imbeaux, I. Ivanova-Stanik, T. Jonsson, H.-J. Klingshirn, C. Konz, A. Kus, N.B. Marushchenko, E. Nardon, S. Nowak, G. Pereverzev, M. Owsiak, E. Poli, Y. Peysson, R. Reimer, J. Signoret, O. Sauter, R. Stankiewicz, P. Strand, I. Voitsekhovitch, E. Westerhof, T. Zok, W. Zwingmann, ITM-TF contributors, the ASDEX Upgrade Team et JET-EFDA Contributors. The European Integrated Tokamak Modelling (ITM) effort : achievements and first physics results. *Nucl. Fusion* 54.4 (2014), p. 043018. doi : 10.1088/0029-5515/54/4/043018.

[6] B. Faugeras, A. Ben Abda, J. Blum et C. Boulbe. Minimization of an energy error functional to solve a Cauchy problem arising in plasma physics : the reconstruction of the magnetic flux in the vacuum surrounding the plasma in a Tokamak. *ARIMA* 15 (2012), p. 37–60.

[7] B. Faugeras, J. Blum, C. Boulbe, P. Moreau et E. Nardon. 2D interpolation and extrapolation of discrete magnetic measurements with toroidal harmonics for equilibrium reconstruction in a Tokamak. *Plasma Phys. Control Fusion* 56 (2014), p. 114010.

[8] P. Hertout, C. Boulbe, E. Nardon, J. Blum, S. Brémond, J. Bucalossi, B. Faugeras, V. Grandgirard et P. Moreau. The CEDRES++ equilibrium code and its application to ITER, JT-60SA and Tore Supra. *Fusion Engineering and Design* 86.6-8 (2011), p. 1045–1048.

[9] H. Heumann, J. Blum, C. Boulbe, B. Faugeras, G. Selig, J.-M. Ané, S. Brémond, V. Grangirard, P. Hertout et E. Nardon. Quasi-static free-boundary equilibrium of toroidal plasma with CEDRES++ : computational methods and applications. *J. Plasma. Phys.* (2015). doi : 10.1017/S0022377814001251.

[10] A. Murari, J. Vega, D. Mazon, G.A Rattà, J. Svensson, S. Palazzo, G. Vagliasindi, P. Arena, C. Boulbe, B. Faugeras, L. Fortuna, D. Moreau et JET-EFDA Contributors. Innovative signal processing and data analysis methods on JET for control in the perspective of next-step devices. *Nucl. Fusion* 50.5 (2010).

[11] S. Dueri, B. Faugeras et O. Maury. Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E : Part 1. Model formulation. *Ecological Modelling* 245 (2012), p. 41–54.

[12] S. Dueri, B. Faugeras et O. Maury. Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E : Part 2. Parameter estimation and sensitivity analysis. *Ecological Modelling* 245 (2012), p. 55 –64.

[13] B. Faugeras, O. Bernard, A. Sciandra et M. Lévy. A mechanistic modelling and data assimilation approach to estimate the carbon/chlorophyll and carbon/nitrogen ratios in a coupled hydrodynamical-biological model. *Nonlinear Processes in Geophysics* 11.4 (2004), p. 515–533.

[14] B. Faugeras et O. Maury. A multi-region nonlinear age-size structured fish population model. *Nonlinear Analysis : Real World Appl.* 6.3 (2005), p. 447–460.

[15] B. Faugeras et O. Maury. An advection-diffusion-reaction size-structured fish population dynamics model combined with a statistical parameter estimation procedure : Application to the Indian Ocean skipjack tuna fishery. *Math. Biosciences and Engineering* 2.4 (2005), p. 719–741.

[16] B. Faugeras et O. Maury. Modelling fish population movements : from an individual-based representation to an advection-diffusion equation. *J. Theor. Biol.* 247 (2007), p. 837–848.

[17] O. Maury, B. Faugeras et V. Restrepo. FASST : A Fully Age-Size and Space-Time structured statistical model for the assessment of tuna populations. *ICCAT Coll. Vol. Sci. Pap.* 57.1 (2005), p. 206–217.

[18] O. Maury, B. Faugeras, Y-J. Shin, J.C. Poggiale, T. Ben Ari et F. Marsac. Modeling environmental effects on the size-structured energy flow through marine ecosystems. Part 1 : the model. *Progress in Oceanography* 74 (2007), p. 479–499.

[19] O. Maury, Y-J. Shin, B. Faugeras, T. Ben Ari et F. Marsac. Modeling environmental effects on the size-structured energy flow through marine ecosystems. Part 2 : simulations. *Progress in Oceanography* 74 (2007), p. 500–514.

[20] B. Faugeras et J. Pousin. Variational asymptotic derivation of an elastic model arising from the problem of 3D automatic segmentation of cardiac images. *Analysis and Applications (AA)* 2.4 (2004), p. 275–307.

[21] B. Faugeras, J. Pousin et F. Fontvieille. An efficient numerical scheme for precise time integration of a diffusion - dissolution / precipitation chemical system. *Math. of Computation* 75.253 (2006), p. 209–222.

[22] C. Faugeras, B. Faugeras, M. Orlita, M. Potemski, R. R. Nair et A. K. Geim. Thermal conductivity of graphene in Corbino membrane geometry. *ACS NANO* 4 (2010), p. 1889.

Chapitres de livre à comité de lecture

[23] J. Blum, C. Boulbe et B. Faugeras. "Real-time Equilibrium Reconstruction in a Tokamak". In : t. 988. Burning Plasma Diagnostics. Varenna, Italy : AIP Conference Proceedings, 2007, p. 420–429.

[24] D. Mazon, J. Blum, C. Boulbe, B. Faugeras, A. Boboc, M. Brix, P. DeVries, S. Sharapov et L. Zabeo. EQUINOX : A Real-Time Equilibrium Code and its Validation at JET. In : *From physics to control through an emergent view*. Sous la dir. de L. Fortuna, A. Fradkov et M. Frasca. T. 15. World Scientific Book Series On Nonlinear Science, Series B. World Scientific, 2010, p. 327–333.

Actes de congrès internationaux à comité de lecture

[25] J. Blum, C. Boulbe et B. Faugeras. "Real-time plasma equilibrium reconstruction in a Tokamak". In : *Journal of Physics : Conference Series. Proceedings of the 6th International Conference on Inverse Problems in Engineering : Theory and Practice*. T. 135. Dourdan (Paris), France, juin 2008, p. 012019.

[26] B. Faugeras, J. Blum et C. Boulbe. "Equilibrium reconstruction from discrete magnetic measurements in a Tokamak". In : *Proceedings of the 6th International Conference PICOF'12*. Ecole Polytechnique, Paris, France, avr. 2012.

[27] D. Mazon, J. Blum, C. Boulbe, B. Faugeras, A. Boboc, M. Brix, P. De Vries, S. Sharapov et L. Zabeo. "Real-time identification of the current density profile in the JET Tokamak : method and validation". In : *Proceedings of the 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference*. T. WeA09.1. Shanghai, P. R. China, déc. 2009, p. 285–290.

[28] D. Mazon, P. Lotte, B. Faugeras, C. Boulbe, J. Blum, F. Saint-Laurent, S. Bremond, P. Moreau, A. Murari et P. Blanchard. "Validation of the new real-time equilibrium code EQUINOX on JET and ToreSupra". In : *Proceedings of the 39th EPS Conference and 16th Int. Congress on Plasma Physics*. Stockholm, Sweden, juil. 2012.

[29] F. Saint-Laurent, B. Faugeras, C. Boulbe, S. Bremond, P. Moreau et J. Blum. "Plasma position control and current profile reconstruction for tokamaks". In : *ICALEPCS Conference proceedings*. Kobe, Japan, oct. 2009.

[30] J. Urban, L.C. Appel, J.F. Artaud, B. Faugeras, J. Havlicek, M. Komm, I. Lupelli et M. Peterka. "Validation of equilibrium tools on the COMPASS tokamak". In : *SOFT 2014*. San Sebastian, Spain, sept. 2014.

Rapports de recherche

[31] B. Faugeras. *Diffuse interface formulations for region based active contour image segmentation*. Rapport de recherche ISRN I3S/RR-2006-33-FR. Laboratoire I3S, Sophia-Antipolis, France : CNRS, août 2006.

[32] F. Imbeaux, T. Aniel, B. Faugeras, P. Moreau et E. Nardon. *Tokamak-generic equilibrium identification and profile reconstruction*. Rapport CEA. CEA, Cadarache, Fance : CEA-IRFM, 2014.

[33] D. Mazon, J. Blum, C. Boulbe, B. Faugeras, M. Baruzzo, A. Boboc, S. Bremond, M. Brix, P. DeVries, S. Sharapov, L. Zabeo et JET-EFDA Contributors. *EQUINOX : A Real-Time Equilibrium Code and its Validation at JET*. EFDA-JET-CP(09)07/01. JET, Culham Oxford, England : EFDA-JET, 2009.

[34] A. Murari, J. Vega, D. Mazon, G.A Rattà, J. Svensson, G. Vagliasindi, J. Blum, C. Boulbe, B. Faugeras et JET-EFDA Contributors. *New Information Processing Methods for Control in Fusion*. EFDA-JET-CP(09)04/02. JET, Culham Oxford, England : EFDA-JET, 2009.

Articles soumis

[35] B. Faugeras. Tokamak plasma boundary reconstruction using toroidal harmonics and an optimal control method (submitted 2015).

Bibliographie

[36] M. Abramowitz et I.A. Stegun. *Handbook of Mathematical Functions*. Washington, DC : National bureau of Standards, 1964.

[37] R. Albanese, J. Blum et O. Barbieri. "On the solution of the magnetic flux equation in an infinite domain". In : *EPS. 8th Europhysics Conference on Computing in Plasma Physics (1986)*. 1986.

[38] R. Albanese, J. Blum et O. De Barbieri. "Numerical studies of the Next European Torus via the PROTEUS code". In : *12th Conf. on Numerical Simulation of Plasmas, San Francisco*. 1987.

[39] J.F. Artaud, V. Basiuk, F. Imbeaux, M. Schneider, J. Garcia, G. Giruzzi, P. Huynh, T. Aniel, F. Albajar, J.M. Ané, A. Bécoulet, C. Bourdelle, A. Casati, L. Colas, J. Decker, R. Dumont, L.G. Eriksson, X. Garbet, R. Guirlet, P. Hertout, G.T. Hoang, W. Houlberg, G. Huysmans, E. Joffrin, S.H. Kim, F. Köchl, J. Lister, X. Litaudon, P. Maget, R. Masset, B. Pégourié, Y. Peysson, P. Thomas, E. Tsitrone et F. Turco. The CRONOS suite of codes for integrated tokamak modelling. *Nuclear Fusion* 50.4 (2010), p. 043001.

[40] O. Aumont et L. Bopp. Globalizing results from ocean in situ iron fertilization studies. *Global Biogeochemical Cycles* 20.2 (2006), GB2017.

[41] C. Besse, B. Bidegaray et S. Descombes. Order estimates in time of splitting methods for the nonlinear schrödinger equation. *SIAM J. Numer. Anal.* 40.5 (2002), p. 26–40.

[42] R.J.H. Beverton et S.J. Holt. *On the Dynamics of of Exploited Fish Populations*. Fish and Fisheries Series 11. Chapman & Hall, 1996.

[43] J. Blum. *Numerical Simulation and Optimal Control in Plasma Physics with Applications to Tokamaks*. Series in Modern Applied Mathematics. Paris : Wiley Gauthier-Villars, 1989.

[44] J. Blum, J. Le Foll et B. Thooris. The self-consistent equilibrium and diffusion code SCED. *Computer Physics Communications* 24 (1981), p. 235 –254. issn : 0010-4655. doi : 10.1016/0010-4655(81)90149-1.

[45] J. Blum et H. Heumann. *Optimal Control for Quasi-Static Evolution of Plasma Equilibrium in Tokamaks*. Rapp. tech. INRIA Sophia Antipolis, Laboratoire Jean Alexandre Dieudonné - JAD, 2014.

[46] B.J. Braams. *Computational studies in Tokamak equilibrium and transport*. Thèse de doct. University of Utrecht, 1986.

[47] D.P. Coster, V. Basiuk, G. Pereverzev, D. Kalupin, R. Zagórksi, R. Stankiewicz, P. Huynh et F. Imbeaux. The European Transport Solver. *Plasma Science, IEEE Transactions on* 38.9 (2010), p. 2085–2092. issn : 0093-3813. doi : 10.1109/TPS.2010.2056707.

[48] S. Dueri, L. Bopp et O. Maury. Projecting the impacts of climate change on skipjack tuna abundance and spatial distribution. *Global Change Biology* (2014), gcb.12460.

[49] S. Dueri et O. Maury. Modelling the effect of marine protected areas on the population of skipjack tuna in the Indian Ocean. *Aquatic Living Resources* 26 (2013), p. 171–178.

[50] Y. Fischer. *Approximation dans des classes de fonctions analytiques généralisées et résolution de problèmes inverses pour les tokamaks*. Thèse de doct. Université de Nice-Sophia Antipolis, 2011.

[51] D.A. Fournier, J. Hampton et J.R. Sibert. MULTIFAN-CL : a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. *Can. J. Fish. Aquat. Sci.* 55 (1998), p. 2105–2116.

[52] D.A. Fournier, H.J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M.N. Maunder, A. Nielsen et J. Sibert. AD Model Builder : using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optim. Methods Softw.* 27 (2012), p. 233–249.

[53] J. Ch. Gilbert et C. Lemaréchal. Some numerical experiments with variable-storage quasi-Newton algorithms. *Mathematical Programming* 45 (1989), p. 407–435.

[54] H. Grad et J. Hogan. Classical Diffusion in a Tokomak. *Physical review letters* 24.24 (1970), p. 1337–1340.

[55] H. Grad, P. N. Hu et D. C. Stevens. Adiabatic evolution of plasma equilibrium. *Proceedings of the National Academy of Sciences* 72.10 (1975), p. 3789–3793.

[56] H. Grad, P. N Hu, D. C Stevens et E. Turkel. "Classical plasma diffusion". In : *Plasma Physics and Controlled Nuclear Fusion Research 1976, Volume 2*. T. 2. 1977, p. 355–365.

[57] H. Grad et H. Rubin. "Hydromagnetic Equilibria and Force-free Fields". In : *2nd U.N. Conference on the Peaceful uses of Atomic Energy*. T. 31. Geneva, 1958, p. 190–197.

[58] V. Grandgirard. *Modélisation de l'équilibre d'un plasma de tokamak*. Thèse de doct. l'Université de Franche-Comté, 1999.

[59] E. Hairer, S.P. Norsett et G. Wanner. *Solving Ordinary Differential Equations I, Nonstiff Problems*. Springer Series in Computational Mathematics. Springer Verlag, 1993.

[60] C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems : Numerical Aspects of Linear Inversion*. Philadelphia : SIAM, 1998.

[61] L. Hascoët et V. Pascual. The Tapenade Automatic Differentiation tool : Principles, Model, and Specification. *ACM Transactions On Mathematical Software* 39.3 (2013).

[62] *Integrated Tokamak Modelling*. `http://portal.efda-itm.eu/`. 2013.

[63] S.C. Jardin. *Computational methods in plasma physics*. Computational Science Series. Chapman et Hall, 2010.

[64] S.A.L.M Kooijman. *Dynamic Energy and Mass Budgets in Biological Systems*. Cambridge University Press, 2000.

[65] K. Kurihara. A new shape reproduction method based on the Cauchy-condition surface for real-time tokamak reactor control. *Fus. Engin. Des.* 51-52 (2000), p. 1049–1057.

[66] F.X. Le Dimet et O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations : theoretical aspects. *Tellus* 38A (1986), p. 97–110.

[67] M. Lévy, A.-S. Kremeur et L. Mémery. *Description of the LOBSTER biogeochemical model implemented in OPA8*. Rapp. tech. LOCEAN - IPSL, 2005.

[68] J.L. Lions. *Optimal control of system governed by partial differential equations*. Springer, Berlin Heildelberg New York, 1971.

[69] E. Maisse. *Analyse et simulations numériques de phénomènes de diffusion - dissolution / précipitation en milieux poreux, appliquées au stockage de déchets*. Thèse de doct. Université Claude Bernard Lyon I, 1998.

[70] G.I. Marchuk. "Splitting and alternating direction methods". In : *Handbook of numerical analysis*. T. I. North-Holland, Amsterdam, 1990, p. 197–462.

[71] O. Maury. An overview of APECOSM, a spatialized mass balanced "Apex Predators ECOsystem Model" to study physiologically structured tuna population dynamics in their ecosystem. *Prog. Oceanog* 84 (2010), p. 113–117.

[72] R. Nouilletas, E. Nardon et S. Brémond. "Robust vertical plasma stabilization of the future tungsten divertor configuration of Tore Supra". In : *19th IFAC World Congress*. Cape Town, South Africa, 2014.

[73] D.P. O'Brien, J.J Ellis et J. Lingertat. Local expansion method for fast plasma boundary identification in JET. *Nuclear Fusion* 33.3 (1993), p. 467–474.

[74] F. P. Orsitto, A. Boboc, P. Gaudio, M. Gelfusa, E. Giovannozzi, C. Mazzotta, A. Murari et JET EFDA Contributors. Analysis of Faraday rotation in JET polarimetric measurements. *Plasma Phys. Control. Fusion* 53 (2011), p. 035001.

[75] P. Pebay, T. Baker et J. Pousin. "Dynamic meshing for finite element based segmentation of cardiac imagery". In : *Fifth world congress on Computational mechanics. Vienna*. 2002.

[76] Q.C. Pham. *Segmentation et mise en correspondance en imagerie cardiaque multimodale conduites par un modèle anatomique bi-cavités du coeur*. Thèse de doct. Institut National Polytechnique de Grenoble, 2002.

[77] F. Saint-Laurent et G. Martin. Real time determination and control of the plasma localisation and internal inductance in Tore Supra. *Fusion Engineering and Design* 56-57 (2001), p. 761–765.

[78] F. Sartori, A. Cenedese et F. Milani. JET real-time object-oriented code for plasma boundary reconstruction. *Fus. Engin. Des.* 66-68 (2003), p. 735–739.

[79] G Selig. *Equilibre évolutif à frontière libre et diffusion résistive dans un plasma de tokamak*. Thèse de doct. Université de Nice Sophia Antipolis, 2012.

[80] V.D. Shafranov. On magnetohydrodynamical Equilibrium configurations. *Soviet Physics JETP* 6.3 (1958), p. 1013.

[81] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.* 5 (1968), p. 506–517.

[82] W.C. Thacker. The role of the Hessian matrix in fitting models to measurements. *Journal of Geophysical Research* 94 (1989), p. 6177–6196.

[83] B. Van Leer. Towards the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection. *J. Comp. Phys.* 23 (1977), p. 276–299.

[84] F. Vincent, P. Clarysse, P. Croisille et I.E. Magnin. "An elastic-based region model and its application to the estimation of the heart deformation in tagged MRI". In : *ICIP-2000, Vancouver, BC, Canada*. 2000, p. 629–632.

[85] J. Wesson. *Tokamaks*. T. 118. International Series of Monographs on Physics. New York : Oxford University Press Inc., Third Edition, 2004.

Recueil d'articles

Article A : [4] J. Blum, C. Boulbe et B. Faugeras. Reconstruction of the equilibrium of the plasma in a Tokamak and identification of the current density profile in real time. *J. Comp. Phys.* 231 (2012), p. 960–980

Journal of Computational Physics 231 (2012) 960–980

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Reconstruction of the equilibrium of the plasma in a Tokamak and identification of the current density profile in real time

J. Blum, C. Boulbe, B. Faugeras *

Laboratoire J.A. Dieudonné, UMR 6621, Université de Nice Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 02, France

ARTICLE INFO

Article history:
Available online 13 April 2011

Keywords:
Inverse problem
Grad–Shafranov equation
Finite elements method
Real-time
Fusion plasma

ABSTRACT

The reconstruction of the equilibrium of a plasma in a Tokamak is a free boundary problem described by the Grad–Shafranov equation in axisymmetric configuration. The right-hand side of this equation is a nonlinear source, which represents the toroidal component of the plasma current density. This paper deals with the identification of this nonlinearity source from experimental measurements in real time. The proposed method is based on a fixed point algorithm, a finite element resolution, a reduced basis method and a least-square optimization formulation. This is implemented in a software called Equinox with which several numerical experiments are conducted to explore the identification problem. It is shown that the identification of the profile of the averaged current density and of the safety factor as a function of the poloidal flux is very robust.

1. Introduction

In fusion experiments a magnetic field is used to confine a plasma in the toroidal vacuum vessel of a Tokamak [1]. The magnetic field is produced by external coils surrounding the vacuum vessel and also by a current circulating in the plasma itself. The resulting magnetic field is helicoidal.

Let us denote by $\mathbf{j}$ the current density in the plasma, by $\mathbf{B}$ the magnetic field and by p the kinetic pressure. The momentum equation for the plasma is

$$\rho \frac{d\mathbf{u}}{dt} + \nabla p = \mathbf{j} \times \mathbf{B},$$

where $\mathbf{u}$ represents the mean velocity of particles and ρ the mass density. At the slow resistive diffusion time scale [2] the term $\rho \frac{d\mathbf{u}}{dt}$ can be neglected compared to ∇p and the equilibrium equation for the plasma simplifies to

$$\mathbf{j} \times \mathbf{B} = \nabla p$$

meaning that at each instant in time the plasma is at equilibrium and the Lorentz force $\mathbf{j} \times \mathbf{B}$ balances the force ∇p due to kinetic pressure. Taking into account the magnetostatic Maxwell equations which are satisfied in the whole space (including the plasma) the equilibrium of the plasma in presence of a magnetic field is described by

$$\mu_0 \mathbf{j} = \nabla \times \mathbf{B}, \tag{1}$$
$$\nabla \cdot \mathbf{B} = 0, \tag{2}$$
$$\mathbf{j} \times \mathbf{B} = \nabla p, \tag{3}$$

* Corresponding author.
E-mail address: Blaise.Faugeras@unice.fr (B. Faugeras).

0021-9991/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2011.04.005

where μ_0 is the magnetic permeability of the vacuum. Ampere's theorem is expressed by Eq. (1) and Eq. (2) represents the conservation of magnetic induction. From the equilibrium equation (3) it is clear that

$$\mathbf{B}\cdot\nabla p = 0 \quad \text{and} \quad \mathbf{j}\cdot\nabla p = 0.$$

Therefore field lines and current lines lie on isobaric surfaces. These isosurfaces form a family of nested tori called magnetic surfaces which enable to define the magnetic axis and the plasma boundary. On the one hand the innermost magnetic surface degenerates into a closed curve and is called magnetic axis and on the other hand the plasma boundary corresponds to the surface in contact with a limiter or to a magnetic separatrix (hyperbolic line with an *X*-point).

The Grad–Shafranov equation [3–5] is a rewriting of Eqs. (1)–(3) under the axisymmetric assumption. Consider the cylindrical coordinate system $(\mathbf{e}_r,\mathbf{e}_\phi,\mathbf{e}_z)$. The magnetic field $\mathbf{B}$ is supposed to be independent of the toroidal angle ϕ. Let us decompose it in a poloidal field $\mathbf{B}_p = B_r\mathbf{e}_r + B_z\mathbf{e}_z$ and a toroidal field $\mathbf{B}_\phi = B_\phi\mathbf{e}_\phi$ (see Fig. 1).

Let us also introduce the poloidal flux

$$\psi(r,z) = \frac{1}{2\pi}\int_D \mathbf{B}\,ds = \int_0^r B_z r\,dr$$

where D is the disc having as circumference the circle centered on the Oz axis and passing through a point (r,z) in a poloidal section. From Eq. (2) one deduces that $\mathbf{B}_p = \frac{1}{r}[\nabla\psi\times\mathbf{e}_\phi]$. Therefore $\mathbf{B}\cdot\nabla\psi = 0$ meaning that ψ is a constant on each magnetic surface and that $p = p(\psi)$.

The same poloidal–toroidal decomposition can be applied to $\mathbf{j}$. From Eq. (1) it is clear that $\nabla\cdot\mathbf{j} = 0$. As for $\mathbf{B}_p$ it is shown that there exists a function f, called the diamagnetic function, such that $\mathbf{j}_p = \frac{1}{r}\left[\nabla\left(\frac{f}{\mu_0}\right)\times\mathbf{e}_\phi\right]$. Since $\mathbf{j}\cdot\nabla p = 0$ then $\nabla f\times\nabla p = 0$ and f is constant on the magnetic surfaces, $f = f(\psi)$.

From Eq. (1) one also deduces that $\mathbf{B}_\phi = \frac{f}{r}\mathbf{e}_\phi$ and $\mathbf{j}_\phi = (-\Delta^*\psi)\mathbf{e}_\phi$ where

$$\Delta^*\cdot = \frac{\partial}{\partial r}\left(\frac{1}{\mu_0 r}\frac{\partial\cdot}{\partial r}\right) + \frac{\partial}{\partial z}\left(\frac{1}{\mu_0 r}\frac{\partial\cdot}{\partial z}\right).$$

To sum up

$$\begin{cases}\mathbf{B} = \mathbf{B}_p + \mathbf{B}_\phi,\\ \mathbf{B}_p = \frac{1}{r}[\nabla\psi\times\mathbf{e}_\phi],\\ \mathbf{B}_\phi = \frac{f}{r}\mathbf{e}_\phi\end{cases} \quad \text{and} \quad \begin{cases}\mathbf{j} = \mathbf{j}_p + \mathbf{j}_\phi,\\ \mathbf{j}_p = \frac{1}{r}[\nabla(\frac{f}{\mu_0})\times\mathbf{e}_\phi]\\ \mathbf{j}_\phi = -\Delta^*\psi\mathbf{e}_\phi.\end{cases}$$

From Eq. (3) one deduces that

$$(\mathbf{j}_p + j_\phi\mathbf{e}_\phi)\times(\mathbf{B}_p + B_\phi\mathbf{e}_\phi) = -\frac{1}{\mu_0 r}B_\phi\nabla f + j_\phi\frac{1}{r}\nabla\psi = \nabla p$$

and since

$$\nabla p = p'(\psi)\nabla\psi \quad \text{and} \quad \nabla f = f'(\psi)\nabla\psi$$

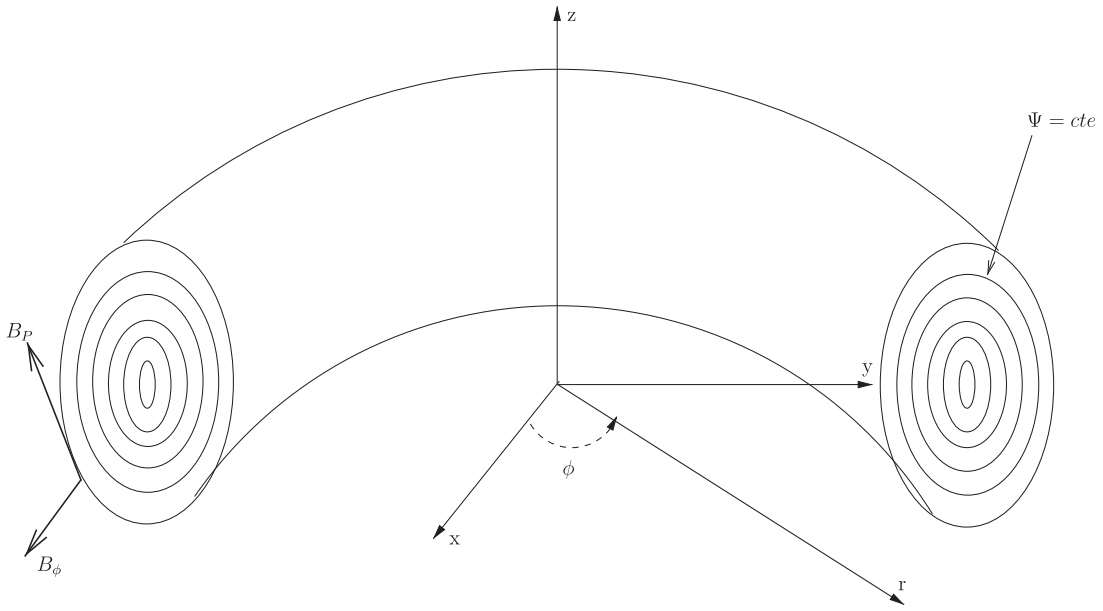

Fig. 1. Toroidal geometry.

the Grad–Shafranov equation valid in the plasma reads

$$-\Delta^*\psi = rp'(\psi) + \frac{1}{\mu_0 r}(ff')(\psi). \tag{4}$$

Thus under the axisymmetric assumption, the three dimensional equilibrium equations (1)–(3) reduce to a two dimensional non linear problem. Note that the right-hand side of Eq. (4) represents the toroidal component j_ϕ of the current density in the plasma which is determined by the unknown functions p' and ff'. In the vacuum there is no current and the poloidal flux satisfies

$$-\Delta^*\psi = 0.$$

In this paper, we are interested in the numerical reconstruction of the equilibrium, i.e. of the poloidal flux ψ and in the identification of the unknown plasma current density [6–8]. In a control perspective this reconstruction has to be achieved in real time from experimental measurements. The main difficulty consists in identifying the functions p' and ff' in the non linear right-hand side source term in Eq. (4). An iterative strategy involving a finite element method for the resolution of the direct problem and a least square optimization procedure for the identification of the non linearity using a decomposition basis is proposed.

Let us give a brief historical background of this problem of the reconstruction of the plasma current density from experimental measurements. In large aspect ratio Tokamaks with circular cross-sections, it was established in [9,10] that the quantities that can be identified from magnetic measurements are the total plasma current I_p and a sum involving the poloidal beta and the internal inductance: $\beta_p + l_i/2$ (see Appendix C). A large number of papers proved the possibility of separating β_p from l_i as soon as the plasma is no longer circular with high-aspect ratio [11–14]. The fact of adding supplementary experimental diagnostics, such as line integrated electronic density and Faraday rotation measurements, has considerably improved the identification of the current density profile [15,6,7]. The knowledge of the flux lines (from density or temperature measurements) enables in principle [16] to determine fully the two functions p' and ff' in the toroidal plasma current density, except in a particular case pointed out by [17] and studied by [18] and referred to as minimum-B equilibria. The difficulty in the reconstruction of the current profile, especially when only magnetic measurements are used, has been pointed out in [19] and is inherent to the ill-posedness of this inverse problem. The theory of variances in equilibrium reconstruction [20] enables to determine by statistical methods what kind of plasma functions can be reconstructed in a robust way. The equilibrium reconstruction problem in the case of anisotropic pressure is treated in [21].

A certain number of mathematical results on the identifiability of the right-hand-side of the Grad–Shafranov equation from Cauchy boundary conditions on the plasma frontier exist and seem unknown from the physical community. They are first dealing with the cylindrical case where the equilibrium equation becomes $-\Delta\psi = p'(\psi)$ and where only one non-linearity has to be identified. It is clear that, if the plasma boundary is circular, then the magnetic field is constant on the plasma boundary and there is an infinity of non-linearities giving this value and the only information coming from the poloidal field on the plasma boundary is the total plasma current. In [22] it was proved that if p' is a real-analytic function, then in a domain with a corner there is only one non-linearity p' corresponding to a given poloidal field on the plasma boundary. Some angles in the proof were excluded but in [23] the proof was given for corners with arbitrary angles (including the 90° *X*-point case). Curiously the case where the plasma boundary is smooth is mathematically more difficult and it has been proved in [24] that, if the plasma is non-circular and if p' is affine in terms of ψ then there exists at most a finite number of affine functions corresponding to the Cauchy boundary conditions. The link with the Schiffer and Pompeiu conjectures which is clearly pointed out in this paper is particularly interesting. In [25] results of unicity for a class of affine functions or for exponential functions are given for special smooth boundaries and results of non-unicity for doublet-type configurations. Finally in [26] identifiability results are given for the full Grad–Shafranov equation in a domain with a corner, with some exceptions for the angle. Of course, in spite of all these identifiability results, the ill-posedness of the reconstruction of the non-linearities from the Cauchy boundary measurements remains and has to be tackled very cautiously.

Section 2 is devoted to the statement of the mathematical problem and to the description of the experimental measurements available. The proposed algorithm is described in Section 3. This methodology has been implemented in a software called Equinox and numerical results using synthetic and real measurements are presented in Section 4.

2. Setting of the direct and inverse problems

2.1. Experimental measurements

Although the unknown functions $p'(\psi)$ and $(ff')(\psi)$ cannot be directly measured in a Tokamak several measurements are available:

- Magnetic measurements: they represent the basic information on which any equilibrium reconstruction relies. Flux loops provide measurements of ψ and magnetic probes provide measurements of the poloidal field $\mathbf{B}_p$ at several points around the vacuum vessel. Let Ω be the domain representing the vacuum vessel and $\partial\Omega$ its boundary. In what follows we assume that we are able to obtain the Dirichlet boundary conditions $\psi = g_D$ and the Neumann boundary conditions $\frac{1}{r}\frac{\partial\psi}{\partial n} = g_N$ at any points of the contour $\partial\Omega$ thanks to a preprocessing of the magnetic measurements. This preprocessing can either

be a simple interpolation between real measurements or be the result of some boundary reconstruction algorithm which computes ψ outside the plasma satisfying $\Delta^*\psi = 0$ under the constraint of the measurements [27–29].

A second set of measurements which can be used as a complement to magnetic measurements are internal measurements:

- Interferometric measurements: they give the values of the integrals along a family of chords C_i of the electronic density $n_e(\psi)$ which is approximately constant on each flux line $\int_{C_i} n_e(\psi)dl = \gamma_i$.
- Polarimetric measurements: they give the value of the integrals

$$\int_{C_i} \frac{n_e(\psi)}{r}\frac{\partial\psi}{\partial n}\,dl = \alpha_i.$$

$\frac{\partial\psi}{\partial n}$ is the normal derivative of ψ along the chord C_i.

Even when using magnetic measurements only for the equilibrium reconstruction the numerical algorithm presented in this paper also uses:

- Current measurement: it gives the value of the total plasma current I_p defined by

$$I_p = \int_{\Omega_p} j_\phi\,dx.$$

Ampere's theorem shows that this quantity can be deduced from magnetic measurements.
- Toroidal field measurement: it gives the value B_0 of the toroidal component of the field in the vacuum at the point $(R_0, 0)$ where R_0 is the major radius of the Tokamak. This is used for the integration of ff' into f and for the computation of the safety factor q (see Appendix B).

2.2. Direct problem

The equilibrium of a plasma in a Tokamak is a free boundary problem. The plasma boundary is determined either as being the last flux line in a limiter L or as being a magnetic separatrix with an X-point (hyperbolic point). The region $\Omega_p \subset \Omega$ containing the plasma is defined by

$$\Omega_p = \{\mathbf{x} \in \Omega,\ \psi(\mathbf{x}) \geqslant \psi_b\},$$

where $\psi_b = \max_L \psi$ in the limiter configuration or $\psi_b = \psi(X)$ when an X-point exists.

In the vacuum region, the right-hand side of Eq. (4) vanishes and the equilibrium equation reads

$$\Delta^*\psi = 0 \quad \text{in } \Omega \setminus \Omega_p.$$

Let us introduce the normalized flux $\bar{\psi} = \frac{\psi-\psi_a}{\psi_b-\psi_a} \in [0,1]$ in Ω_p with $\psi_a = \max_{\Omega_p}\psi$, $A(\bar{\psi}) = \frac{R_0}{\lambda}p'(\psi)$ and $B(\bar{\psi}) = \frac{1}{\lambda\mu_0 R_0}(ff')(\psi)$. This is introduced so that the non dimensional and unknown functions A and B are defined and identified on the fixed interval $[0,1]$. Imposing Dirichlet boundary conditions the final equilibrium equation is expressed as the boundary value problem:

$$\begin{cases} -\Delta^*\psi = \lambda\left[\frac{r}{R_0}A(\bar{\psi}) + \frac{R_0}{r}B(\bar{\psi})\right]\chi_{\Omega_p} & \text{in } \Omega \\ \psi = g_D & \text{on } \partial\Omega \end{cases} \tag{5}$$

The free boundary aspect of the problem reduces to the particular non linearity appearing through χ_{Ω_p} the characteristic function of Ω_p. The parameter λ is a scaling factor used to ensure that the given total current value I_p is satisfied

$$I_p = \lambda\int_{\Omega_p}\left[\frac{r}{R_0}A(\bar{\psi}) + \frac{R_0}{r}B(\bar{\psi})\right]dx. \tag{6}$$

2.3. Inverse problem

The inverse problem consists in the identification of functions A and B from the measurements available. It is formulated as a least-square minimization problem

$$\begin{cases} \text{Find } A^*,\ B^*,\ n_e^* \text{ such that :} \\ J(A^*, B^*, n_e^*) = \inf\ J(A, B, n_e). \end{cases} \tag{7}$$

If magnetic measurements only are used the formulation only needs the A and B variables and the J_1 and J_2 terms in Eq. (8) below are not needed. When polarimetric and interferometric measurements are used, the electronic density $n_e(\bar{\psi})$ also has to be identified even if it does not appear in Eq. (5). The cost function J is defined by

$$J(A, B, n_e) = J_0 + J_1 + J_2 + J_\varepsilon. \tag{8}$$

J_0 describes the misfit between computed and measured tangential component of $\mathbf{B}_p$

$$J_0 = \frac{1}{2}\sum_{k=1}^{N}(w_k)^2\left(\frac{1}{r}\frac{\partial\psi}{\partial n}(M_k) - g_N(M_k)\right)^2,$$

where N is the number of points M_k of the boundary $\partial\Omega$ where the magnetic measurements are given.

$$J_1 = \frac{1}{2}\sum_{k=1}^{N_c}\left(w_k^{polar}\right)^2\left(\int_{C_k}\frac{n_e(\bar{\psi})}{r}\frac{\partial\psi}{\partial n}\,dl - \alpha_k\right)^2$$

and

$$J_2 = \frac{1}{2}\sum_{k=1}^{N_c}\left(w_k^{inter}\right)^2\left(\int_{C_k} n_e(\bar{\psi})dl - \gamma_k\right)^2.$$

N_c is the number of chords over which interferometry and polarimetry measurements are given. The weights w give the relative importance of the different measurements used. The influence of the choice of the weights on the results of the identification was extensively studied in [7]. As a consequence of the ill-posedness of the identification of A, B and n_e, a Tikhonov regularization term J_ε is introduced [30] where

$$J_\varepsilon = \frac{\varepsilon_A}{2}\int_0^1 [A''(x)]^2\,dx + \frac{\varepsilon_B}{2}\int_0^1 [B''(x)]^2\,dx + \frac{\varepsilon_{n_e}}{2}\int_0^1 \left[n_e''(x)\right]^2 dx$$

and ε_A, ε_B and ε_{n_e} are the regularization parameters.

The values of the different weights and parameters introduced in the cost function are discussed in Section 4.

It should be noticed here that magnetic measurements provide Dirichlet and Neumann boundary conditions. The choice was made to use the Dirichlet boundary conditions in the resolution of direct problem and to include the Neumann boundary conditions in the cost function formulated to solve the inverse problem. This is arbitrary and another solution could have been chosen.

3. Algorithm and numerical resolution

3.1. Overview of the algorithm

The aim of the method is to reconstruct the equilibrium and the toroidal current density in real time. At each time step determined by the availability of new measurements during a discharge, the algorithm consists in constructing a sequence $(\psi^n, \Omega_p^n, A^n, B^n, \lambda^n)$ converging to the solution vector $(\psi, \Omega_p, A, B, \lambda)$. The unknown function n_e may be added too if interferometry and polarimetry measurements are used. The sequence is obtained through the following iterative loop:

- Starting guess: $\psi^0, \Omega_p^0, A^0, B^0$ and λ^0 known from the previous time step solution.
- Step 1 – Optimisation step: compute λ^{n+1} satisfying (6)

$$\lambda^{n+1} = I_p \Big/ \int_{\Omega_p^n}\left[\frac{r}{R_0}A^n(\bar{\psi}^n) + \frac{R_0}{r}B^n(\bar{\psi}^n)\right]dx$$

then compute $A^{n+1}(\bar{\psi}^n)$ and $B^{n+1}(\bar{\psi}^n)$ using the least square procedure detailed in Section 3.2.2.
- Step 2 – Direct problem step: compute ψ^{n+1} solution to

$$\begin{cases} -\Delta^*\psi^{n+1} = \lambda^{n+1}\left[\frac{r}{R_0}A^{n+1}(\bar{\psi}^n) + \frac{R_0}{r}B^{n+1}(\bar{\psi}^n)\right]\chi_{\Omega_p^n} & \text{in } \Omega, \\ \psi^{n+1} = g_D \quad \text{on } \partial\Omega \end{cases} \tag{9}$$

and the new plasma domain Ω_p^{n+1}.
- n: = n + 1. If the process has not converged return to Step 1 else $(\psi, \Omega_p, A, B, \lambda) = (\psi^n, \Omega_p^n, A^n, B^n, \lambda^n)$. The process is supposed to have converged when the relative residu $\frac{\|\psi^{n+1}-\psi^n\|}{\|\psi^n\|}$ is small enough.

At each iteration of the algorithm, an inverse problem corresponding to the optimization step and an approximated direct Grad–Shafranov problem have to be solved successively. In Eq. (9), $\bar{\psi}^n$ is known and since the right-hand side does not depend on ψ^{n+1} the boundary value problem (9) is linear.

In the next section the numerical methods used to solve the two problems corresponding to steps 1 and 2 are detailed.

3.2. Numerical resolution

3.2.1. The finite element method for the direct problem

The resolution of the direct problem is based on a classical P^1 finite element method [31]. Let us consider the family of triangulation τ_h of Ω, and V_h the finite dimensional subspace of $H^1(\Omega)$ defined by

$$V_h = \{ v_h \in H^1(\Omega),\ \ v_{h|T} \in P^1(T),\ \ \forall T \in \tau_h \}$$

and introduce $V_h^0 = V_h \cap H_0^1(\Omega)$. The discrete variational formulation of the boundary value problem (9) reads

$$\begin{cases} \text{Find } \psi_h \in V_h \text{ with } \psi_h = g_D \text{ on } \partial\Omega \text{ such that} \\ \forall v_h \in V_h^0, \quad \int_\Omega \frac{1}{\mu_0 r} \nabla\psi_h \cdot \nabla v_h dx = \int_{\Omega_p} \lambda \left[\frac{r}{R_0} A(\overline{\psi^*}) + \frac{R_0}{r} B(\overline{\psi^*}) \right] v_h \, dx, \end{cases} \tag{10}$$

where ψ^* represents the known value of ψ at the previous iteration. Numerically the Dirichlet boundary conditions are imposed using the method consisting in computing the stiffness matrix $\widehat{K}$ of the Neumann problem and modifying it. Consider (v_i) a basis of V_h then $\widehat{K}_{ij} = \int_\Omega \frac{1}{\mu_0 r} \nabla v_i \nabla v_j \, dx$. The modifications consist in replacing the rows corresponding to each boundary node setting 1 on the diagonal terms and 0 elsewhere. At each iteration only the right-hand side of the linear system in which the Dirichlet boundary conditions appear has to be modified. The linear system corresponding to Eq. (10) can be written in the form

$$K \cdot \Psi = y + g, \tag{11}$$

where K is the $n \times n$ modified stiffness matrix, Ψ is the unknown vector of size n (the number of nodes of the finite elements mesh), y is the vector associated with the modified right-hand side of Eq. (10) and g is the vector corresponding to the Dirichlet boundary conditions.

The matrix K is sparse and let LU be its decomposition. The inverse matrix K^{-1} is not sparse. The linear system (11) is inverted using the LU decomposition since it is computationally cheaper than using the full inverse matrix K^{-1} which is nevertheless needed for the optimization step of the algorithm in Eq. (15) below.

The vector y depends on functions A and B which are determined in the optimization step. Functions A, B and n_e are decomposed on a finite dimensional basis $(\Phi_i)_{i=1,\dots,m}$of functions defined on $[0,1]$

$$A(x) = \sum_i^m a_i \Phi_i(x), \quad B(x) = \sum_i^m b_i \Phi_i(x) \quad \text{and} \quad n_e(x) = \sum_i^m c_i \Phi_i(x).$$

The vector y reads

$$y = Y(\overline{\psi^*})u, \tag{12}$$

where $u = (a_1, \dots, a_m, b_1, \dots, b_m) \in \mathbb{R}^{2m}$ is the vector of the components of functions A and B in the basis (Φ_i). The matrix Y of size $n \times 2m$ is defined as follows. Each row i of Y is decomposed as

$$Y_{ij}(\overline{\Psi^*}) = \begin{cases} \int_{\Omega_p} \lambda \frac{r}{R_0} \Phi_j(\overline{\psi^*}) v_i \, dx & \text{if } 1 \leqslant j \leqslant m, \\ \int_{\Omega_p} \lambda \frac{R_0}{r} \Phi_{j-m}(\overline{\psi^*}) v_i \, dx & \text{if } m+1 \leqslant j \leqslant 2m. \end{cases}$$

3.2.2. Detailed numerical algorithm

One equilibrium computation corresponds to one instant in time during a pulse. The quasi-static approximation consists in considering that at each instant the Grad–Shafranov equation is satisfied. During a pulse successive equilibrium configurations are computed with a time resolution Δt corresponding to the acquisition time of measurements:

- Initialization before the discharge: the modified stiffness matrix K, its LU decomposition as well its inverse K^{-1} are computed once for all and stored.
- Consider that the equilibrium at time $t - \Delta t$ is known and that a new set of measurements is acquired at time t.
- Computation of the new equilibrium at time t through the iterative loop briefly described in the previous section and detailed below:

The equilibrium from the previous time step is used as a first guess in the iterative loop.

Step 1 – Optimization step During the optimization step, n_e is first estimated from interferometric measurements and A and B are computed in a second time.

- Compute the electronic density n_e based on the equilibrium of the previous iteration $\bar{\psi}^*$ using a least square formulation for the minimum of J_2 with Tikhonov regularization and solving the associated normal equation: The flux $\bar{\psi}^*$ is given

$$n_e(x) = \sum_{j=1}^{m} v_j \phi_j(x).$$

The interferometric measurements for $i = 1, \ldots, n_c$ are

$$\gamma_i \approx \int_{C_i} n_e(\bar{\psi}^*) dl = \sum_j v_j \int_{C_i} \phi_j(\bar{\psi}^*) dl = \sum_j v_j B_{ij}.$$

The cost functional reads

$$J(v) = \frac{1}{2} \sum_i \left(w_i^{inter}\right)^2 \left(\sum_j B_{ij} v_j - \gamma_i\right)^2 + \frac{\varepsilon}{2} v^T \Lambda v = \frac{1}{2} \|D^{1/2}(Bv - \gamma)\|^2 + \frac{\varepsilon}{2} v^T \Lambda v,$$

where $D^{1/2} = diag(w_i^{inter})$ and the regularization matrix Λ is defined by

$$\Lambda_{ij} = \int_0^1 \Phi_i''(x) \Phi_j''(x) dx$$

and Φ_i'' is the second order derivative of the basis function Φ_i.
It is minimized solving the associated normal equation

$$(\alpha^2 (D^{1/2}B)^T (D^{1/2}B) + \hat{\varepsilon}\Lambda)\hat{v} = \alpha (D^{1/2}B)^T D^{1/2} \gamma. \tag{13}$$

Since $n_e \approx 10^{19}$ m^{-3} an adimensionalizing parameter $\alpha = 10^{19}$ m^{-3}, such that $v = \alpha \hat{v}$, is introduced in order to precondition the linear system which is inverted using LU decomposition, as well as a reasonable prescribed value for the non dimensional regularization parameter $\hat{\varepsilon} = \alpha^2 \varepsilon$.

- Compute λ^{n+1} satisfying Eq. (6). In the right-hand side y, λ appears in the product λu. In order to avoid any divergence issue due to the non uniqueness of λ (for all $\alpha, \lambda u = (\lambda\alpha)(\frac{u}{\alpha})$) the degrees of freedom (dofs) u are scaled by $m = \max(|a_i|)$, u is replaced by $\frac{1}{m}u$ and λ by $m\lambda$.
- Compute A and B. In order to approximate A and B, suppose n_e is known and consider the discrete approximated inverse problem

$$\begin{cases} \text{Find } u \text{ minimizing :} \\ J(u) = \frac{1}{2}\|C(\psi^*)\Psi - d\|_D^2 + \frac{\varepsilon}{2} u^T \Lambda u. \end{cases} \tag{14}$$

where $C(\psi^*)$ is the observation operator and d the vector of experimental measurements. The first term in J is the discrete version of $J_0 + J_1$. The second one corresponds to the first two terms of the Tikhonov regularization J_ε with $\varepsilon_A = \varepsilon_B = \varepsilon$ which will always be assumed in order for functions A and B to play a symmetric role.

Let us denote by l the number of measurements available ($l = N + N_c$ if magnetic and polarimetric measurements are used) and by D the diagonal matrix made of the weights w_k and w_k^{polar}, the norm $\|\cdot\|_D$ is defined by $\forall \mathbf{x} \in \mathbb{R}^l \|\mathbf{x}\|_D^2 = (D\mathbf{x}, \mathbf{x}) = (D^{1/2}\mathbf{x}, D^{1/2}\mathbf{x})$.

$C(\psi^*)$ is a sparse matrix of size $l \times n$ and can be viewed as a vector composed of two blocks C_0 of size $N \times n$ and independent of ψ^* and $C_1(\psi^*)$ of size $N_c \times n$ corresponding respectively to J_0 and J_1. That is to say that multiplication of the kth row of C_0 by ψ gives the kth Neumann boundary condition approximation

$$(C_0)_k \Psi \approx \left(\frac{1}{r}\frac{\partial \psi}{\partial n}\right)(M_k).$$

The block $C_1(\psi^*)$ depends on ψ^* through the $n_e(\psi^*)$ function. The multiplication of the kth row of $C_1(\psi^*)$ by Ψ gives the kth polarimetric measurements approximation

$$(C_1(\psi^*))_k \Psi \approx \int_{c_k} \frac{n_e(\psi^*)}{r} \frac{\partial \psi}{\partial n} dl.$$

The matrix Λ is of size $2m \times 2m$ and is block diagonal composed of two blocks Λ_1 and Λ_2 of size $m \times m$, with

$$(\Lambda_1)_{ij} = (\Lambda_2)_{ij} = \int_0^1 \Phi_i''(x) \Phi_j''(x) dx.$$

Using Eqs. (11) and (12) problem (14) becomes

$$J(u) = \frac{1}{2}\|C(\psi^*)\Psi - d\|_D^2 + \frac{\varepsilon}{2} u^T \Lambda u = \frac{1}{2}\|C(\psi^*)K^{-1}Y(\overline{\psi^*})u + (C(\psi^*)K^{-1}g - d)\|_D^2 + \frac{\varepsilon}{2} u^T \Lambda u = \frac{1}{2}\|Eu - f\|_D^2 + \frac{\varepsilon}{2} u^T \Lambda u,$$

where $E = C(\psi^*)K^{-1}Y(\overline{\psi^*})$ and $f = -C(\psi^*)K^{-1}h + d$. Setting $\widetilde{E} = D^{1/2}E$, problem (14) reduces to solve the normal equation

$$(\widetilde{E}^T\widetilde{E} + \varepsilon\Lambda)u = \widetilde{E}^T f \tag{15}$$

whose solution is denoted by u^*.

Step 2 – Direct problem step. Update the dofs u and update the flux ψ by solving the linear system

$$K\psi = Y(\bar{\psi}^*)u^* + g \tag{16}$$

using the LU decomposition of matrix K. Update Ω_p possibly computing the position of the X-point if the plasma is not in a limiter configuration.

Finally it should be noticed that this algorithm is particularly well adapted to real-time applications. Indeed during the computations the expensive operations are the updates of matrices C and Y as well as the computation of products CK^{-1} and $CK^{-1}Y$ which appear in Eq. (15). In order to reduce computation time the K^{-1} matrix is precomputed and only the ψ-dependent part of C is dealt with. The resolution of the direct problem, Eq. (16), is cheap since the LU decomposition of the K matrix is also precomputed.

4. Numerical results

4.1. Twin experiment with noise free magnetic measurements

In this section we assume that the poloidal flux corresponding to an equilibrium configuration ψ is given on the boundary Γ. These Dirichlet boundary conditions can either be real measurements or can be the output from some equilibrium simulation code. In a first step we also assume to know functions p' and ff' (or A and B). In what follows these reference functions are given point by point. It is then possible to run a direct simulation to compute ψ on Ω (see Fig. 2) and thus $\frac{1}{r}\frac{\partial\psi}{\partial n}$ on Γ which can then be used as measurements in an inverse problem resolution.

In this first experiment the magnetic measurements are free of noise. The identification algorithm is initialized using the first guess functions are $A(x) = B(x) = 1 - x$ and $\lambda = 1$. The poloidal flux ψ is initially a constant on Ω. The weights in the misfit part of the cost function J_0 related to magnetic measurements are defined by $w_k = \frac{1}{\sqrt{N}\sigma}$. Since the error on magnetic measurements are of about one percent we define $\sigma = 0.01B_m$ where B_m is a mean magnetic field value which thanks to Ampere's theorem can be defined as $B_m = \frac{\mu_0 I_p}{|\Gamma|}$.

The functions A and B are decomposed in a function basis defined on the interval $[0,1]$. Several basis have been tested (piecewise affine functions, polynomials, B-splines and wavelets) in order to verify that the result of the identification does not depend on the decomposition basis. This is the case as long as the dimension of the basis is large enough. In the remaining part of this paper each function is decomposed in the same basis of 8 B-splines [32]. The boundary condition $A(1) = B(1) = 0$ is imposed.

The computations are carried out for several values of the regularization parameters ε ranging from 10^{-10} to 1. We are interested in the ability of the method to recover functions A and B and thus the current density profile averaged over the magnetic surfaces (see Appendix A):

$$R_0\left\langle \frac{j(r,\bar{\psi})}{r} \right\rangle = \lambda A(\bar{\psi}) + \lambda R_0^2 \left\langle \frac{1}{r^2} \right\rangle B(\bar{\psi})$$

and the safety factor q (see Appendix B).

As can be seen from Fig. 3 the optimal choice for ε is of about 10^{-5} for which functions A and B are well recovered. For smaller values some oscillations appear because the regularization is not strong enough and on the contrary greater values lead to less precision in the recovery of the unknown functions since regularization is too strong. In the second column the relative errors on the identified functions are plotted.

Fig. 4 shows an important point. Almost whatever the chosen value of ε is, i.e. whatever the quality of the identification of A and B is, the identified averaged current density $R_0\left\langle \frac{j(r,\bar{\psi})}{r} \right\rangle$ as well as the safety factor q are always well recovered and the relative errors are one order of magnitude smaller than for functions A and B. The same kind of observation was made in [8] where the identified functions A and B seemed to be rather sensitive to perturbations whereas the averaged current density was very stable.

In Table 1, the evolution of the relative residu on ψ, A, B and λ versus the number of iterations is given. It demonstrates numerically the convergence of the algorithm in this case where a value of 10^{-6} is used as stop condition. The algorithm needs 10 iterations to converge. It is interesting to notice that even though the first guess is not particularly well chosen the relative residu on ψ at the second iteration has already fallen to 4%. In real applications when simulating a whole pulse the first guess for the computation of the equilibrium at t is the equilibrium computed at $t - \Delta t$ and two iterations are enough to ensure a good convergence of the algorithm.

4.2. Twin experiment with noisy magnetic measurements

Figs. 5 and 6 show the results of the same type of numerical experiment but with noisy measurements. Each magnetic input, m representing either ψ or $\frac{1}{r}\frac{\partial\psi}{\partial n}$ at a point of the domain boundary Γ, is perturbed with a one percent noise normally distributed, $m_\eta = m + \eta$ with $\eta \sim N(m, 0.01m)$. For each chosen value of the regularization parameter the algorithm is run 200 times with measurements randomly perturbed as above. Then for each function λA, $\lambda R_0^2\left\langle \frac{1}{r^2} \right\rangle B$, $R_0\left\langle \frac{j(r,\bar{\psi})}{r} \right\rangle$ and q, a mean function and a standard deviation function are computed.

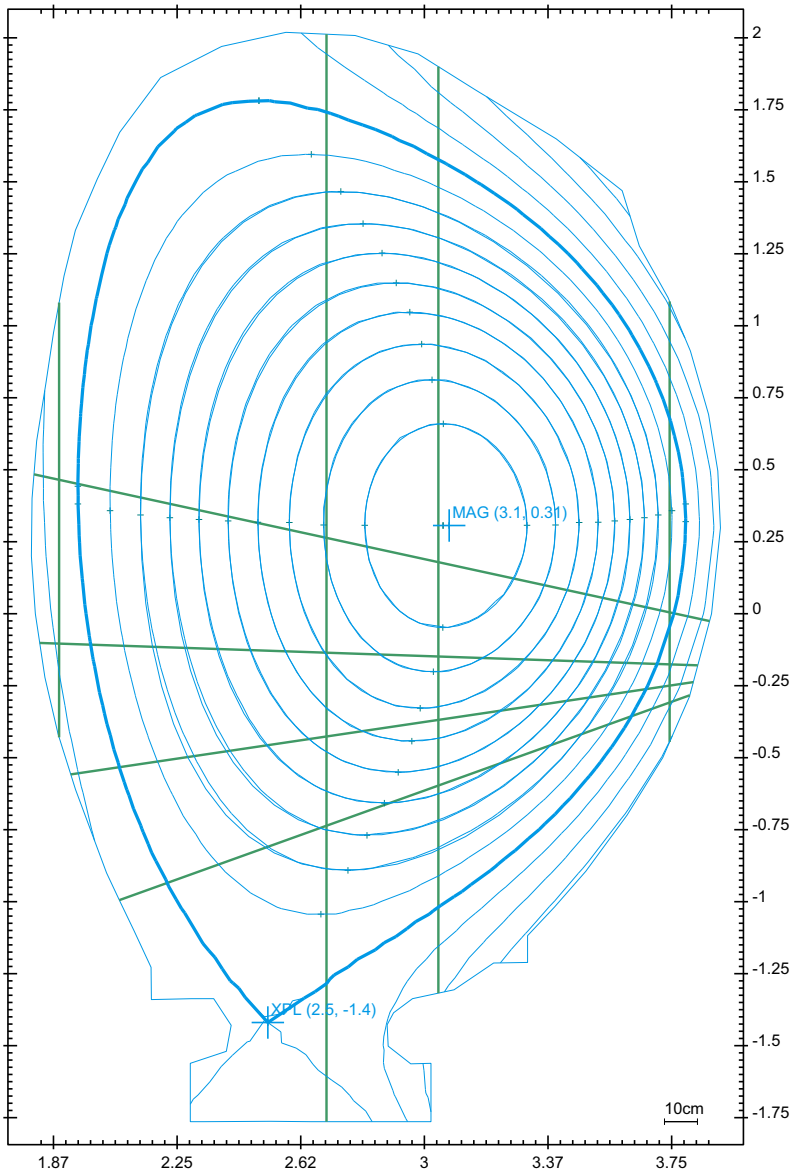

Fig. 2. An equilibrium configuration for the Tokamak JET from which twin experiments are performed. The domain Ω and its boundary Γ (external blue line) are shown. Isoflux are plotted from $\bar{\psi} = 0$ (magnetic axis) to $\bar{\psi} = 1$ (plasma boundary represented by the thick blue line running through the X-point $(2.5, -1.4)$) by step of $\Delta\bar{\psi} = 0.1$ Interferometry and polarimetry chords appear in green. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

In comparison with the noise free case the regularization parameter needs to be significantly increased to values of at least $\varepsilon = 10^{-2}$ and for a safer convergence of the algorithm to $\varepsilon = 10^{-1}$. For smaller values the algorithm either does not converge or gives very oscillating identified functions.

The mean error on the reconstructed functions is always smaller in the interval $\bar{\psi} \in [0.5, 1]$ than in the interval $[0, 0.5]$. This is due to the fact that magnetic measurements are external to the plasma and do not provide enough information to properly reconstruct the functions in the innermost part of the plasma.

As ε increases the variability or the standard deviation on the identified functions decreases. With small ε the algorithm can find very different functions depending on the perturbations of the measurements. With $\varepsilon = 10^{-2}$ the variability in the identified functions A and B is strong however the mean identified functions are close to the exact reference ones. On the other hand with $\varepsilon = 1$ the variability of the identified functions is strongly reduced but they are quite different from the exact reference functions in the interval $[0, 0.5]$.

It is worth noticing that in all cases the resulting safety factor q and averaged current density $R_0\left\langle \frac{j(r,\bar{\psi})}{r} \right\rangle$ are well recovered. The remark of the preceding section on the identifiability of the averaged current density still holds: it is quite well recovered even if functions A and B taken separately are not well identified. The mean error on the current density profile is almost always smaller than the mean errors on functions A and B. Moreover this error does not change very much between the different cases and particularly between the $\varepsilon = 10^{-1}$ and the $\varepsilon = 1$ cases. This implies that for a large interval of ε the value of

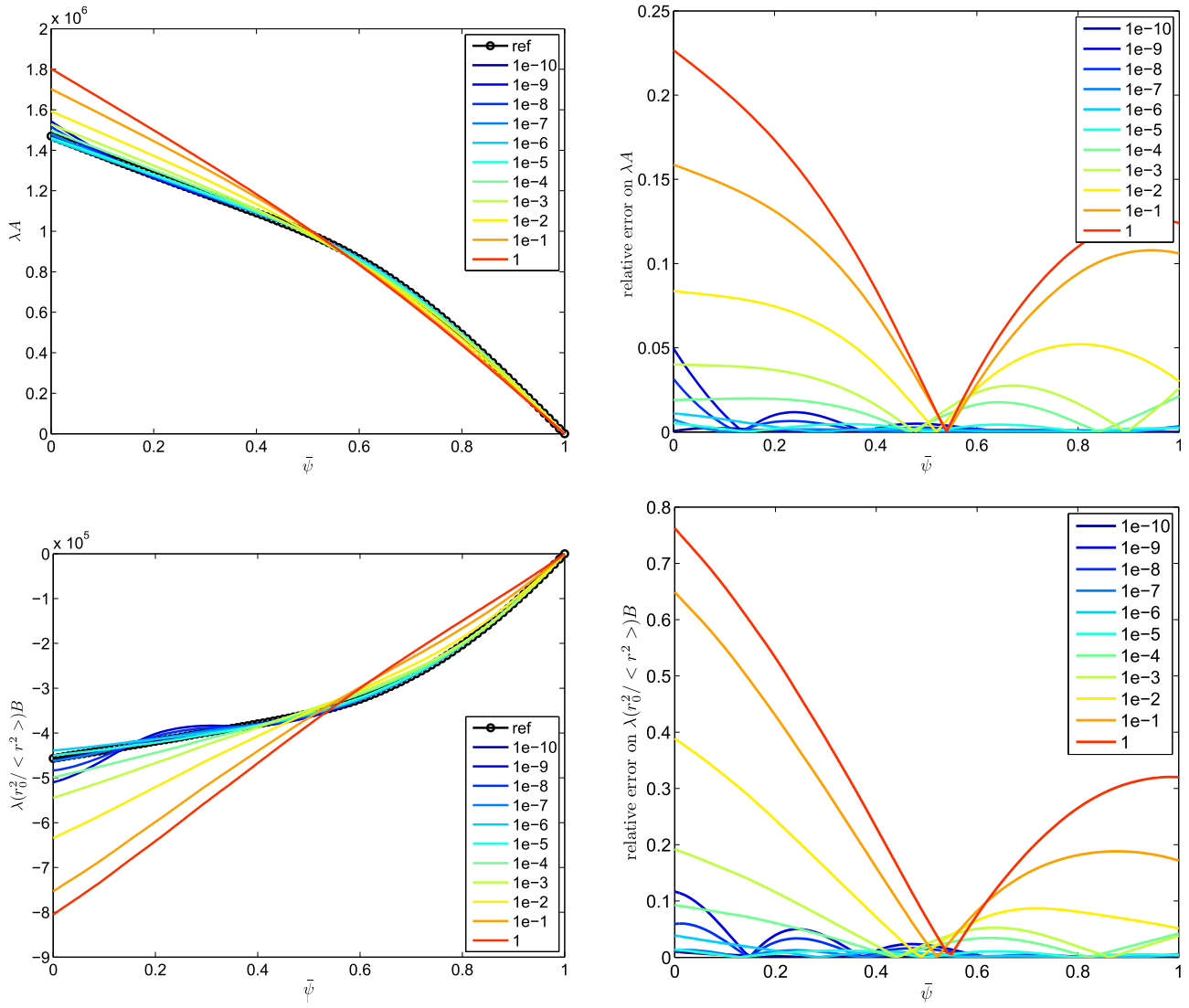

Fig. 3. Twin experiment with noise free measurements and different regularization parameters ε ranging from 10^{-10} to 1. Left column: identified functions $\lambda A(\bar{\psi})$ and $\lambda R_0^2\langle\frac{1}{r^2}\rangle B(\bar{\psi})$ for each different ε value, and the known reference functions (almost superimposed with the $\varepsilon = 10^{-5}$ curve). Right column: relative errors.

the part of the cost function related to magnetic measurements J_0 is almost constant. Therefore it is difficult to find an optimal value for the regularization parameter. For example the L-curve method [33] for the determination of the regularization parameter can hardly be used and gives some results which are not very reliable since the L-curves are not well behaved and the location of the corner is not clear. The "L" is an almost vertical line. This is due to the fact that, in a large interval of ε values, an increase in ε implies a important decrease in the regularization term $\frac{1}{2}(u^*(\varepsilon))^T Au^*(\varepsilon)$ but does not lead to a significative increase in the misfit term $J_0(u^*(\varepsilon))$.

4.3. Twin experiment with noisy magnetic, interferometric and polarimetric measurements

In this last twin experiment, interferometric and polarimetric measurements are also used. At first a reference density profile, $n_e(x)$ is prescribed point by point on $[0,1]$, as well as the same reference A and B functions as in the previous twin experiments. Then similar to the preceding section the equilibrium is computed from given Dirichlet boundary condition. A set of artificial magnetic, interferometric and polarimetric measurements is generated. Finally several twin experiments with a 1% noise are performed and some statistics are computed. The weights related to interferometric and polarimetric measurements in the cost function are defined as

- $w_k^{polar} = \frac{1}{\sqrt{N_c}\sigma^{polar}}$, with $\sigma^{polar} = 10^{-1}$ rad.
- $w_k^{inter} = \frac{1}{\sqrt{N_c}\sigma^{inter}}$, with $\sigma^{inter} = 10^{18}$ m^{-3}.

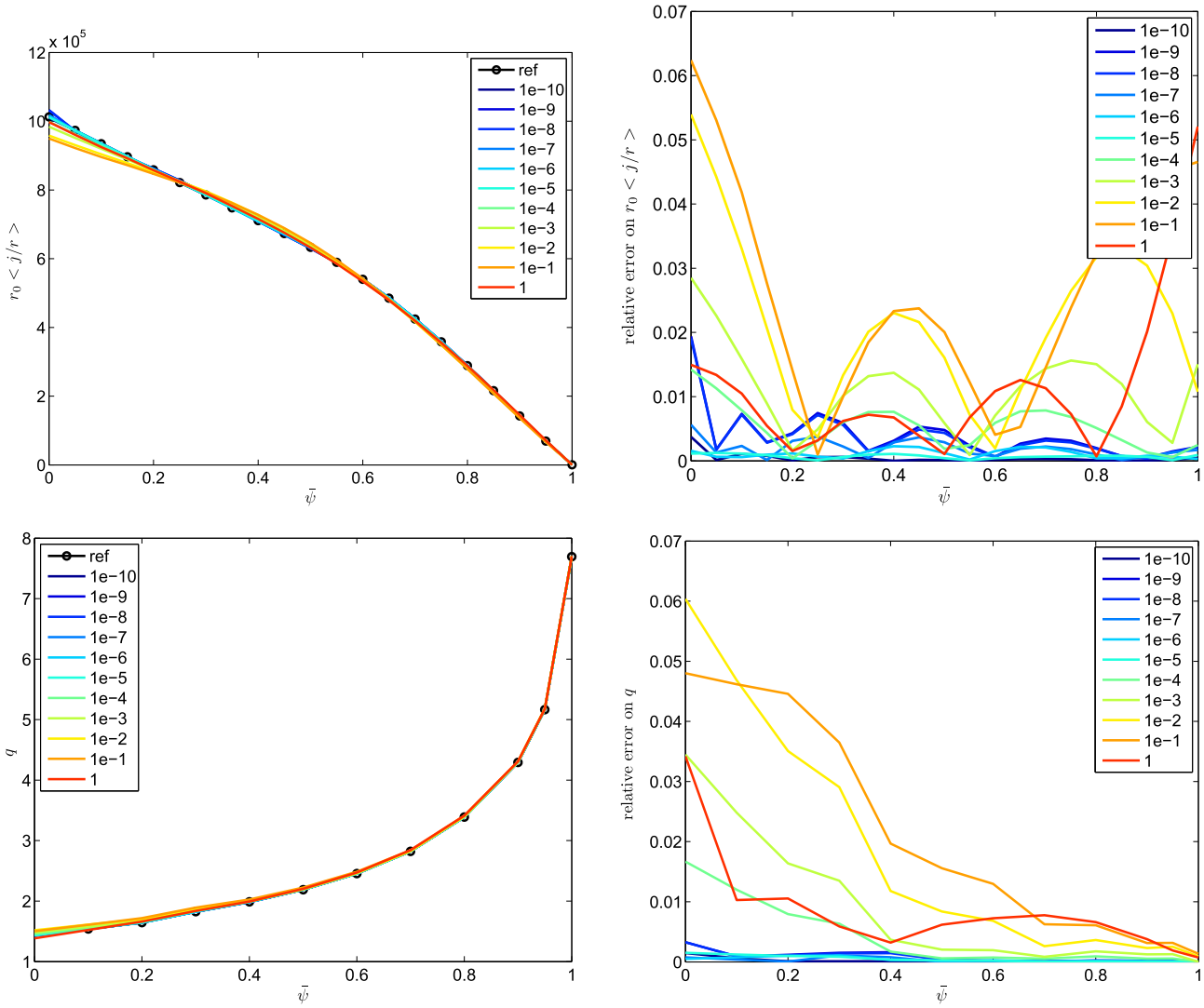

Fig. 4. Twin experiment with noise free measurements and different regularization parameters ε ranging from 10^{-10} to 1. Left column: resulting identified averaged current density $R_0\left\langle\frac{j(r,\bar{\psi})}{r}\right\rangle$, safety factor q for each ε value and the corresponding known reference values. Right column: relative errors.

Table 1
Numerical convergence of the algorithm.

Iteration n	$\frac{\|\psi^{n+1}-\psi^n\|}{\|\psi^n\|}$	$\frac{\|A^{n+1}-A^n\|}{\|A^n\|}$	$\frac{\|B^{n+1}-B^n\|}{\|B^n\|}$	$\frac{\|\lambda^{n+1}-\lambda^n\|}{\|\lambda^n\|}$
1	2.64809	6.07599	5.3509	0.100127
2	0.0408642	1.19473	1.42619	9.24968
3	0.0733385	1.83005	1.47338	0.563235
4	0.0404254	0.884617	1.0359	0.108107
5	0.00539736	4.79091	4.37571	0.826455
6	0.000349811	0.127626	0.180449	0.0889022
7	1.58606e−05	0.0262942	0.0246657	0.0263
8	5.67036e−06	0.00294791	0.0024952	0.00315952
9	1.4533e−06	0.000339986	0.000273055	0.000362224
10	6.19066e−07	6.41923e−05	6.51076e−05	6.29838e−05

The determination of the regularization parameter for the density function n_e is far less a problem than for functions A and B since for example the L-curve method works quite well in this case (see Fig. 10 in the next section) and the n_e function is well recovered as shown in Fig. 9. The regularization parameter for the density function is set to $\varepsilon_{ne} = 10^{-2}$.

The statistical results of the twin experiments are shown in Figs. 7 and 8 for three different values of ε. The use of interferometric and polarimetric measurements adds supplementary constraints on the A and B functions. The variability in the recovered functions is less important than in the case where only magnetics are used particularly for $\bar{\psi} \in [0, 0.5]$. This is not

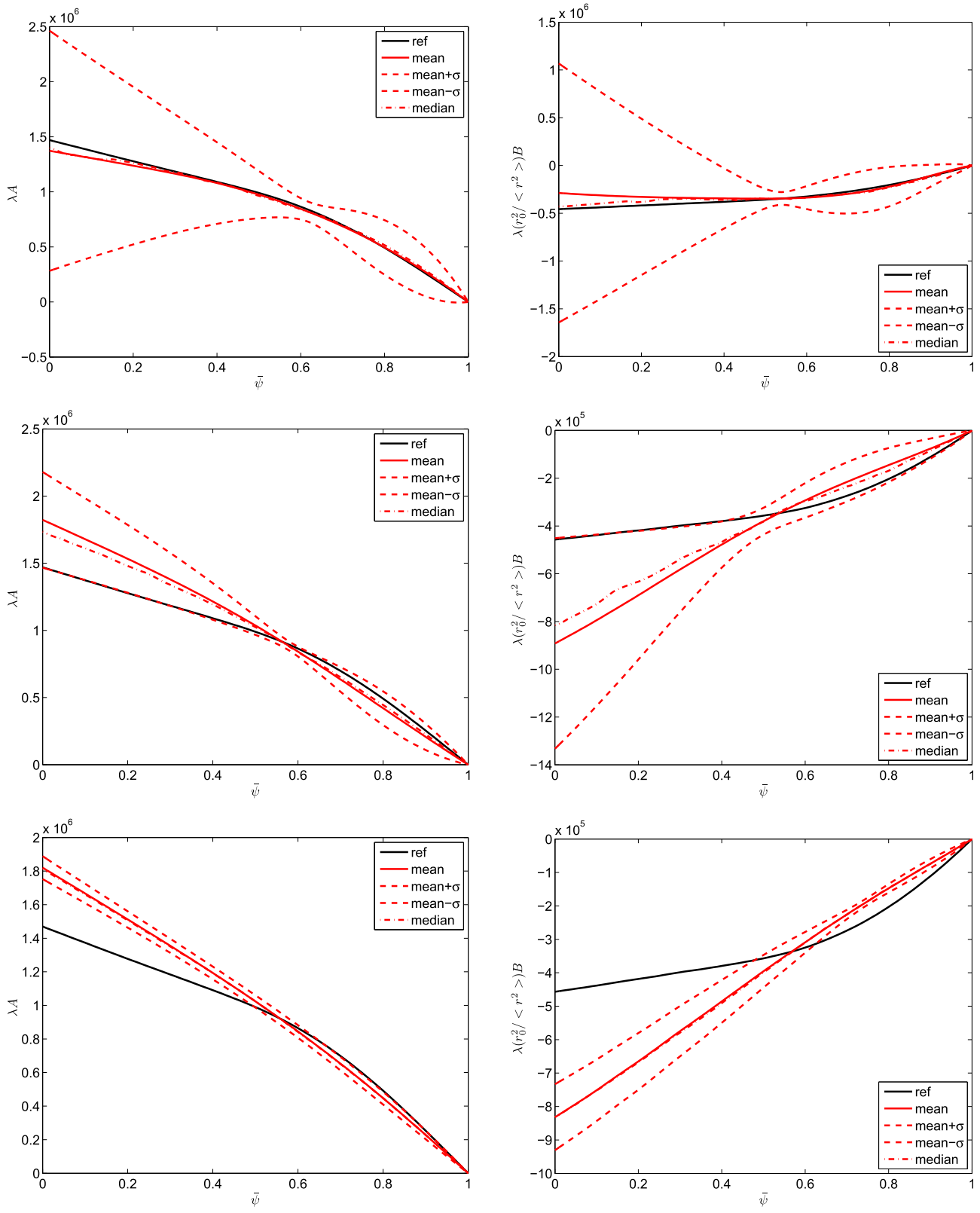

Fig. 5. Statistical results of the identification experiments with noisy magnetic measurements. Row 1: $\varepsilon = 10^{-2}$, row 2: $\varepsilon = 10^{-1}$, row 3: $\varepsilon = 1$. Column 1: function $\lambda A(\bar{\psi})$ and column 2: $\lambda R_0^2 \langle \frac{1}{r^2} \rangle B(\bar{\psi})$. For each function the reference value from which the unperturbed measurements were computed is given in black and the mean identified function in red. The mean ± standard deviation functions are shown in dashed red. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

surprising since the new measurements are internal and bring some information contained inside the plasma domain. Nevertheless it is not enough to perfectly reconstruct independently the A and B functions. This does not prevent an excellent recovery of the averaged current density profile and of the safety factor q. This phenomenon already observed in the magnetics case is emphasized here where the variability of the recovered profiles has decreased.

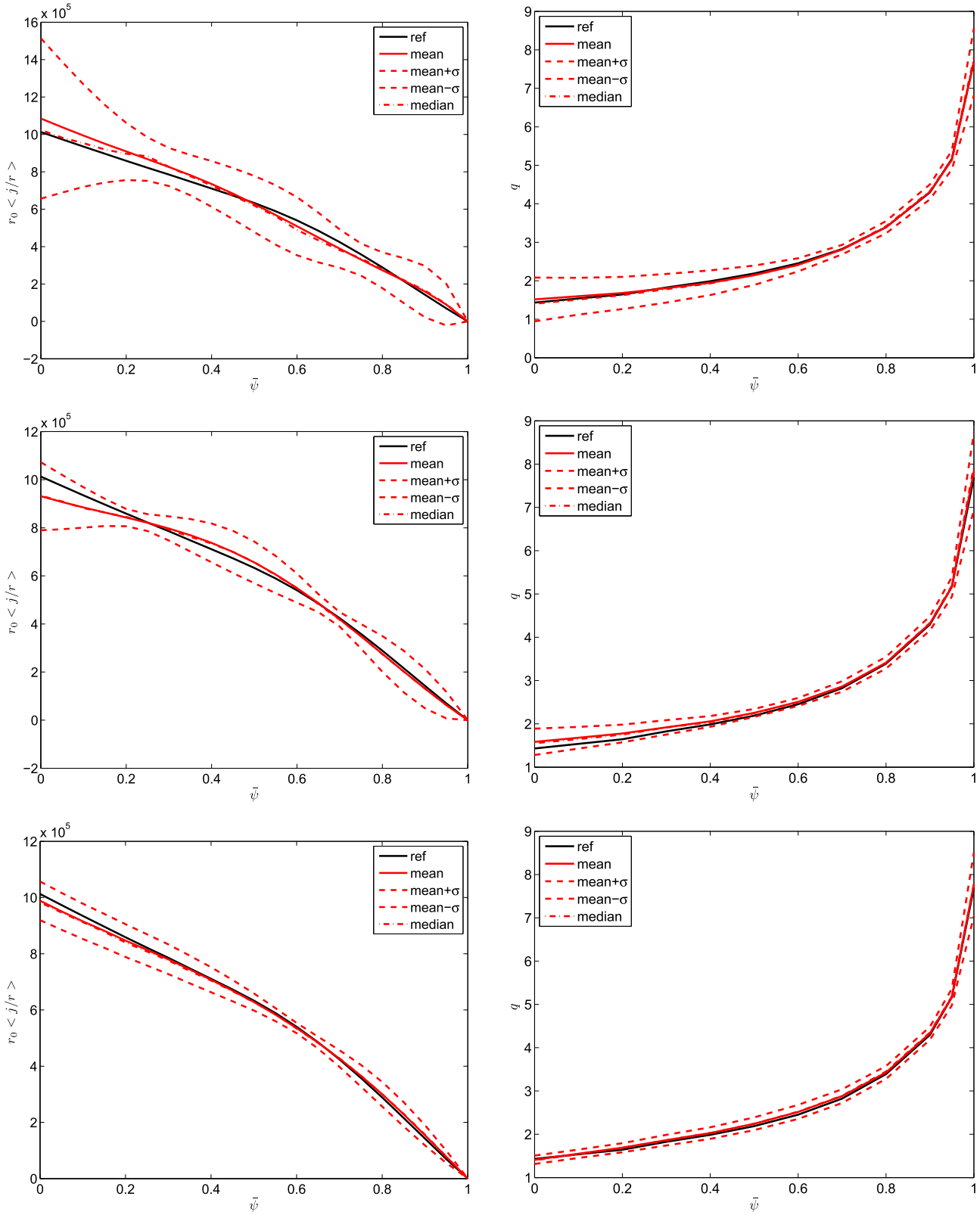

Fig. 6. Statistical results of the identification experiments with noisy magnetic measurements. Row 1: $\varepsilon = 10^{-2}$, row 2: $\varepsilon = 10^{-1}$, row 3: $\varepsilon = 1$. Column 1: $R_0\left\langle \frac{j(r,\bar{\psi})}{r} \right\rangle$, and column 2: safety factor q. For each function the reference value is given in black and the mean identified function in red. The mean ± standard deviation functions are shown in dashed red. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

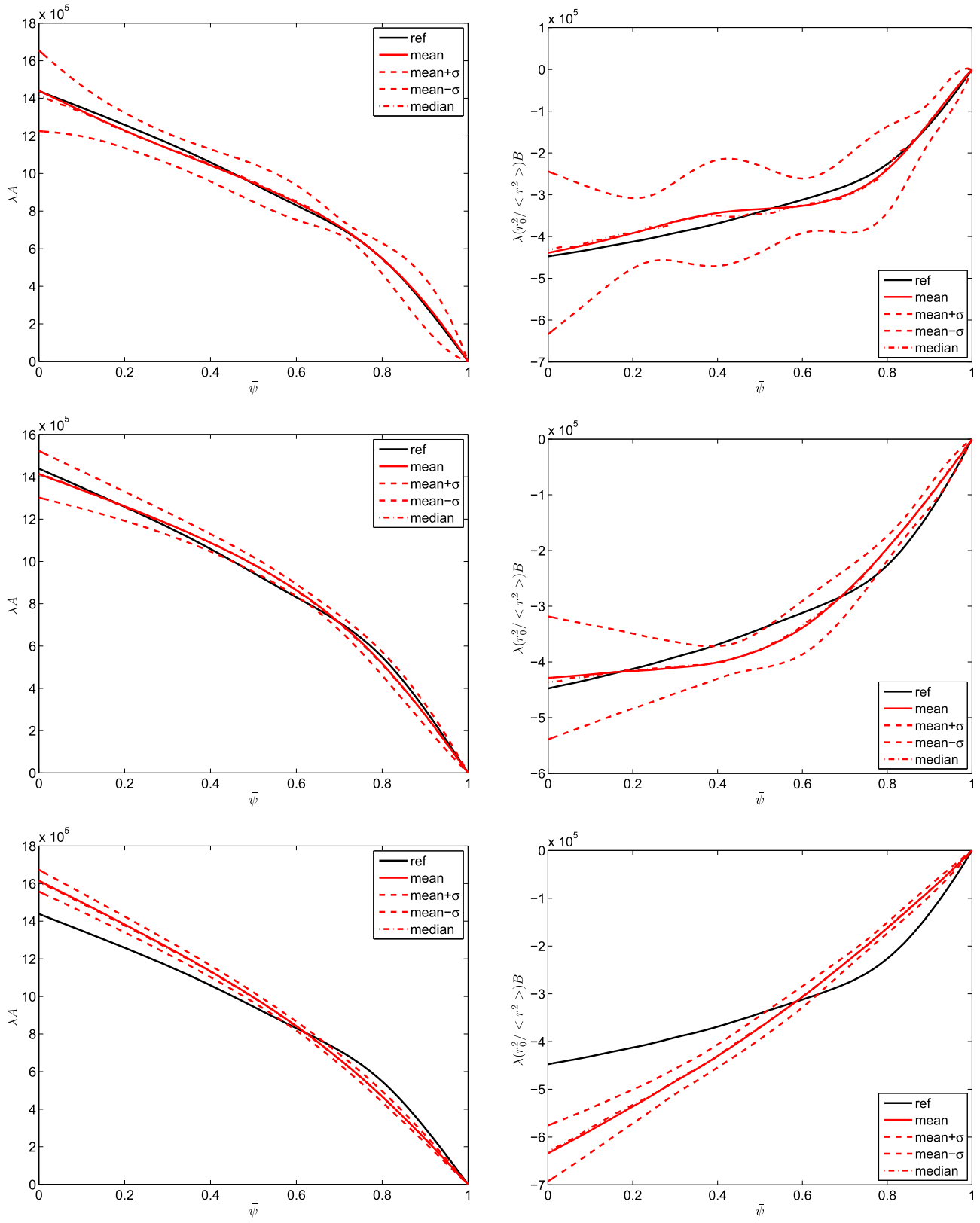

Fig. 7. Statistical results of the identification experiments with noisy measurements (magnetics, interferometry and polarimetry). Row 1: $\varepsilon = 10^{-2}$, row 2: $\varepsilon = 10^{-1}$, row 3: $\varepsilon = 1$. Column 1: function $\lambda A(\bar{\psi})$, and column 2: $\lambda R_0^2 \langle \frac{1}{r^2} \rangle B(\bar{\psi})$. For each function the reference value from which the unperturbed measurements were computed is given in black and the mean identified function in red. The mean ± standard deviation functions are shown in dashed red. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

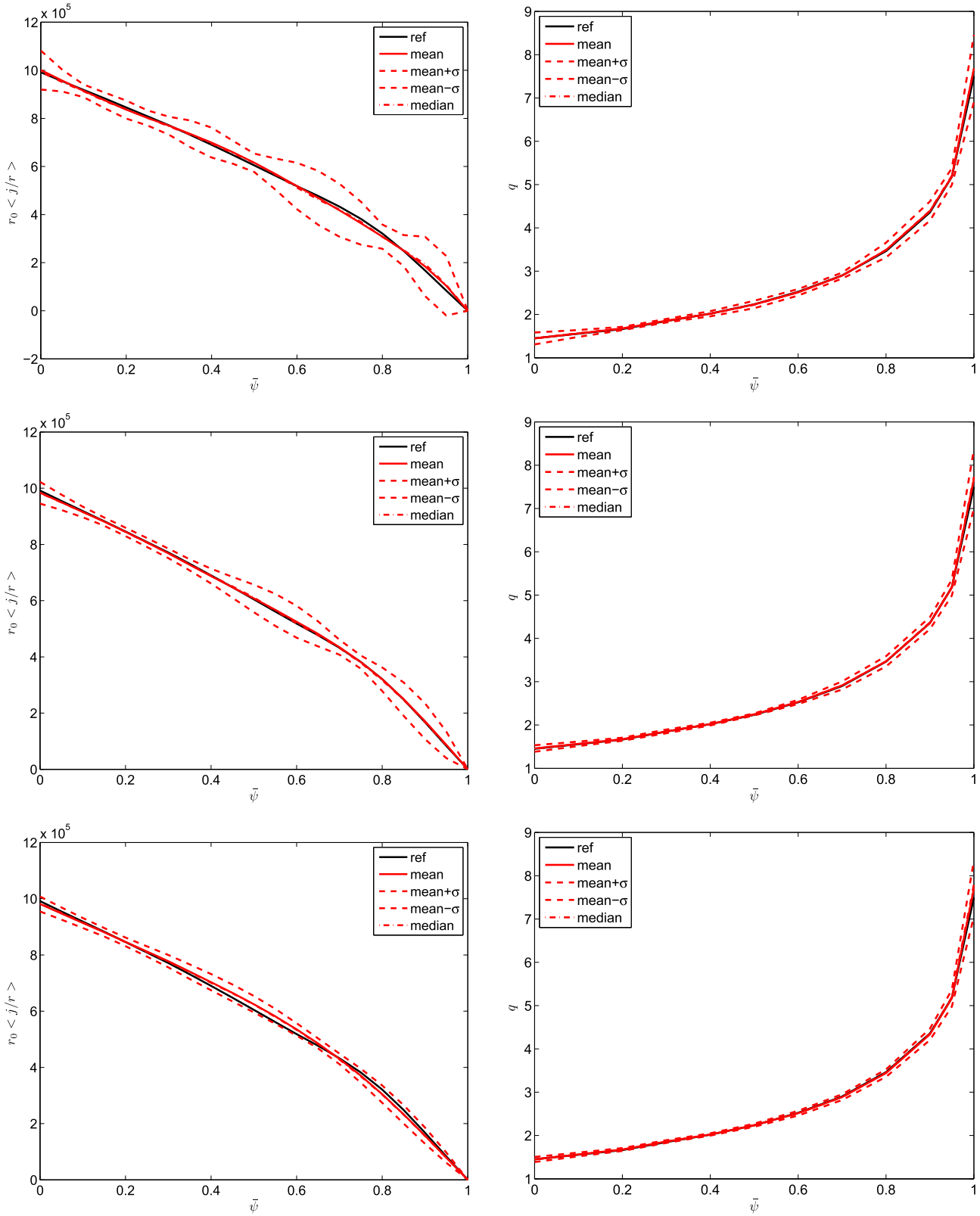

Fig. 8. Statistical results of the identification experiments with noisy measurements (magnetics, interferometry and polarimetry). Row 1: $\varepsilon = 10^{-2}$, row 2: $\varepsilon = 10^{-1}$, row 3: $\varepsilon = 1$. Column 1: $R_0\left\langle \frac{j(r,\bar{\psi})}{r} \right\rangle$, and column 2: safety factor q. For each function the reference value is given in black and the mean identified function in red. The mean ± standard deviation functions are shown in dashed red. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

4.4. A real pulse

The algorithm detailed in this paper has been implemented in a C++ software called Equinox developed in collaboration with the Fusion Department at Cadarache for Tore Supra and JET. Equinox can be used on the one hand for precise studies in which the computing time is not a limiting factor and on the other hand in a real-time framework to reconstruct the successive plasma equilibrium configurations during a whole pulse. For the time being it is used on JET and ToreSupra pulses, it has also been tested on the Tokamak TCV and can potentially be used on any Tokamak.

During the real time analysis of a whole pulse an equilibrium is reconstructed from new measurements with a time step of $\Delta t = 100$ ms. For each equilibrium reconstruction the number of iterations of the algorithm is set to 2. This enables fast enough computations while a very good precision is achieved since the initial guess for an equilibrium computation at time t is the equilibrium computed at time $t - \Delta t$. After 1 iteration a typical value for the relative residu on ψ is of 10^{-2} and it is of 10^{-3} after two iterations. Table 2 gives the size of the finite elements mesh used at ToreSupra and at JET as well as typical computation times on a laptop computer.

The choice of the regularization parameters is crucial since it determines the balance between the fit to the data and the regularity of the identified functions. It is also difficult as is shown in the preceding section. Ideally they should be determined for each equilibrium reconstruction. However this is not possible in a real-time application and the regularization parameters have to be set a priori to a constant value. From the twin experiments presented in the preceding sections it is quite clear that a good value for the regularization parameter ε is in the range $[10^{-2},1]$. By trial and error on different pulses using magnetics, interferometry and polarimetry, it appeared that a value of $\varepsilon = 5 \cdot 10^{-2}$ gave good results.

As for the identification of functions A and B the choice of a good regularization parameter for the identification of n_e is crucial. However in this case the L-curve method works quite well and it was used to determine the regularization parameters ε_{n_e} a priori on a number of equilibria for a few shots. The obtained values showed little variation and the choice of a mean value $\varepsilon = 0.01$ proved to be efficient. Fig. 10 shows an example of an L-curve computed for the identification of n_e.

Concerning real pulses at JET we refer to [34,35] in which a validation of Equinox is performed using many different pulses. This validation includes a posteriori comparison of the position of rational q surfaces computed from Equinox and deduced from soft X-rays measurements. The validation is satisfactory and shows again that when solving the inverse problem the use of interferometry, polarimetry and even Motional Stark Effect measurements at JET improves the location of rational q surfaces.

Here we only present an example of the output from Equinox on a ToreSupra pulse. Fig. 11 shows the equilibrium computed at time 20.408 s for ToreSupra pulse number 36,182 using magnetic measurements as well as interferometric and

Table 2
Typical mesh size and computation time for ToreSupra and JET.

	ToreSupra	JET
Finite element mesh		
Number of triangles	1382	2871
Number of nodes	722	1470
Computation time (1.80 GHz)		
One equilibrium	20 ms	60 ms

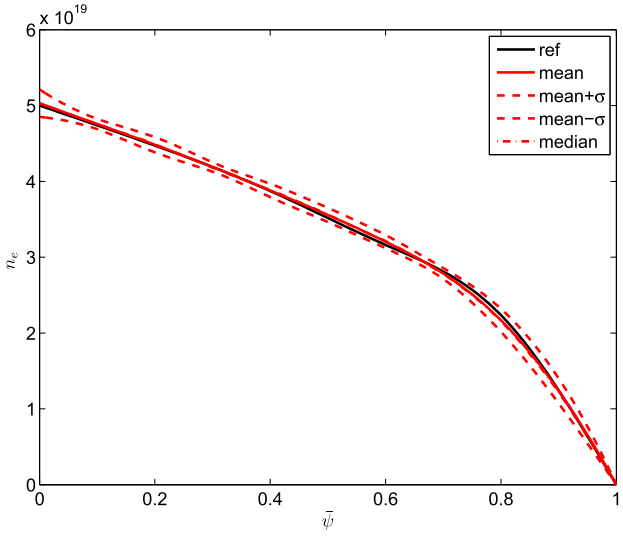

Fig. 9. Statistical results for the identification of the density function n_e with noisy interferometric measurements.

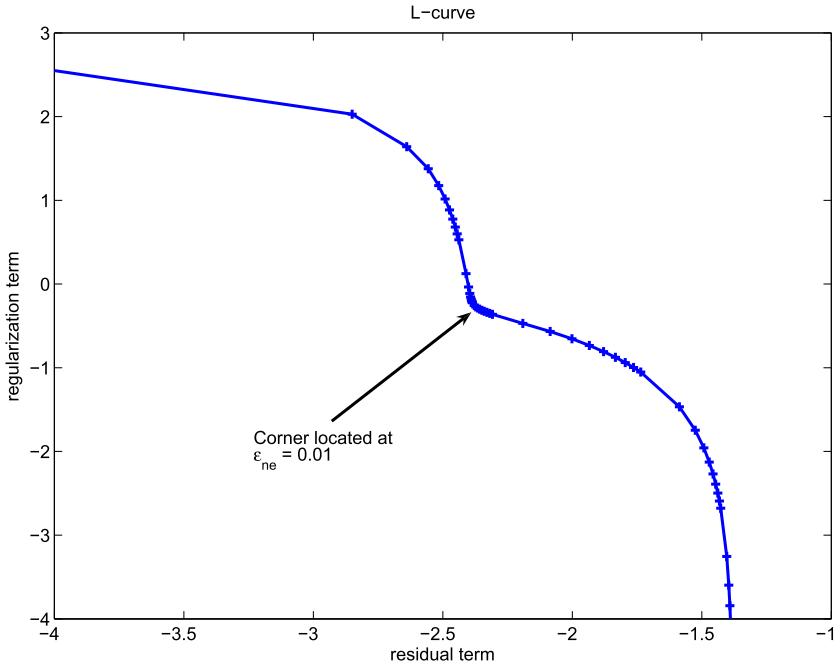

Fig. 10. Typical L-curve for the determination of ε_{n_e}. It is a plot of the parametric curve $x(\varepsilon_{n_e}) = \log\left(\frac{1}{2}\|D^{1/2}(Bv^*(\varepsilon_{n_e}) - \gamma)\|^2\right)$, $y(\varepsilon_{n_e}) = \log\left(\frac{1}{2}(v^*(\varepsilon_{n_e}))^T A v^*(\varepsilon_{n_e})\right)$ where $v^*(\varepsilon_{n_e})$ is the solution to Eq. (13). Hansen's algorithm [33] locates a corner at $\varepsilon_{n_e} = 0.01$.

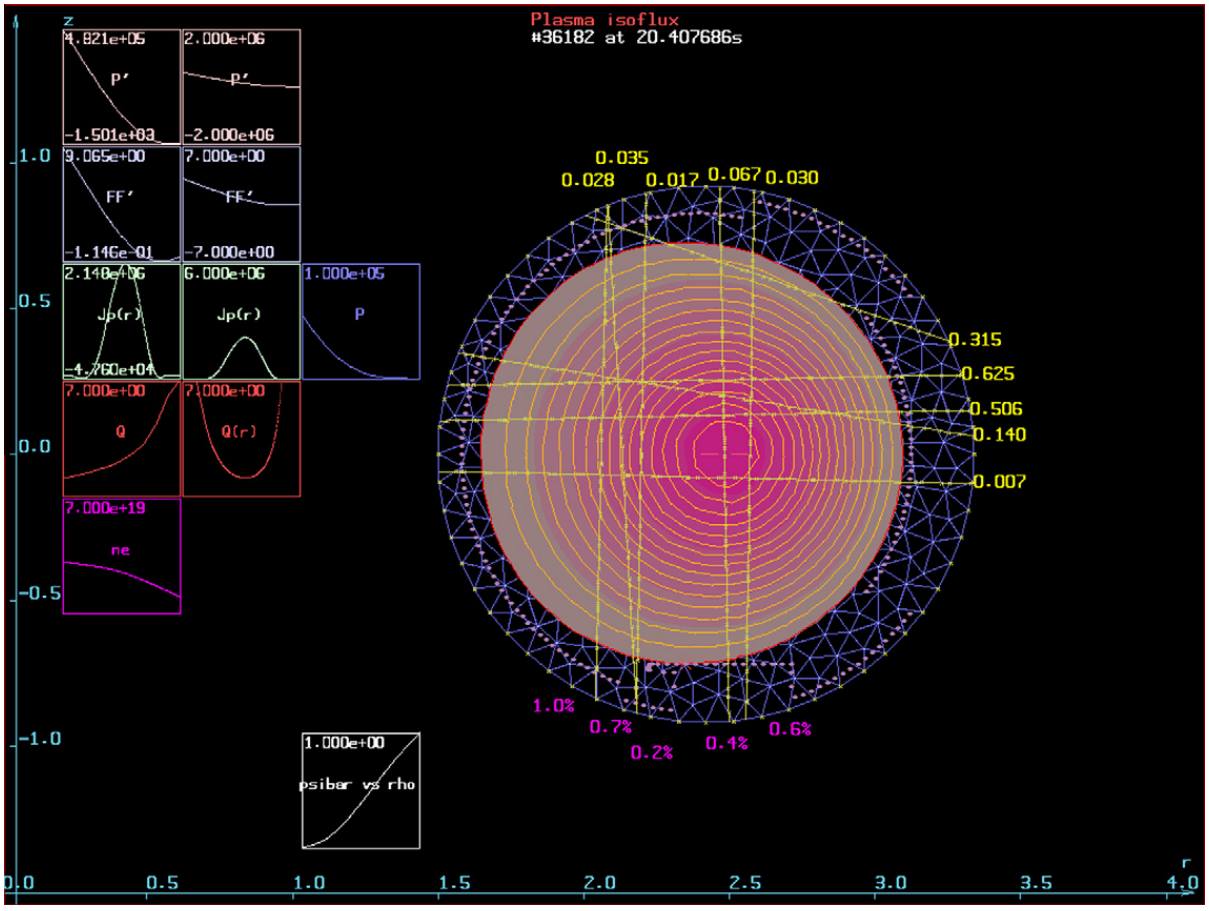

Fig. 11. Graphical output from Equinox. Reconstructed equilibrium at time 20.408 s for ToreSupra pulse number 36,182. Magnetic, interferometry and polarimetry measurements are used. See text for more details.

Table 3
β_p and l_i computed by Equinox and by Apolo for ToreSupra shot 36,182 at t = 20.408 s.

	β_p	l_i	$\beta_p + \frac{l_i}{2}$
Equinox	0.62	1.66	1.45
Apolo	0.70	1.55	1.47

polarimetric measurements. One can observe the position of the plasma in the vacuum vessel. Isoflux lines are displayed from the magnetic axis to the boundary. The interferometry and polarimetry chords are displayed. For each chord the error between computed and measured interferometry is given in purple. These errors are about 1% for the active chords. The polarimetry absolute errors are given in yellow. Different graphs are plotted on the left hand side of the display. On the first row the identified function A, and corresponding functions p' and p. On the second row the identified function B and corresponding function ff'. The third row gives the toroidal current density j_ϕ in the equatorial plane and the fourth one shows the safety factor q. Finally on the fifth row the identified n_e function is plotted.

It is of importance to compute the kinetic energy poloidal β_p parameter and the internal inductance l_i. In Equinox these equilibrium parameters are computed following the equations of Appendix C. For ToreSupra they are computed in the code Apolo [28] from the Shafranov integrals and from the toroidal plasma flux. The agreement between the two methods is good as shown in Table 3. The relative errors on β_p and l_i are about 10% while it is of about 1% on the sum $\beta_p + \frac{l_i}{2}$.

Finally it should be noticed that at ToreSupra or JET there does not exist reliable enough pressure measurements to be used in an inverse equilibrium reconstruction. The electron pressure p_e can be reasonably estimated from interferometry for the density n_e and Thomson scattering and Electron Cyclotron Emission for the temperature T_e. On the contrary very large uncertainties on the ion quantities n_i and T_i make the ion pressure p_i and thus the total pressure $p = p_e + p_i$ unusable in a real-time identification algorithm such as the one presented here. Moreover the quantity really important in order to constrain the identification of the p' term would be the pressure gradient on which the error bars are even larger.

5. Conclusion

We have presented an algorithm for the identification of the current density profile in the Grad–Shafranov equation and the equilibrium reconstruction from experimental measurements in real time. We have shown thanks to several twin experiments that even though the unknown functions A and B (or p' and ff') taken separately might not be always exactly identified the resulting averaged current density and safety factor seem to be always well identified. We have also shown that the use of internal polarimetric measurements improves the quality of the identification but is still not enough to perfectly identify both A and B. Finally we have introduced the software Equinox in which this methodology is developed. This work constitutes a step towards the real-time control of the safety factor and of the averaged current density profile in a Tokamak plasma which will be essential in nuclear fusion reactors.

Acknowledgments

The authors are grateful to Kristoph Bosak who developed a first version of the code Equinox. Although it has now been thoroughly modified this version was an essential basis to start from.

The authors would also like to thank all colleagues from the CEA at Cadarache in France involved in a collaboration between the University of Nice and the CEA through the LRC (Laboratoire de Recherche Conventionné). Discussions with Francois Saint-Laurent and Sylvain Bremond were particularly helpful. Emmanuel Joffrin initiated the real-time approach and Didier Mazon helped introducing us at JET where different people are also involved. In particular Luca Zabeo provided magnetic input data from the boundary code Xloc for Equinox and the work of Fabio Piccolo and Robert Felton is essential to implement Equinox on JET real-time system.

Appendix A. Average over magnetic surfaces

The method of averaging over the magnetic surfaces is detailed in [14, p. 242]. The average $\langle A \rangle$ of an arbitrary quantity A on a magnetic surface S is defined as

$$\langle A \rangle = \frac{\partial}{\partial V} \int_V A \, dV,$$

where V is the volume inside S. This notion of average has the following property:

$$\langle A \rangle = \frac{\int_{C_{\bar{\psi}}} \frac{A \, dl}{B_p}}{\int_{C_{\bar{\psi}}} \frac{dl}{B_p}}$$

where $C_{\bar{\psi}}$ is a closed contour $\bar{\psi} = cte \in (0,1)$ and $B_p = \frac{1}{r}\|\nabla\psi\|$.

Appendix B. Safety factor q

The safety factor is so called because of the role it plays in determining stability [1, p. 111]. It can be seen as the ratio of the variation of the toroidal angle needed for one magnetic field line to perform one poloidal turn

$$q = \frac{\Delta\phi}{2\pi}.$$

Since q is the same for all magnetic field lines on a magnetic surface it is a function of ψ (or $\bar{\psi}$). The expression of q used for computations is the following

$$q(\bar{\psi}) = \frac{1}{2\pi}\int_{C_{\bar{\psi}}} \frac{B_\phi}{rB_p} dl$$

where $C_{\bar{\psi}}$ is a closed contour $\bar{\psi} = cte \in (0,1)$, $B_\phi = \frac{f}{r}$ and

$$f(\psi) = \sqrt{(B_0R_0)^2 + \int_{\psi_b}^{\psi} (f^2)'(y)dy}.$$

Appendix C. Poloidal β_p and Internal inductance l_i

The full 3D plasma domain is denoted by D. The plasma domain in the poloidal section by Ω_p and its boundary $\partial\Omega_p = \Gamma_p$. Let us define $R_g = \frac{1}{2}(R_{left} + R_{right})$.

Surface and perimeter of a poloidal section. Let us define $S_p = \int_{\Omega_p} ds$ and $L_p = \int_{\Gamma_p} dl$. For a circular plasma of radius a: $L_p = 2\pi a$, $S_p = \pi a^2$ and $S_p = \frac{L_p^2}{4\pi}$. Even for non-circular plasma the following quantity is used:

$$\widehat{S}_p = \frac{L_p^2}{4\pi}. \tag{17}$$

Plasma volume

$$V_p = \int_D dv = \int_0^{2\pi}\int_{\Omega_p} r\,d\phi\,ds = 2\pi\int_{\Omega_p} r\,ds. \tag{18}$$

The following approximation can be used:

$$\widehat{V}_p = 2\pi R_g \widehat{S}_p. \tag{19}$$

Poloidal β_p. The ratio $\beta = \frac{p}{B^2/2\mu_0}$ represents the efficiency of the confinement of the plasma pressure by the magnetic field. The poloidal beta is defined as the ratio of the mean kinetic pressure of the plasma to its magnetic pressure [1, p. 116]:

$$\beta_p = \frac{\bar{p}}{B_{pa}^2/2\mu_0}, \tag{20}$$

where

$$\bar{p} = \frac{\int_D p\,dv}{\int_D dv} = \frac{\int_{\Omega_p} pr\,ds}{\int_{\Omega_p} r\,ds} \tag{21}$$

and

$$B_{pa} = \frac{\int_{\Gamma_p} B_p\,dl}{\int_{\Gamma_p} dl} = \frac{\mu_0 I_p}{L_p}. \tag{22}$$

Let us define the internal kinetic energy

$$W = \frac{3}{2}\int_D p\,dv.$$

We have

$$W = \frac{3}{2}\bar{p}V_p = \frac{3}{2}\frac{B_{pa}^2}{2\mu_0}V_p\beta_p$$

and from Eqs. (22), (19) and (17) follows that [1, p. 504]

$$W = \frac{3}{8}\mu_0 R_g I_p^2 \beta_b.$$

Then β_p can be approximated by

$$\beta_p = \frac{\frac{3}{2}\bar{p}V_p}{\frac{3}{8}\mu_0 R_g I_p^2} \tag{23}$$

which the default β_p computed by Equinox.

Internal inductance l_i

The internal inductance l_i of the plasma characterizes the current density profile [1, p. 120,14, p. 44]:

$$l_i = \frac{\overline{B}_p^2}{B_{pa}^2}, \tag{24}$$

where

$$\overline{B}_p^2 = \frac{\int_D B_p^2 \, dv}{\int_D dv}.$$

In Equinox the computation of l_i is done as follows:

$$l_i = \frac{\overline{B}_p^2 V_p}{B_{pa}^2 V_p}.$$

Using Eqs. (22), (19) and (17) leads to

$$l_i = \frac{\overline{B}_p^2 V_p}{\frac{\mu_0^2}{2} R_g I_p^2} \tag{25}$$

which is the default computation of l_i in Equinox.

References

[1] J. Wesson, Tokamaks, International Series of Monographs on Physics, third ed., vol. 118, Oxford University Press Inc., New York, 2004.
[2] H. Grad, J. Hogan, Classical diffusion in a tokomak, Physical Review Letters 24 (24) (1970) 1337–1340.
[3] H. Grad, H. Rubin, Hydromagnetic equilibria and force-free fields, in: 2nd U.N. Conference on the Peaceful uses of Atomic Energy, Geneva, vol. 31, 1958, pp. 190–197.
[4] V. Shafranov, On magnetohydrodynamical equilibrium configurations, Soviet Physics JETP 6 (3) (1958) 1013.
[5] C. Mercier, The MHD approach to the problem of plasma confinement in closed magnetic configurations, Lectures in Plasma Physics, Commission of the European Communities, Luxembourg, 1974.
[6] L. Lao, J. Ferron, R. Geoebner, W. Howl, H. St. John, E. Strait, T. Taylor, Equilibrium analysis of current profiles in Tokamaks, Nuclear Fusion 30 (6) (1990) 1035.
[7] J. Blum, E. Lazzaro, J. O'Rourke, B. Keegan, Y. Stefan, Problems and methods of self-consistent reconstruction of Tokamak equilibrium profiles from magnetic and polarimetric measurements, Nuclear Fusion 30 (8) (1990) 1475.
[8] J. Blum, H. Buvat, An inverse problem in plasma physics: the identification of the current density profile in a Tokamak, in: Biegler, Coleman, Conn, Santosa (Eds.), IMA Volumes in Mathematics and its Applications, Large Scale Optimization with Applications, Part 1: Optimization in Inverse Problems and Design, vol. 92, 1997, pp. 17–36.
[9] V.D. Shafranov, Determination of the parameters β_p and l_i in a Tokamak for arbitrary shape of plasma pinch cross-section, Plasma Physics 13 (9) (1971) 757.
[10] L. Zakharov, V. Shafranov, Equilibrium of a toroidal plasma with noncircular cross-section, Soviet Physics Technical Physics 18 (2) (1973) 151–156.
[11] J. Luxon, B. Brown, Magnetic analysis of non-circular cross-section Tokamaks, Nuclear Fusion 22 (6) (1982) 813–821.
[12] D. Swain, G. Neilson, An efficient technique for magnetic analysis for non-circular, high-beta Tokamak equilibria, Nuclear Fusion 22 (8) (1982) 1015–1030.
[13] L. Lao, Separation of β_p and l_i in Tokamaks of non-circular cross-section, Nuclear Fusion 25 (11) (1985) 1421.
[14] J. Blum, Numerical Simulation and Optimal Control in Plasma Physics with Applications to Tokamaks, Series in Modern Applied Mathematics, Wiley Gauthier-Villars, Paris, 1989.
[15] F. Hofmann, G. Tonetti, Tokamak equilibrium reconstruction using faraday rotation measurements, Nuclear Fusion 28 (10) (1988) 1871.
[16] J. Christiansen, J. Taylor, Determination of current distribution in a Tokamak, Nuclear Fusion 22 (1982) 111.
[17] B. Braams, The interpretation of Tokamak magnetic diagnostics, Plasma Physics and Controlled Fusion 33 (1991) 715.
[18] M. Bishop, J. Taylor, Degenerate toroidal MHD equilibria and minimum B, Physics of Fluids 29 (1986) 1444.
[19] V. Pustovitov, Magnetic diagnostics: general principles and the problem of reconstruction of plasma current and pressure profiles in toroidal systems, Nuclear Fusion 41 (6) (2001) 721.
[20] L. Zakharov, J. Lewandoski, E. Foley, F. Levinton, H. Yuh, V. Drozdov, D. McDonald, The theory of variances in equilibrium reconstruction, Physics of Plasmas 15 (9) (2008) 092503.
[21] W. Zwingmann, L.-G. Eriksson, P. Stubberfield, Equilibrium analysis of Tokamak discharges with anisotropic pressure, Plasma Physics and Controlled Fusion 43 (11) (2001) 1441–1456.
[22] M. Beretta, M. Vogelius, An inverse problem originating from magnetohydrodynamics, Archive for Rational Mechanics and Analysis 115 (2) (1991) 137–152.
[23] M. Beretta, M. Vogelius, An inverse problem originating from magnetohydrodynamics III. Domains with corners of arbitrary angles, Asymptotic Analysis 11 (1992) 289–315.
[24] M. Vogelius, An inverse problem for the equation $\Delta u = -cu - d$, Annales Institut Fourier, Grenoble 44 (4) (1994) 1181–1209.
[25] A. Demidov, A.Y. Kochurov, A.S. Popov, To the problem of the recovery of nonlinearities in equations of mathematical physics, Journal of Mathematical Sciences 163 (1) (2009) 46–77.

[26] M. Beretta, M. Vogelius, An inverse problem originating from magnetohydrodynamics II. The case of the Grad–Shafranov equation, Indiana University Mathematics Journal 41 (1992) 1081–1118.
[27] D. O'Brien, J. Ellis, J. Lingertat, Local expansion method for fast plasma boundary identification in JET, Nuclear Fusion 33 (3) (1993) 467–474.
[28] F. Saint-Laurent, G. Martin, Real time determination and control of the plasma localisation and internal inductance in Tore Supra, Fusion Engineering and Design 56-57 (2001) 761–765.
[29] F. Sartori, A. Cenedese, F. Milani, JET real-time object-oriented code for plasma boundary reconstruction, Fusion Engineering and Design 66–68 (2003) 735–739.
[30] A. Tikhonov, V. Arsenin, Solutions of Ill-posed Problems, Winston, Washington, DC, 1977.
[31] P. Ciarlet, The Finite Element Method For Elliptic Problems, North-Holland, 1980.
[32] C. De Boor, A Practical Guide To Splines, Springer-Verlag, 1978.
[33] P. Hansen, Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems, Numerical Algorithms 20 (1999) 195–196.
[34] D. Mazon, J. Blum, C. Boulbe, B. Faugeras, A. Boboc, M. Brix, P. De Vries, S. Sharapov, L. Zabeo, Real-time identification of the current density profile in the JET Tokamak: method and validation, in: Proceedings of the 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, Shanghai, PR China, vol. WeA09.1, 2009, pp. 285–290.
[35] D. Mazon, J. Blum, C. Boulbe, B. Faugeras, A. Boboc, M. Brix, P. DeVries, S. Sharapov, L. Zabeo, Equinox: a real-time equilibrium code and its validation at JET, in: L. Fortuna, A. Fradkov, M. Frasca (Eds.), From Physics to Control Through an Emergent View, World Scientific Book Series on Nonlinear Science, Series B, vol. 15, World Scientific, 2010, pp. 327–333.

Article B : [28] D. Mazon, P. Lotte, B. Faugeras, C. Boulbe, J. Blum, F. Saint-Laurent, S. Bremond, P. Moreau, A. Murari et P. Blanchard. “Validation of the new real-time equilibrium code EQUINOX on JET and ToreSupra”. In : *Proceedings of the 39th EPS Conference and 16th Int. Congress on Plasma Physics*. Stockholm, Sweden, juil. 2012

Validation of the new real-time equilibrium code EQUINOX on JET and Tore Supra

D. Mazon[1], P. Lotte[1], B. Faugeras[2], C. Boulbe[2], J. Blum[2],
F. Saint-Laurent[1], S. Bremond[1], P. Moreau[1], A. Murari[3], P. Blanchard[4]
and EFDA JET Contributors, JET-EFDA, Culham Science Centre, Abingdon, OX14 3DB, UK*

[1] CEA, IRFM 13108 St Paul Lez Durance, France
[2] Laboratoire J-A Dieudonné (UMR 66 21), Université de Nice Sophia-Antipolis, CNRS Parc Valrose 06108 Nice Cedex 02 France
[3] Consorzio RFX – Associazione EURATOM ENEA per la Fusione, Padova, Italy
[4]JET-EFDA, Culham Science Centre, OX14 3DB, Abingdon, UK

**See the Appendix of F. Romanelli et al. Proc. 22nd IAEA Fusion Energy Conference,*

1. Introduction

The shape of the plasma current density profile, direct output of an equilibrium reconstruction, is known to play a leading role in triggering and sustaining high performance regimes. In the perspective of improving the control of these regimes, the objective is thus to develop real-time methods and algorithms that reconstruct the magnetic equilibrium in the perspective to use their outputs for feedback purposes.

The real time equilibrium reconstruction code EQUINOX, which solves the Grad Shafranov equation, has been recently rewritten and installed in both JET and Tore Supra (TS) real time control systems. This new version provides much more flexibility in terms of parameters tuning and constraints. Indeed in addition to the magnetic measurements it may consider as internal constraints MSE, polarimetry, and potentially others such as Soft X-rays measurements and/or plasma pressure profiles for magnetic axis determination. The calculation time, when internal constraints are included, is about 50ms on both machines, which is short enough to allow feed back control on the plasma current on medium and large devices.

2. Overview of the used RT resolution techniques

The problem of the equilibrium of a plasma in a Tokamak is a free boundary problem in which the plasma boundary is defined as the last closed magnetic flux surface. Inside the plasma, the equilibrium equation in an axisymmetric configuration is the Grad-Shafranov equation:

$$-\Delta^*\psi = rp'(\psi) + \frac{1}{\mu_0 r}(ff')(\psi) \quad \text{with} \quad \Delta^* = \frac{\partial}{\partial r}\left(\frac{1}{\mu_0 r}\frac{\partial}{\partial r}\right) + \frac{\partial}{\partial z}\left(\frac{1}{\mu_0 r}\frac{\partial}{\partial z}\right) \qquad (1)$$

Where μ_0 is the magnetic permeability of the vacuum, $\psi(r,z)$ the poloidal flux, r, z the Cartesian coordinates. The right hand side of this equation is a non-linear source which represents the toroidal component of the plasma current density. The goal of a real-time equilibrium code is to identify not only the plasma boundary but also the flux surface geometry outside and inside the plasma, the current density profile and derive the safety factor 'q' and other important parameters from the obtained equilibrium. In order to meet the real-time requirements, a new version of the EQUINOX [1] code has been designed and implemented in C++ using a finite element method, a non linear fixed point algorithm associated to a least square optimization procedure. Tokamak specific softwares like FELIX/XLOC [2] (or APOLO [3] at Tore Supra) provide to the EQUINOX code the boundary conditions (discrete poloidal flux values on the first wall of the vacuum vessel) in real-time. By means of least-square minimization of the difference between measurements and the simulated ones the code identifies the source term of the non linear Grad-Shafranov equation. The experimental measurements that enable the identification are the magnetics at the vacuum vessel, the interferometric and polarimetric measurements on several chords and the motional Stark effect measurements (only at JET). The finite element solver uses triangles interpolation, the calculation being limited to the vacuum chamber. A careful implementation inside the MARTe framework [4] at JET leads to execution time less than 50ms per iteration on a 2GHz PC, complemented with excellent robustness and very good precision (+/- 1cm compared to FELIX-XLOC code) of plasma boundary for an equilibrium code. Examples of reconstructed equilibria at Tore Supra and JET are provided in Fig.1:

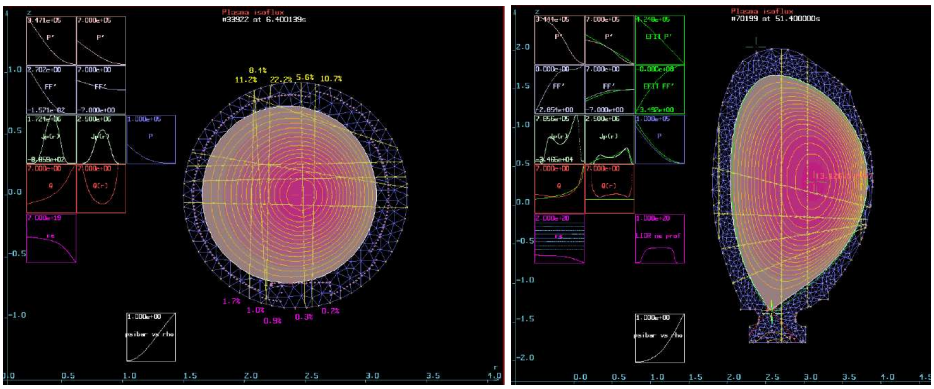

Fig.1: Examples of Equilibrium reconstruction: left Tore Supra case (#33922 at 6.4s) right JET case (#70199 at 51.4s).

3. Code validation at JET

Using a validated database of 150 pulses (shots with or without the new ITER Like Wall) well representative of JET operational space ($1.12<I_p<3.09$MA, $1.68<B_T<3.42$T, $0.06<\delta<0.51$), EQUINOX has been first fully and carefully benchmarked against the

online plasma boundary shape reconstruction code XLOC, the off line equilibrium code EFIT [5] and MHD signatures. Statistical analysis confirmed the relevance of the EQUINOX reconstruction (Fig 2) for the reconstruction of global parameters.

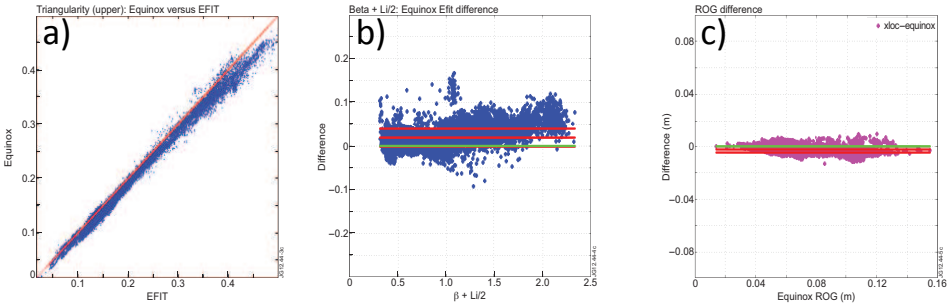

Fig.2 Statistical comparisons between EQUINOX and EFIT [5] for a) triangularity b) Shafranov Shift and XLOCc) Right Outer Gap (ROG)). Horizontal green line zero reference, horizontal red lines standard deviations

Validation has also been performed on specific shots to check the dynamical response of the code but also to validate the accuracy of the reconstruction when internal measurements are used (Fig.3).

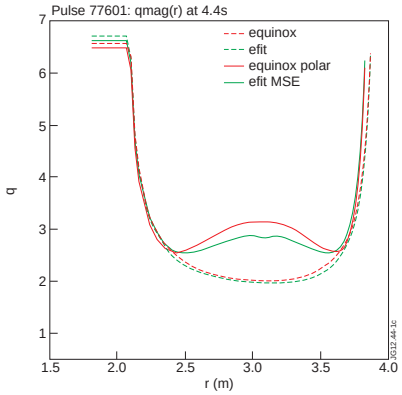

Fig 3 Comparison between EQUINOX and EFIT of q profile magnetic only (dotted lines), polarimetry and MSE (green plain lines) #77601, Ip=1.7MA BT=2.6T, 3MW LHCD (Lower Hybrid Current Drive), 6MW ICRH (Ion Cyclotron Resonance Heating), 20 MW NBI.

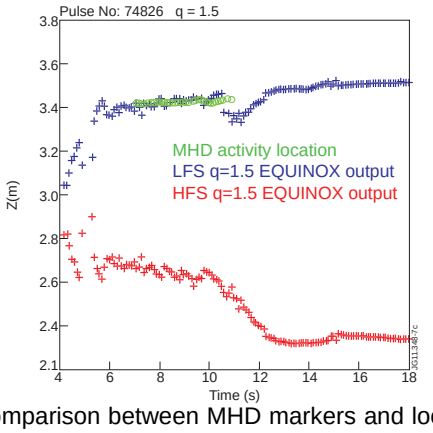

Fig 4 Comparison between MHD markers and location of q=1.5 (low and high field side) obtained in real time from EQUINOX (constrained with polarimetry), (#74826, 19MW NBI, I_p=1.6MA, B_T=2T)

Independent analysis of the database provides identification of MHD mode and their location. Fig. 4 shows the perfect agreement between EQUINOX and mode location (q=1.5) identified from Electron Cyclotron Emission (ECE) and magnetic measurements.

4. Code validation at Tore Supra

The validation of Equinox on Tore Supra has started and will follow the same methodology as JET. The new version of Equinox takes into account the polarimetry data. Indeed, this diagnostic is of crucial importance at Tore Supra where shots can last several minutes, these durations being presently much too large for continuous MSE measurements. Equinox input parameters have been tuned by calculating

plasma equilibria for some typical shots of the last campaign, and compared with EFIT and with the current diffusion code CRONOS calculations. They have also been compared with results from APOLO code that controls the plasma position in real time, taking information from the poloïdal generators and the magnetic diagnostics. Figure 5.a shows an example of q profiles obtained by EQUINOX, EFIT and CRONOS for a sawtooth discharge with 5MW of Ion Cyclotron Radiofrequency Heating. EQUINOX and EFIT both using polarimetry are in a very good agreement, whereas slight differences can be seen with CRONOS, but the difference looks reasonable since these codes are based on different principles. Figure 5.b shows the evolution of the rational q surfaces position with time for the 3 calculations, still in good agreement. When possible the comparison with MHD information is performed. For instance in this figure, the sawtooth inversion radius derived from the ECE diagnostic is indicated. The tuning of EQUINOX now needs to be tested on a larger database of shots, and this code will be available for the next campaign, the new q profile control algorithm tools being developed in parallel.

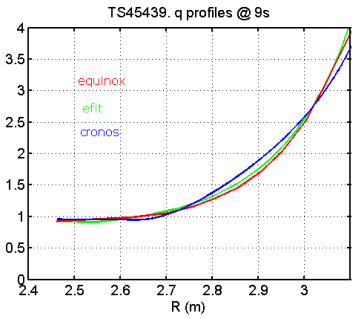

Fig 5.a Comparison of q profiles obtained by EQUINOX, EFIT, and the current diffusion code CRONOS

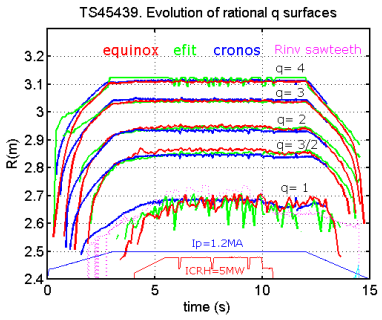

Fig 5.b Comparison of rational q surfaces evolution with time for EQUINOX, EFIT and CRONOS. The sawtooth inversion radius is indicated.

5. Conclusions and perspectives

The EQUINOX code is now available in real time in both JET and Tore Supra tokamaks and will be used for q profile feedback control experiments. The full validation of the real time reconstruction provides now a good base for real-time control but more generally systematic physics analysis. This code is also available inside the Integrated Tokamak Modelling platform which makes EQUINOX a potentially very powerful tool to predict equilibrium and current profile evolution in ITER or DEMO.

References

[1] J. Blum et al, 2012 JCP **231** 960-980
[2] F. Sartori et al, 2003 Fusion Eng Design **66-68** 735
[3] F. Saint-Laurent et al, 2009, Proceedings of the 12th Int Conf on Accelerator and large Physics Control Systems Kobe Japan
[4] A. Neto et al, 2010, IEEE Transactions **57** 479
[5] L. Lao, 1990 Nuclear Fusion **30** 1035
[6] J.F. Artaud *et al,* 2010 *Nucl. Fusion* **50** 043001

Acknowledgements

This work was supported by EURATOM and carried out within the frameworkof the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

Article C : [7] B. Faugeras, J. Blum, C. Boulbe, P. Moreau et E. Nardon. 2D interpolation and extrapolation of discrete magnetic measurements with toroidal harmonics for equilibrium reconstruction in a Tokamak. *Plasma Phys. Control Fusion* 56 (2014), p. 114010

IOP Publishing

Plasma Physics and Controlled Fusion

Plasma Phys. Control. Fusion **56** (2014) 114010 (11pp)

doi:10.1088/0741-3335/56/11/114010

2D interpolation and extrapolation of discrete magnetic measurements with toroidal harmonics for equilibrium reconstruction in a tokamak

Blaise Faugeras[1], **Jacques Blum**[1], **Cedric Boulbe**[1], **Philippe Moreau**[2] **and Eric Nardon**[2]

[1] Laboratoire J.A. Dieudonné, UMR 7351, Université Nice Sophia-Antipolis, Parc Valrose, 06108 Nice Cedex 02, France

[2] CEA, IRFM, F-13108, Saint-Paul-lez-Durance, France

E-mail: Blaise.Faugeras@unice.fr

Received 11 September 2013, revised 15 November 2013

Accepted for publication 21 November 2013

Published 17 October 2014

Abstract

We present a method based on the use of toroidal harmonics and on a modelization of the poloidal field coils and divertor coils for the 2D interpolation and extrapolation of discrete magnetic measurements in a tokamak. The method is generic and can be used to provide the Cauchy boundary conditions needed as input by a fixed domain equilibrium reconstruction code like Equinox (Blum *et al* 2012 *J. Comput. Phys.* **231** 960–80). It can also be used to extrapolate the magnetic measurements in order to compute the plasma boundary itself. The proposed method and algorithm are detailed in this paper and results from numerous numerical experiments are presented. The method is foreseen to be used in the real-time plasma control loop on the WEST tokamak (Bucalossi *et al* 2011 *Fusion Eng. Des.* **86** 684–8).

Keywords: tokamak, plasma equilibrium, plasma boundary, toroidal harmonics, magnetic measurements, inverse problem

(Some figures may appear in colour only in the online journal)

1. Introduction

Equilibrium reconstruction codes are fundamental for the analysis and the control of fusion experiments in a tokamak [3]. The state variable of interest in the modelization of such an equilibrium under the usual axisymmetric assumption is the poloidal flux $\psi(r, z)$, which is related to the poloidal magnetic field by the relation $B = (1/r)(-\partial_z\psi, \partial_r\psi)$ in the cylindrical coordinate system (r, ϕ, z). The basic inputs used to achieve the numerical reconstruction of the equilibrium are magnetic measurements taken at several locations surrounding the vacuum vessel.

Basically equilibrium codes can be of two types. The first one is the full domain type in which the reconstruction is performed in the whole right-half plane ($r > 0$) and relies on the use of Green's functions. A drawback of this method is that nonlinear ferromagnetic structures, which can be present in certain tokamaks, are complicated to deal with. An iron model has to be introduced [4] and these codes can hardly run in real time. The second type of code is the bounded domain one in which computations are performed in a fixed domain containing the plasma but restricted to a limited region. A difficulty is that Cauchy boundary conditions ($g = \psi$, $h = \partial_n\psi$) have to be provided on a fixed closed contour defining the boundary of the computation domain (this contour, called Γ in the remaining part of this paper, can link some of the B probes, for example). These boundary conditions have to be computed from the discrete magnetic measurements. If these measurements are numerous enough and regularly located on a smooth contour, a direct linear interpolation can be considered [5, 6]. However, this method is not robust in case of defective sensors. Morevover, it cannot be used in today's machines like JET, in which the measurement points

are scattered in some annular region surrounding the vacuum vessel. Another approach that can be considered is to use the plasma boundary identification code of a particular machine if it exists. Such a code will, in fact, compute the poloidal flux ψ in the vacuum surrounding the plasma and can be used to evaluate ψ and its normal derivative on any given fixed closed contour Γ. This is the approach followed until now for the equilibrium reconstruction code Equinox [1]. The computations rely on the boundary conditions g and h provided by the plasma boundary reconstruction codes Xloc at JET [7, 8] and the Apolo code at Tore Supra [9]. The main drawback of this approach is that these two boundary reconstruction codes are extremely machine dependent and are not transportable to a generic platform such as the ITM [10]. Moreover, concerning the particular case of Tore Supra, a new numerical method has to be developed since the machine is going to be upgraded to WEST [2].

The aim of this paper is to investigate the possibility of using toroidal harmonics to perform the 2D interpolation and extrapolation of magnetic measurements in an annular domain surrounding the plasma to compute Cauchy boundary conditions on a given fixed closed contour Γ in order to be able to run in a second step an equilibrium reconstruction code such as Equinox in the bounded domain limited by Γ. In fact, at a given instant in time, the fictitious inner boundary of the annular domain could be defined as being the plasma boundary itself, and the data interpolation problem is very closely connected to the ill-posed inverse problem of the identification of the plasma boundary. The latter is a Cauchy problem for the elliptic equation $\Delta^*\psi = 0$, and various solution methods have been proposed to solve it, to compute ψ in the vacuum surrounding the plasma and to identify the plasma boundary (see [11] for a review). The ill-posed nature of the problem usually imposes the use of a regularization technique and an *a priori* representation of plasma current density and the flux it generates, called the internal solution, for example using filaments of current or a fictitious current sheet, or also a decomposition in toroidal harmonics. The latter seems particularly attractive since these functions provide explicit solutions to the equation $\Delta^*\psi = 0$. Toroidal harmonics [12, 13] were used in a number of papers in the plasma physics literature in the 1980s and 1990s [14–20] and more recently in [21].

Apart from [16, 21], authors using toroidal harmonics do not use any regularization procedure. Our point of view is that the small number of harmonics needed to represent the flux in the vacuum is in itself a regularizing procedure. Our numerical experiments confirm this point. In fact the number of toroidal harmonics used to represent the internal solution in some way can be seen as the regularization parameter. A too small number might lead to a smooth solution that possibly does not fit the data very well and does not give a very accurate plasma boundary, whereas a too large number might lead to an irregular solution that fits the data well but gives an irregular plasma boundary.

Another ingredient of the method to which the solution is quite sensitive is the location of the pole of the toroidal coordinates system. In fact, together with the number of internal toroidal harmonics it is the only parameter of the internal solution that can be tuned, and curiously it is generally kept fixed in the literature apart from in [19], in which an optimization method is proposed to identify a proper location of the pole. In this paper we propose two simple methods to do so.

Finally, in order to represent the flux ψ in the vacuum, some authors use a pure decomposition in toroidal harmonics [14–16, 20] whereas others add a term coming from the modelization of the flux generated by the poloidal field coils [17, 19]. In this paper, we discuss this point together with the impact of the presence of nonlinear ferromagnetic material. Our numerical experiments show that it is important to take into account the divertor coils.

The paper is organized as follows. In the next section we introduce notations for a number of domains and contours that are needed. Section 3 deals with the decomposition of the flux in toroidal harmonics. In section 4, the proposed numerical algorithm is presented, and in the last section a number of numerical results are presented.

2. Mathematical setting

In order to get into the details of this work we first need to briefly recall the equilibrium reconstruction problem and introduce a number of contours and domains. Therefore, a schematic representation of a poloidal cross-section of a tokamak is shown in figure 1 and is described below.

The unknown plasma free boundary domain is noted Ω_p. The plasma boundary is the isoflux line whose value is defined either by the contact with the limiter or as a magnetic separatrix (a hyperbolic line with an X-point represented by the dashed line inside the limiter contour in figure 1). Poloidal field coils and divertor coils are denoted as Ω_{C_k}. Poloidal field probes measure the local value of the poloidal magnetic field, and flux loops and saddle loops measure the local value of the flux ψ. Flux and saddle loops, represented by triangles, and B probes, represented by cross-circles, are shown surrounding the limiter contour. All these measurement points can be included in a fictitious annular domain D, which neither contains the coils nor the plasma. The inner boundary of D can, for example, be chosen to be the limiter contour. The presence of divertor coils can impose the choice of a somehow tortured outer boundary (the dashed line labelled $\partial\Omega$). The outer boundary of D also defines a domain Ω including D, Ω_p and the vacuum region lying between the plasma and D. Eventually all these different domains will be included in the larger domain Ω_0, outside of which nonlinear magnetic material like iron might be present. Its boundary is noted $\partial\Omega_0$.

Depending on the domain, the poloidal flux satisfies the partial differential equation

$$-\Delta^*\psi = 0 \qquad \text{in } D \tag{1}$$

or

$$-\Delta^*\psi = j_p(\psi, r)\chi_{\Omega_p} \qquad \text{in } \Omega \tag{2}$$

or

$$-\Delta^*\psi = j_p(\psi, r)\chi_{\Omega_p} + \sum_k j_{C_k}\chi_{\Omega_{C_k}} \qquad \text{in } \Omega_0 \tag{3}$$

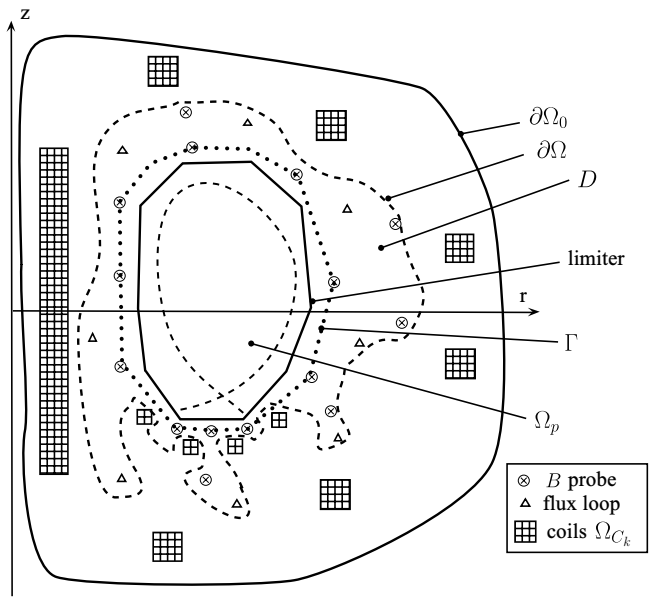

Figure 1. Schematic representation of a poloidal cross-section of a tokamak. See the text for details.

where the differential operator

$$\Delta^* = \frac{\partial}{\partial r}\left(\frac{1}{\mu_0 r}\frac{\partial}{\partial r}\right) + \frac{\partial}{\partial z}\left(\frac{1}{\mu_0 r}\frac{\partial}{\partial z}\right),$$

is linear, the right-hand side represents the toroidal component of the local current density and χ is the indicator function. If iron is present outside Ω_0 (like for JET and Tore Supra), μ_0 is no longer a constant but a function of $|B|$ and the operator Δ^* becomes nonlinear. This is why we restrict ourselves to Ω_0 where Δ^* is linear. In $\Omega_0 \setminus \{\Omega_p \bigcup \Omega_{C_k}\}$, the current is null. In the coils Ω_{C_k}, the densities j_{C_k} are measured and known. In the plasma Ω_p, the current density is unknown but according to the Grad–Shafranov equation takes the form

$$j_p(\psi, r) = r p'(\psi) + \frac{1}{\mu_0 r}(ff')(\psi) \tag{4}$$

in which p' and ff' are unknown functions to be identified by an equilibrium reconstruction code.

As explained in the introduction, one of our goals is to compute Cauchy conditions ($g = \psi$, $h = \partial_n \psi$) on the contour Γ in order to provide inputs to the reconstruction code Equinox. Indeed, let us recall that this code solves the following problem:

find functions A and B defined on $[0, 1]$ which minimize the following regularized cost function

$$J(A, B) = \int_\Gamma (\partial_n \psi - h)^2 \, \mathrm{d}s + \epsilon \left(\int_0^1 (A''(x))^2 + (B''(x))^2 \, \mathrm{d}x \right)$$

where ψ satisfies

$$-\Delta^* \psi = \lambda \left(\frac{r}{r_0} A(\bar{\psi}) + \frac{r_0}{r} B(\bar{\psi}) \right) \chi_{\Omega_p(\psi)}, \qquad \text{in } D_\Gamma$$
$$\psi = g \qquad \text{on } \Gamma.$$

Here λ is a scaling factor, r_0 is a constant, A and B are related to p' and ff', $\bar{\psi}$ is a normalized flux and D_Γ is the domain contained inside Γ. Equinox implements a finite element discretization method and identifies the full equilibrium (plasma boundary and current density) in the fixed domain D_Γ. This computation completely relies on the boundary conditions g and h deduced from the magnetic measurements.

3. Decomposition of the poloidal flux ψ in the annular domain *D*

In this section we recall the principle of the decomposition of the flux in toroidal harmonics in the region D. Moreover, we show that the nonlinearity induced by the presence of iron outside Ω_0 does not restrict the possibility of using a modelization of the flux generated by the different coils.

3.1. Toroidal harmonics

The toroidal coordinates system [13, 22] or bipolar coordinates system (if we ignore the angular toroidal variable) $(\zeta, \eta) \in \mathbb{R}_*^+ \times [0, 2\pi]$ about the pole $F_0 = (r_0, z_0)$ is related to the cylindrical coordinates system (r, z) by

$$r = \frac{r_0 \sinh\zeta}{\cosh\zeta - \cos\eta} \qquad \text{and} \qquad z - z_0 = \frac{r_0 \sin\eta}{\cosh\zeta - \cos\eta}.$$

In what follows, we assume that F_0 lies inside the region surrounded by the annular domain D and more precisely inside the plasma domain where the homogeneous equation, $-\Delta^*\psi = 0$, is not satisfied. In the domain D, this equation is satisfied. It is known that explicit solutions to this equation in an annular domain can be found in toroidal coordinates using a quasi separation of variables technique (see [23, 24] for details on the computations). Moreover the family of solutions found is complete [21, 24]. That is to say that, given any regular enough Dirichlet boundary condition u on ∂D, the solution to the boundary value problem

$$\begin{cases} -\Delta^*\psi = 0 & \text{in } D \\ \psi = u & \text{on } \partial D \end{cases} \tag{5}$$

can be uniquely decomposed as

$$\begin{cases} \psi = \psi_{\text{ext}} + \psi_{\text{int}} \\ \psi_{\text{ext}} = \dfrac{r_0 \sinh\zeta}{\sqrt{\cosh\zeta - \cos\eta}} \\ \quad \times \left[\displaystyle\sum_{n=0}^{\infty} a_n^e Q^1_{n-1/2}(\cosh\zeta)\cos(n\eta) \right. \\ \quad \left. + \displaystyle\sum_{n=1}^{\infty} b_n^e Q^1_{n-1/2}(\cosh\zeta)\sin(n\eta) \right] \\ \psi_{\text{int}} = \dfrac{r_0 \sinh\zeta}{\sqrt{\cosh\zeta - \cos\eta}} \\ \quad \times \left[\displaystyle\sum_{n=0}^{\infty} a_n^i P^1_{n-1/2}(\cosh\zeta)\cos(n\eta) \right. \\ \quad \left. + \displaystyle\sum_{n=1}^{\infty} b_n^i P^1_{n-1/2}(\cosh\zeta)\sin(n\eta) \right] \end{cases} \tag{6}$$

where $P^1_{n-1/2}$ and $Q^1_{n-1/2}$ are the associated Legendre functions of first and second kind, of degree one and half integer order [25], also called toroidal harmonics when evaluated at point $\cosh\zeta$. Functions $P^1_{n-1/2}$ have a singularity when $\zeta \to \infty$; that is to say at point F_0 and therefore ψ_{int} represents the flux generated by currents flowing inside D. On the contrary, functions $Q^1_{n-1/2}$ are singular when $\zeta \to 0$, that is to say on the axis $r = 0$, and therefore ψ_{ext} represents the flux generated by currents flowing outside D.

3.2. Including information from the knowledge of the currents in the coils

Let us denote by $x = (r, z)$ a point in the poloidal plane. For any scalar fields ψ and ϕ on Ω, two integrations by parts of the quantity $\phi\Delta^*\psi$ lead to the so called Green's second identity (or Green's theorem)

$$\int_\Omega (\phi\Delta^*\psi - \psi\Delta^*\phi)\,\mathrm{d}x = \int_{\partial\Omega} \frac{1}{\mu_0 r}\left(\frac{\partial\phi}{\partial n}\psi - \frac{\partial\psi}{\partial n}\phi\right)\mathrm{d}s. \tag{7}$$

Let $\bar{x} \in D$ and $G(x, \bar{x})$ be the free space Green's function which satisfies $-\Delta^* G(x, \bar{x}) = \delta(x - \bar{x})$ in the whole half plane $r > 0$ and $G(x, \bar{x}) \to 0$ when $|x| \to \infty$ or $r \to 0$.

The important point here is that even if the region external to Ω_0 contains nonlinear materials such as iron, the restriction of G to Ω can still be used as the function ϕ in equation (7). If ψ is chosen to be the solution to equation (2), one gets

$$\begin{aligned}\psi(\bar{x}) = &\int_{\Omega_p} j_p(\psi(x), r)G(x, \bar{x})\,\mathrm{d}x \\ &+ \int_{\partial\Omega} \frac{1}{\mu_0 r}\left(\frac{\partial G}{\partial n}(x, \bar{x})\psi(x) - \frac{\partial\psi}{\partial n}(x)G(x, \bar{x})\right)\mathrm{d}s\end{aligned} \tag{8}$$

and this leads again to a decomposition of the type $\psi = \psi_{\text{int}} + \psi_{\text{ext}}$ with

$$\begin{cases}\psi_{\text{int}}(\bar{x}) = \int_{\Omega_p} j_p(\psi(x), r)G(x, \bar{x})\,\mathrm{d}x \\ \psi_{\text{ext}}(\bar{x}) = \displaystyle\int_{\partial\Omega} \frac{1}{\mu_0 r}\left(\frac{\partial G}{\partial n}(x, \bar{x})\psi(x) - \frac{\partial\psi}{\partial n}(x)G(x, \bar{x})\right)\mathrm{d}s\end{cases} \tag{9}$$

ψ_{int} is another expression for the flux generated by currents running inside D and ψ_{ext} for those running outside. Moreover Green's theorem can also be applied in Ω_0 (the region including the plasma and the coils). One then gets the following expression for ψ in D: $\psi = \psi_{\text{int}} + \psi_{\text{ext}} + \psi_C$ with

$$\begin{cases}\psi_{\text{int}}(\bar{x}) = \int_{\Omega_p} j_p(\psi(x), r)G(x, \bar{x})\,\mathrm{d}x \\ \psi_{\text{ext}}(\bar{x}) = \displaystyle\int_{\partial\Omega_0} \frac{1}{\mu_0 r}\left(\frac{\partial G}{\partial n}(x, \bar{x})\psi(x) - \frac{\partial\psi}{\partial n}(x)G(x, \bar{x})\right)\mathrm{d}s \\ \psi_C(\bar{x}) = \displaystyle\sum_k \int_{C_k} j_{C_k} G(x, \bar{x})\,\mathrm{d}x\end{cases} \tag{10}$$

where ψ_C represents the contribution of the coils to the total flux. In the annular domain D, $\tilde{\psi} = \psi - \psi_C = \psi_{\text{int}} + \psi_{\text{ext}}$ still satisfies

$$\begin{cases}-\Delta^*\tilde{\psi} = 0 & \text{in } D \\ \tilde{\psi} = \psi|_{\partial D} - \psi_C|_{\partial D} & \text{on } \partial D\end{cases} \tag{11}$$

and can thus be decomposed in toroidal harmonics.

This shows that the knowledge of the currents j_{C_k} in the coils can be used in the representation of the flux in the region D in the presence or absence of iron outside Ω_0. In fact, if the coils are located very close to the measurement points such as the divertor coils, it is necessary to modelize them. Their contribution to the flux in D can theoretically be written as a series of toroidal harmonics, but many of them are needed in practice. This can be critical compared to the number of measurements, and the numerical resolution of the problem might become difficult.

4. Numerical method

Let us now present the numerical method which we implemented. At each discrete time step during a discharge, the magnetic measurements available are of three types:

- Flux loops provide N_f flux measurements at points x_i^f such that $\psi_i^{\text{meas}} \approx \psi(x_i^f)$;
- Saddle loops provide N_s flux variation measurements between two points such that $\delta_i\psi^{\text{meas}} \approx \psi(x_i^1) - \psi(x_i^2)$;
- B probes provide N_B measurements of the poloidal field at points x_i^B and directions d_i such that $B_i^{\text{meas}} \approx B(x_i^B).d_i$.

The first step of the algorithm consists in subtracting from the measurements a numerical approximation of the contribution from the coils.

$$\begin{cases}\tilde{\psi}_i^{\text{meas}} = \psi_i^{\text{meas}} - \hat{\psi}_C(x_i^f), \quad \text{for } i = 1, \dots N_f \\ \delta_i\tilde{\psi}^{\text{meas}} = \delta_i\psi^{\text{meas}} - (\hat{\psi}_C(x_i^1) - \hat{\psi}_C(x_i^2)), \\ \quad \text{for } i = 1, \dots N_s \\ \tilde{B}_i^{\text{meas}} = B_i^{\text{meas}} - \hat{B}_C(x_i^B).d_i, \quad \text{for } i = 1, \dots N_B\end{cases} \tag{12}$$

Here the contribution from each coil C_k is computed as follows. The known current density is given as $j_{C_k} = I_k/S_k$, where I_k is the total current in the coil and S_k its surface. The coil is divided into n_k subcoils $C_{k,l}$ of equal surface S_k/n_k and center $c_{k,l}$ on which the integrals are numerically evaluated as

$$\int_{C_{k,l}} \frac{I_k}{S_k} G(x, \bar{x})\,\mathrm{d}x \approx \frac{I_k}{n_k} G(c_{k,l}, \bar{x}). \tag{13}$$

This in fact consists of considering the contribution of the coils as a sum of the contributions from filaments of current

$$\hat{\psi}_C(\bar{x}) = \sum_k \sum_l \frac{I_k}{n_k} G(c_{k,l}, \bar{x}). \tag{14}$$

The second step consists of truncating the toroidal harmonics expansion of $\tilde{\psi}$ to approximate it by

$$\begin{cases} \hat{\psi} = \hat{\psi}_{\text{ext}} + \hat{\psi}_{\text{int}}, \\ \hat{\psi}_{\text{ext}} = \dfrac{r_0 \sinh\zeta}{\sqrt{\cosh\zeta - \cos\eta}} \\ \quad \times \Big[\displaystyle\sum_{n=0}^{n_a^e} a_n^e Q^1_{n-1/2}(\cosh\zeta)\cos(n\eta) \\ \quad + \displaystyle\sum_{n=1}^{n_b^e} b_n^e Q^1_{n-1/2}(\cosh\zeta)\sin(n\eta) \Big] \\ \hat{\psi}_{\text{int}} = \dfrac{r_0 \sinh\zeta}{\sqrt{\cosh\zeta - \cos\eta}} \\ \quad \times \Big[\displaystyle\sum_{n=0}^{n_a^i} a_n^i P^1_{n-1/2}(\cosh\zeta)\cos(n\eta) \\ \quad + \displaystyle\sum_{n=1}^{n_b^i} b_n^i P^1_{n-1/2}(\cosh\zeta)\sin(n\eta) \Big] \end{cases} \tag{15}$$

and to evaluate each of the terms in the expansion of $\hat{\psi}$ and of the associated field $\hat{B}$ at the different measurement points in order to form a least squares cost function

$$J(u) = \sum_{i=1}^{N_f} \frac{(\hat{\psi}_i(u) - \tilde{\psi}_i^{\text{meas}})^2}{\sigma_f^2} + \sum_{i=1}^{N_s} \frac{(\delta_i\hat{\psi}(u) - \delta_i\tilde{\psi}^{\text{meas}})^2}{\sigma_s^2} + \sum_{i=1}^{N_B} \frac{(\hat{B}_i(u) - \tilde{B}_i^{\text{meas}})^2}{\sigma_B^2} \tag{16}$$

depending on

$$u = (a_0^e, \ldots, a_{n_a^e}^e, b_1^e, \ldots, b_{n_b^e}^e, a_0^i, \ldots, a_{n_a^i}^i, b_1^i, \ldots, b_{n_b^i}^i)$$

the unknown coefficients of the expansion in toroidal harmonics. The weights σ_f, σ_s and σ_B correspond to the assumed measurement errors. J is quadratic in u and is minimized by solving the associated normal equation to find the optimal set of coefficients u_{opt}.

In these computations the expressions for $\hat{\psi}_C$ and $\hat{B}_C$ are explicit [26]. The expression for $\hat{B}$ is also explicit. The numerical evaluation of half-integer order associated Legendre functions is not straightforward. We use the algorithm and the computer routine DTORH1 provided with [27]. This code enables an accurate and fast evaluation of the set $P^m_{n-1/2}(x)$, $Q^m_{n-1/2}(x)$ for real $x > 1$, integers $m \geqslant 0$ and $n = 0, \ldots N$.

Once u_{opt} is computed, an approximation of the flux can be obtained at any point of the vacuum surrounding the plasma by $\psi(x) = \hat{\psi}(x) + \hat{\psi}_C(x)$. In particular, one can evaluate ψ and its normal derivative on a fixed closed contour Γ in order to provide Cauchy boundary conditions to a fixed bounded domain equilibrium reconstruction code. Of course one can also identify the plasma boundary as the largest closed flux surface inside the limiter contour.

Such a procedure provides meaningful results if the pole F_0 lies inside the unknown plasma region and not too close to the boundary. The most natural choice is to put the pole at the location of the magnetic axis, but as the plasma boundary it is unknown, we propose the following procedure. At the first time step, the pole is located at $(r_0, 0)$ where r_0 is the major radius of the tokamak. Then at time step t^{n+1}, the pole is located at the position of the magnetic axis computed at the previous time step t^n. This magnetic axis position is computed exactly if an equilibrium reconstruction code like Equinox is run at each time step. If this is not the case and one is only interested in the plasma boundary identification problem, or by an equilibrium reconstruction at a given instant in time, it can be approximated by the current center (r_c, z_c) defined as moments of the plasma current density [11, 28]. These quantities can be precisely computed as integrals on the contour Γ at every point of which the flux ψ and the field B can be evaluated:

$$I_p := \int_{D_\Gamma} j_p \,\mathrm{d}x = \int_\Gamma \frac{1}{\mu_0} B_s \,\mathrm{d}s,$$

$$z_c I_p := \int_{D_\Gamma} z j_p \,\mathrm{d}x = \int_\Gamma \frac{1}{\mu_0}(-r\log r B_n + z B_s)\,\mathrm{d}s,$$

$$r_c^2 I_p := \int_{D_\Gamma} r^2 j_p \,\mathrm{d}x = \int_\Gamma \frac{1}{\mu_0}(2rz B_n + r^2 B_s)\,\mathrm{d}s.$$

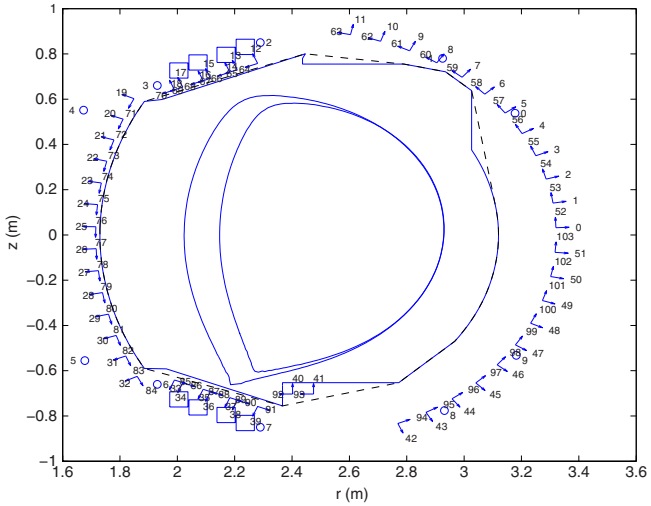

Figure 2. Poloidal section of the WEST tokamak. The two plasma boundaries correspond to case 1 (large plasma) and case 2 (smaller plasma). The B probes represented by arrows are numbered from 0 to 103 and the flux loops represented by small circles are numbered from 0 to 9. The four bottom divertor coils are shown as well as the top ones. The limiter contour is also plotted as well as its convex hull (dashed line), which will be used as the contour Γ, that is to say the boundary of the computation domain for the equilibrium reconstruction code Equinox.

5. Numerical results

5.1. Twin experiments for WEST

In view of the upgrade of the tokamak Tore Supra to WEST, we have conducted several numerical experiments to test the method. The code Cedres++ [29] is run to simulate four WEST equilibria. In the first case the X-point position is very close to the limiter whereas in the second one the plasma is smaller. Configurations 3 and 4 are limiter configurations. From these simulations the equivalent of magnetic measurements are extracted: 10 flux loops measurements and 104 Bprobes measurements (see figure 2). The reconstruction of the plasma

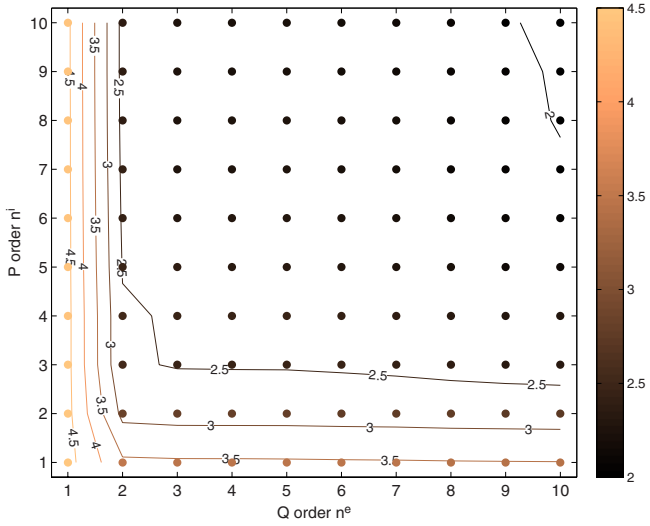

Figure 3. Contour and scatter plot of $\log(J(u_{\text{opt}}))$ as a function of the maximum order of the associated Legendre functions of first kind n^i and second kind n^e used for the representation of the flux.

boundary and the equilibrium can then be performed using these measurements as inputs, as well as the currents running in the coils. Using the notations of equation (16), we take $N_f = 10$, $\sigma_f = 10^{-3}\left(\frac{\text{Wb}}{2\pi}\right)$ and $N_B = 104$, $\sigma_B = 10^{-3}$ (T).

Let us first concentrate on the reconstruction of the flux in a vacuum for case 1 with the algorithm using toroidal harmonics (TH). Unless specified, we always take into account the values of the currents in the coils. Here we want to reconstruct a single equilibrium and thus do not have *a priori* at our disposal the knowledge of the magnetic axis position. As described in section 4, we proceed in two steps. First, we run TH setting the pole of the toroidal coordinate system to $P_0 = (r_0, 0)$, compute the current center $P_1 = (r_c, z_c)$ and re-run TH setting the pole to these new coordinates. This mimics the fact that during a whole pulse reconstruction the current center at the previous time step would play the role of P_0, and if P_1 is too far from P_0 then the pole of the coordinate system is modified.

A first natural question which has to be answered is how many toroidal harmonics should be used to represent the flux ψ in the vacuum. In all the computations we choose the maximum order of the toroidal harmonics used in equation (15) to be $n^e := n^e_a = n^e_b$ and $n^i := n^i_a = n^i_b$. From figure 3 it appears that the value of the cost function at the optimal point, which is an indicator of the quality of the fit to the measurements, decreases very rapidly as we increase the maximum order from $n^e = n^i = 1$ to 4. Above this value, the benefit of adding new degrees of freedom is much less significant and the plot shows an almost flat region for orders greater than 4.

As a consequence we make the choice $n^e = n^i = 4$. This corresponds to the minimum number of degrees of freedom needed to obtain a good fit to the measurements. Numerical values for the optimal cost and corresponding root mean square (rms) errors are given for different choices of n^e and n^i in table 1. The corresponding computed plasma boundaries are also shown in figure 4. As already mentioned, adding interior functions (column (4, 9)) or exterior functions (column

Table 1. Minimization results for the default choice $(n^e = 4, n^i = 4)$, as well as choices (4, 9), (9, 4) and (30, 4) without using any representation of the flux generated by the divertor and poloidal field coils (no c).

(n^e, n^i)	(4, 4)	(4, 9)	(9, 4)	(30, 4) no c
cost $J(u_{\text{opt}})$	2.783e + 02	1.666e + 02	2.004e + 02	3.352e + 02
rms B (T)	1.614e−03	1.233e−03	1.343e−03	1.624e−03
rms Flux $\left(\frac{\text{Wb}}{2\pi}\right)$	5.814e−04	6.663e−04	7.832e−04	1.657e−03

(9, 4)) does not significantly modify the rms. However, in the first case it deteriorates the plasma boundary reconstruction (figure 4). This is due to the fact that interior functions are involved with the ill-posed character of the inverse boundary reconstruction problem. The only regularization mechanism lies in the small number of toroidal harmonics used to represent the flux. The blow up of the interior harmonics accentuates with their order and the zone where the computed solution is not relevant spreads around the pole of the coordinate system, even reaching the plasma boundary in this case. This phenomenon disappears in the plasma boundaries computed by Equinox in all cases (see figure 5) which is due to the fact that the reconstruction of the boundary is not an ill-posed inverse problem in Equinox in which the equation for ψ is solved also in the plasma and the free boundary problem is a particular nonlinearity of the model.

The last column of table 1 shows the interest of using a modelization of the flux generated by the divertor and poloidal field coils. If we do not use this information in this particular case, a value of n^e of at least 30 has to be taken to achieve a fit to the measurements comparable to the one obtained with the choice (4, 4).

Figure 6 shows the fit to the measurements for the choice $n^e = n^i = 4$. It appears that the largest errors on the B probes measurements happen for those in the range 32–40 and 84–92, which correspond to the ones located close to the X-point. These numerical experiments therefore suggest that if some sensors could be added to the design of WEST, it would desirable to put them, if possible, in the region of the divertor.

From table 2 it can be seen that the reconstruction of the X-point position is quite accurate (up to a few mm) with the default choice $(n^e = 4, n^i = 4)$. More interestingly, it is also accurate with the choice (4, 9) where the plasma boundary shows some oscillations. This is still true for many other choices of (n^e, n^i) and is thus satisfying because it makes the determination of the X-point only very slightly dependant on the tuning of the TH algorithm. Table 2 also shows that the magnetic axis computed by Equinox is very close to the one given by Cedres, and that the computed pole for the toroidal coordinate system is also a good approximation of the magnetic axis position.

Finally, in order to get some insight into the impact of a noisy sensor on the reconstruction of the X-point position with the TH algorithm, we have conducted 104 + 10 numerical reconstructions, each time applying an offset on a different sensor. The results are displayed in figure 7. Adding an offset of 10 mT on a B-coil or of 200 mWb on a flux loop perturbates the X-point position by about a maximum of 1 mm. Again,

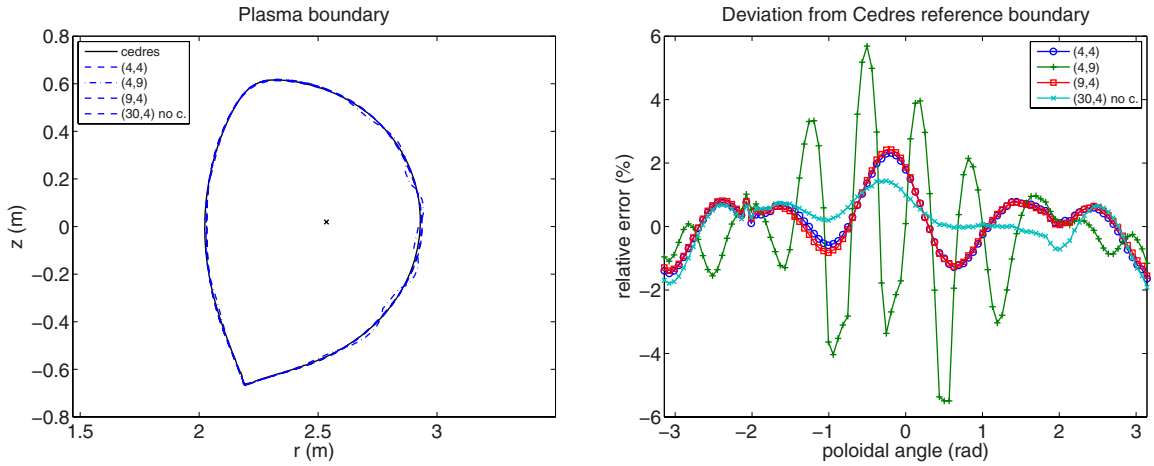

Figure 4. Left: plasma boundaries. Boundaries computed with $(n^e = 4, n^i = 4)$ or $(n^e = 9, n^i = 4)$ and $(n^e = 30, n^i = 4)$ without any PF coils modelization (no c) are almost superimposed with the reference boundary computed with Cedres. The boundary computed with $(n^e = 4, n^i = 9)$ shows some irregularity. Right: corresponding relative deviation from the Cedres++ boundary, $100(\rho - \rho_{\text{Cedres}})/\rho_{\text{Cedres}}$ as a function of the poloidal angle θ. The center of the polar coordinate system (ρ, θ) is the magnetic axis from Cedres++ (shown in the left-hand figure).

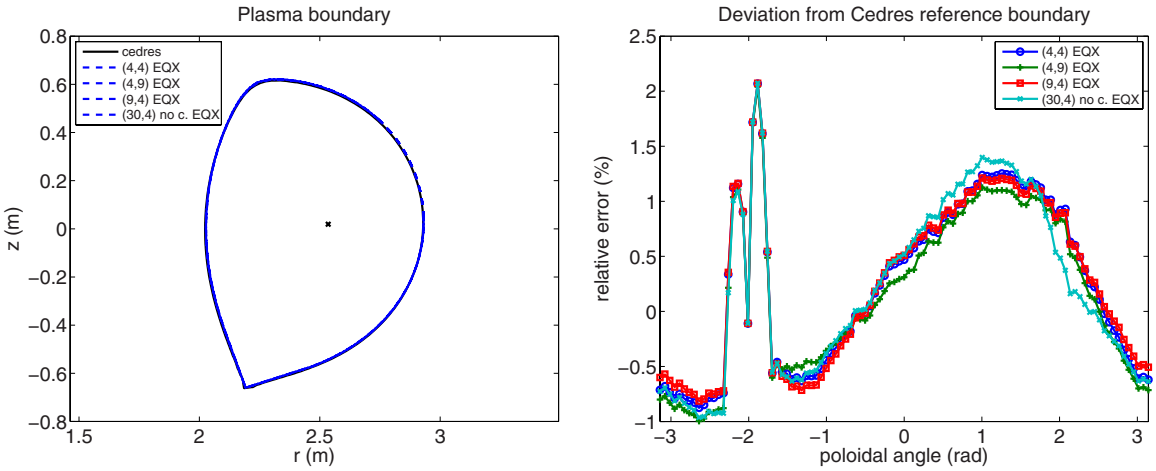

Figure 5. Left: plasma boundaries. Boundaries computed by Equinox (EQX) with $(n^e = 4, n^i = 4)$ or $(n^e = 9, n^i = 4)$ and $(n^e = 30, n^i = 4)$ without any PF coils modelization (no c) are almost superimposed with the reference boundary computed with Cedres++. The boundary computed with Equinox $(n^e = 4, n^i = 9)$ does not show any irregularities. Right: corresponding relative deviation from the Cedres++ boundary, $100(\rho - \rho_{\text{Cedres}})/\rho_{\text{Cedres}}$ as a function of the poloidal angle θ. The center of the polar coordinate system (ρ, θ) is the magnetic axis from Cedres++ (shown in the left-hand figure).

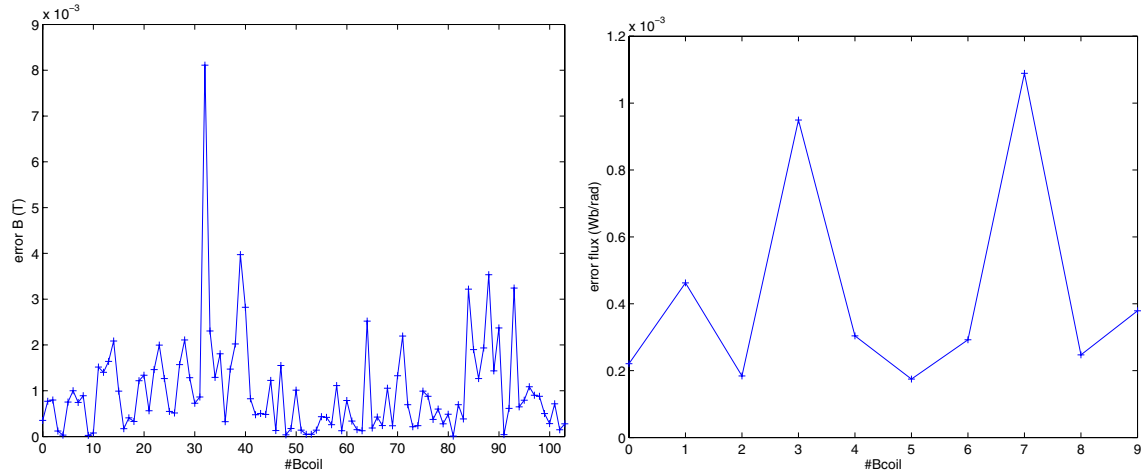

Figure 6. Comparison between measured and reconstructed values for WEST case 1 using $(n^e = 4, n^i = 4)$.

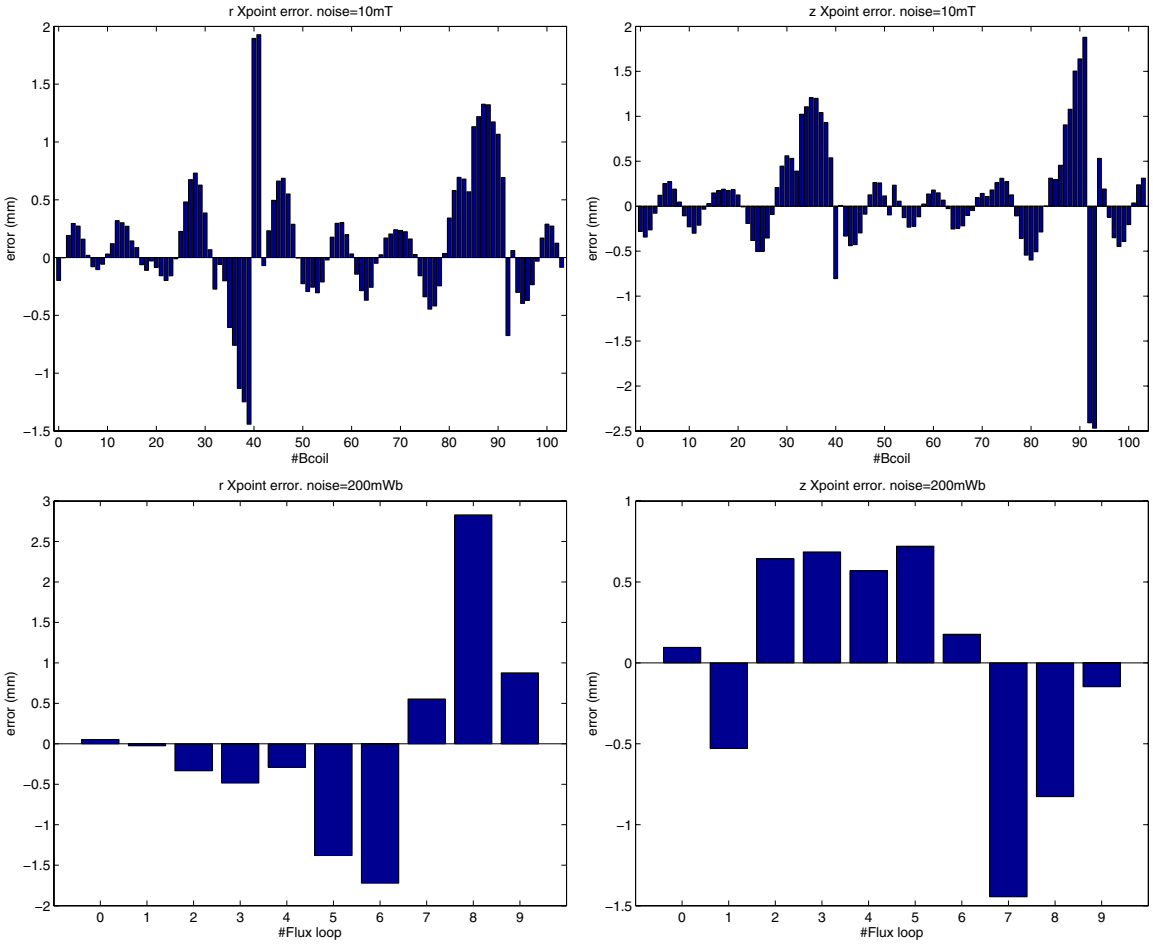

Figure 7. Error introduced on the X-point position reconstruction by adding a 10 mT offset on a single B-coil measurement or a 200 mWb offset on a single flux-loop measurement.

naturally the X-point position is more dependent on sensors that are in the divertor region than on others.

The numerical results for case 2 (the smaller plasma) are very similar to those presented above for case 1. It should be mentioned, however, that the plasma boundary reconstructed by the TH algorithm with the default ($n^e = 4, n^i = 4$) choice presents a small concavity on the high field side. Nevertheless it is small (a distance of maximum 2 cm from the Cedres boundary) and again disappears in the plasma boundary computed by Equinox (see figure 8). In cases 3 and 4 with a limiter configuration, the plasma boundary reconstructions are accurate (see figure 8).

Table 2. Rows 1 and 2: distance between the X-point given by Cedres and the one computed by the Toroidal Harmonics algorithm (TH) or the one re-computed by Equinox (EQX). Row 3: distance between the current center used as the pole of the toroidal coordinate system in TH and the magnetic axis given by Cedres. Row 4: distance between the magnetic axis given by Cedres and computed by Equinox.

(n^e, n^i)	(4, 4)	(4, 9)	(9, 4)	(30, 4) no c
$\|\|X\text{pt}_{\text{TH}} - X\text{pt}_{\text{Cedres}}\|\|$ (mm)	6.7	8.6	7.4	7.8
$\|\|X\text{pt}_{\text{EQX}} - X\text{pt}_{\text{Cedres}}\|\|$ (mm)	5.4	5.4	5.4	5.4
$\|\|C_{\text{TH}} - \text{Mag}_{\text{Cedres}}\|\|$ (mm)	18.5	18.3	18.5	19.3
$\|\|\text{Mag}_{\text{EQX}} - \text{Mag}_{\text{Cedres}}\|\|$ (mm)	4.4	3.6	4.5	4.0

5.2. Computing time

In view of the possible use of this method for the real-time control of the plasma position and shape in the WEST tokamak, it is important to evaluate the computing time for one boundary reconstruction. Each evaluation of the flux or the field at a given point has a cost in terms of computing time because the evaluation of the toroidal harmonics as well as the elliptic integrals involved in the expression of the flux generated by a filament of current have one. Therefore, in order to have an efficient code, all these functions are precomputed and stored in tables. The evaluation of a function then just involves a linear interpolation between two entries of a table. A second point concerns parallelism. Many loops in the code (matrix assembly, integrals computations, boundary points computation, ...) can be parallelized. We have used OpenMP to do so. The program is tested on a laptop with two quadcore processors running at 2.4 GHz. With this material configuration, the code takes about 2 ms for one boundary reconstruction as shown in table 3. Although this is already in the range of computing time needed for the real-time control of the plasma on WEST, this result could still be improved using more threads or even GPU as proposed in [30].

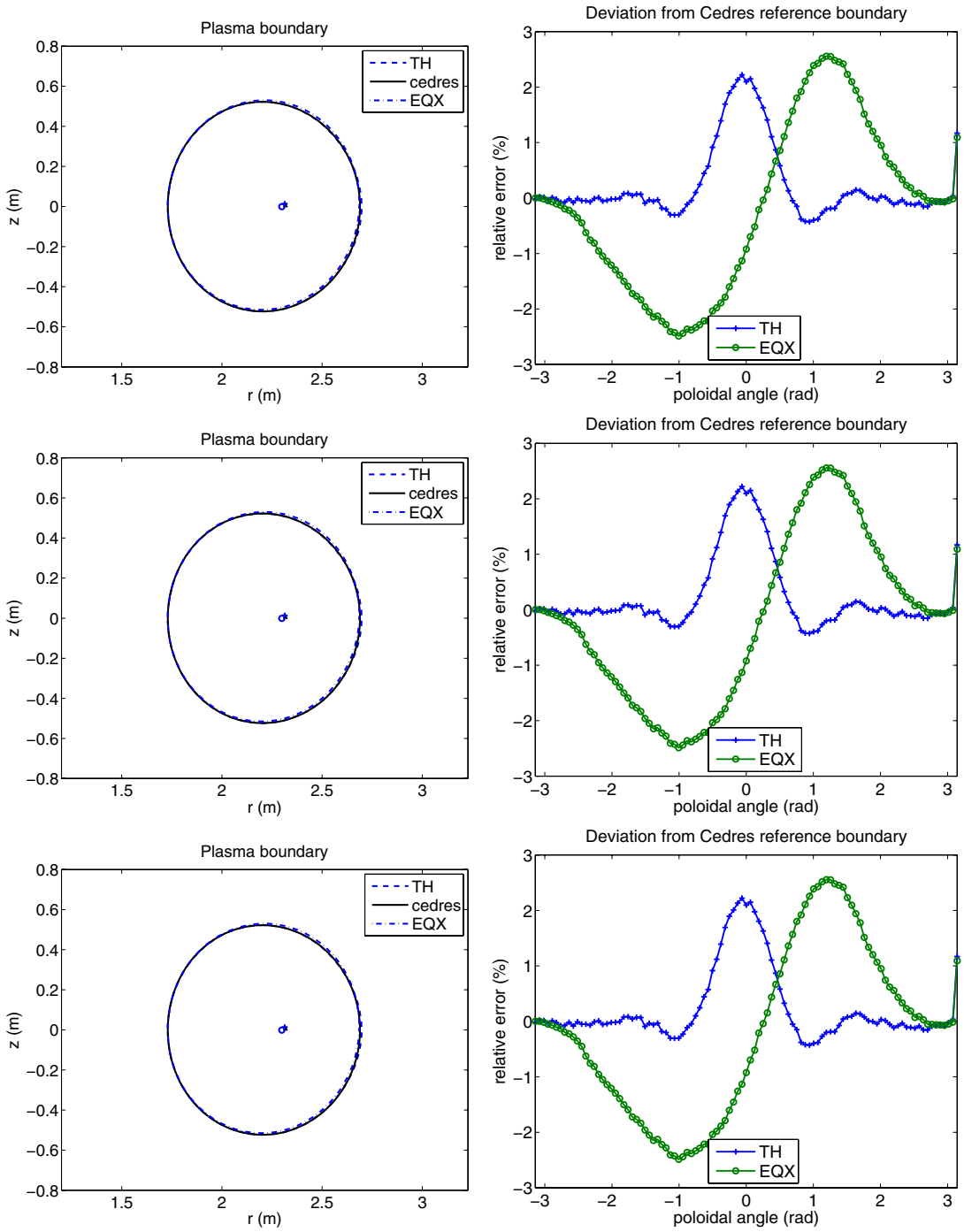

Figure 8. Row 1: case 2. Left: Cedres++ reference and boundaries reconstructed with toroidal harmonics (TH) and Equinox (EQX). The Cedres++ and Equinox magnetic axis as well as the computed plasma center taken as the pole of the toroidal coordinate system (circle) are also shown. Right: corresponding relative deviation from the Cedres++ boundary, $100(\rho - \rho_{\text{Cedres}})/\rho_{\text{Cedres}}$ as a function of the poloidal angle θ. The center of the polar coordinate system (ρ, θ) is the magnetic axis from Cedres++ (shown in the left-hand figure). Row 2: the same for case 3. Row 3: the same for case 4.

5.3. Equinox on the ITM platform

The method presented in this paper has also been tested on data provided in the ITM database. The aim is to be able to run the equilibrium reconstruction code Equinox directly from the discrete magnetic measurements.

During a first initialization phase all the geometric inputs are read from the database: the limiter contour, the PF-coils geometry and location, the B-probes orientation and location, and the flux and saddle loops location. The convex hull of the limiter contour is computed. This is the Γ contour which is the boundary of the computation domain for the finite element part of Equinox. From this contour, a mesh is generated. Of course many other contours could be used as the boundary, and if desired one can define one's own, point by point. However the convex hull of the limiter has the advantage of being computed automatically.

Table 3. Wall-clock computing time. One boundary point corresponds to the following operations: update the normal equation matrix as the pole of the toroidal coordinate system changes, solve the normal equation, compute the new current center (to be given to the next time step), and compute the point defining the boundary isoflux value ψ_b (either limiter point or X-point). Then add to this the computation of 9, 19, 29 or 59 boundary points for the next columns.

Nbr of bnd pts	1	10	20	30	60
Comp. time (ms)	1.09	1.23	1.44	1.54	1.98

Then comes the time stepping. Each time step is made of two stages. In the first one at the discrete time t^n, the pole of the toroidal coordinate system is set to the magnetic axis location computed at time t^{n-1}. The contribution of the different PF coils to the flux is computed and subtracted from the magnetic measurements. The residuals are then fitted to a truncated series of toroidal harmonics.

Once this is done, the flux can be evaluated at any point of an unknown annular domain surrounding the plasma and therefore clearly on the contour Γ. We are thus able to compute Cauchy boundary conditions on Γ. Note that even if it is possible in principle, we do not compute the plasma boundary at this stage. Indeed, we want to run the finite element method of Equinox on a fixed domain which does not need to be re-meshed at each time step. The plasma boundary is thus computed during this second stage, along with all the parameters which characterize a plasma equilibrium (including the magnetic axis, which will be used at the next time step), which are then copied to the ITM database.

6. Conclusion

We have presented in this paper a method based on the use of toroidal harmonics and on a modelization of the poloidal field coils and divertor coils for the 2D interpolation of discrete magnetic measurements.

The method completely relies on the classical assumptions that the equilibrium is axisymmetric and that a negligible amount of the total current density flows in the plasma existing in the region of the sensors (i.e. $\Delta^*\psi = 0$ holds in this region). If the first assumption was to be defaulted with non-negligible 3D effects [31–35], the method might be destabilized. The same conclusion holds if the second assumption was to fail since the decomposition of the flux in a series of toroidal harmonics with constant coefficients is not exact anymore.

However under these assumptions our numerical results show that the method is quite stable even though it does not involve a classical regularization procedure. This is due to the fact that the ill-posed part of the method, that is to say the computation of the internal solution, only relies on the choice of the pole of the toroidal coordinate system and on the number of internal toroidal harmonics used to approximate the flux. Our numerical experiments show that the magnetic axis is a good and easy-to-compute choice for the first point, and, concerning the second point, that only a few toroidal harmonics are needed to accurately approximate the flux.

The method is generic and can be used to provide Cauchy boundary conditions needed as the input by a fixed domain equilibrium reconstruction code like Equinox. This is implemented in the ITM version of Equinox. The method can also be used to extrapolate the magnetic measurements to compute the X-point position and the plasma boundary. It is foreseen to be used in the real-time plasma control loop on the WEST tokamak.

Acknowledgments

We would like to thank the two anonymous reviewers for their constructive criticism from which the paper has benefitted significantly.

References

[1] Blum J, Boulbe C and Faugeras B 2012 Reconstruction of the equilibrium of the plasma in a tokamak and identification of the current density profile in real time *J. Comput. Phys.* **231** 960–80
[2] Bucalossi J *et al* 2011 Feasibility study of an actively cooled tungsten divertor in Tore Supra for ITER technology testing *Fusion Eng. Des.* **86** 684–8
[3] Wesson J 2004 *Tokamaks* (*International Series of Monographs on Physics* vol 118) 3rd edn (New York: Oxford University Press)
[4] O'Brien D P, Lao L L, Solano E R, Garribba M, Taylor T S, Cordey J G and Ellis J J 1992 Equilibrium analysis of iron core tokamaks using a full domain method *Nucl. Fusion* **32** 1351–60
[5] Blum J 1989 *Numerical Simulation and Optimal Control in Plasma Physics with Applications to Tokamaks* (*Series in Modern Applied Mathematics*) (New York/Paris: Wiley/Gauthier-Villars)
[6] Faugeras B, Ben Abda A, Blum J and Boulbe C 2012 Minimization of an energy error functional to solve a cauchy problem arising in plasma physics: the reconstruction of the magnetic flux in the vacuum surrounding the plasma in a tokamak *ARIMA J.* **15** 37–60
[7] O'Brien D P, Ellis J J and Lingertat J 1993 Local expansion method for fast plasma boundary identification in JET *Nucl. Fusion* **33** 467–74
[8] Sartori F, Cenedese A and Milani F 2003 JET real-time object-oriented code for plasma boundary reconstruction *Fusion Eng. Des.* **66–68** 735–9
[9] Saint-Laurent F and Martin G 2001 Real time determination and control of the plasma localisation and internal inductance in Tore Supra *Fusion Eng. Des.* **56–57** 761–5
[10] Integrated Tokamak Modelling 2013 http://portal.efda-itm.eu/
[11] Braams B J 1991 The interpretation of tokamak magnetic diagnostics *Plasma Phys. Control. Fusion* **33** 715–48
[12] Fock V A 1932 *Fiz. Zh. Sovietunion* **1** 215
[13] Morse P M and Feshbach H 1953 *Methods of Theoretical Physics* (Cambridge: Cambridge University Press)
[14] Lee D K and Peng Y-K M 1981 An approach to rapid plasma shape diagnostics in tokamaks *J. Plasma Phys.* **25** 161–73
[15] Deshko G N, Kilovataya T G, Kuznetsov Y K, Pyatov V N and Yasin I V 1983 Determination of the plasma column shape in a tokamak from magnetic measurements *Nucl. Fusion* **23** 1309–17
[16] Bondarenko S P, Golant V E, Gryaznevich M P, Kuznetsov Y K, Pyatov V N, Taran V S, Shakhovets K G and Yasin I V 1984 Measurement of the shape of the plasma column in the tuman-3 tokamak *Sov. J. Plasma Phys.* **10** 520–4

[17] Alladio F and Crisanti F 1986 Analysis of MHD equilibria by toroidal multipolar expansions *Nucl. Fusion* **26** 1143–63
[18] Van Milligen B P 1990 Exact relations between multipole moments of the flux and moments of the toroidal current density in tokamaks *Nucl. Fusion* **30** 157–60
[19] Kurihara K 1992 Improvement of tokamak plasma shape identification with a Legendre–Fourier expansion of the vacuum poloidal flux function *Fusion Technol.* **22** 334–49
[20] Van Milligen B P and Lopez Fraguas A 1994 Expansion of vacuum magnetic fields in toroidal harmonics *Comput. Phys. Commun.* **81** 74–90
[21] Fischer Y 2012 Identification de paramètres magnétiques à l'intérieur d'un tokamak *J. Eur. Syst. Automatisés* **46** 611–32
[22] Lebedev N N 1972 *Special Functions and their Applications* (New York: Dover)
[23] Braams B J 1986 Computational studies in Tokamak equilibrium and transport *PhD Thesis* University of Utrecht
[24] Fischer Y 2011 Approximation dans des classes de fonctions analytiques généralisées et résolution de problèmes inverses pour les tokamaks *PhD Thesis* Université de Nice-Sophia Antipolis
[25] Abramowitz M and Stegun I A 1964 *Handbook of Mathematical Functions* (Washington, DC: National Bureau of Standards)
[26] Durand E 1968 *Magnétostatique* (Paris: Masson et Cie)
[27] Segura J and Gil A 2000 Evaluation of toroidal harmonics *Comput. Phys. Commun.* **124** 104–22
[28] Zakharov L E and Shafranov V D 1973 Equilibrium of a toroidal plasma with non-circular cross section *Sov. Phys. Tech. Phys.* **18** 151–8
[29] Hertout P, Boulbe C, Nardon E, Blum J, Brémond S, Bucalossi J, Faugeras B, Grandgirard V and Moreau P 2011 The CEDRES++ equilibrium code and its application to ITER, JT-60SA and Tore Supra *Fusion Eng. Des.* **86** 1045–8
[30] Yue X N, Xiao B J, Luo Z P and Guo Y 2013 Fast equilibrium reconstruction for tokamak discharge control based on GPU *Plasma Phys. Control. Fusion* **55** 9
[31] Cooper W A, Graves J P and Sauter O 2011 Helical ITER hybrid scenario equilibria *Plasma Phys. Control. Fusion* **53** 024002
[32] Cooper W A *et al* 2013 Bifurcated helical core equilibrium states in tokamaks *Nucl. Fusion* **53** 073021
[33] Lazerson S and Chapman I T 2013 STELLOPT modeling of the 3D diagnostic response in ITER *Plasma Phys. Control. Fusion* **55** 084004
[34] Hanson J D *et al* 2013 Non-axisymmetric equilibrium reconstruction for stellarators, reversed field pinches and tokamaks *Nucl. Fusion* **53** 083016
[35] Turnbull A D *et al* 2013 Comparisons of linear and nonlinear plasma response models for non-axisymmetric perturbations *Phys. Plasmas* **20** 056114

Article D : [6] B. Faugeras, A. Ben Abda, J. Blum et C. Boulbe. Minimization of an energy error functional to solve a Cauchy problem arising in plasma physics : the reconstruction of the magnetic flux in the vacuum surrounding the plasma in a Tokamak. *ARIMA* 15 (2012), p. 37–60

Minimization of an energy error functional to solve a Cauchy problem arising in plasma physics: the reconstruction of the magnetic flux in the vacuum surrounding the plasma in a Tokamak

Blaise Faugeras* — Amel Ben Abda** — Jacques Blum* — Cedric Boulbe*

* Laboratoire J.A. Dieudonné, UMR 6621, Université de Nice Sophia Antipolis,
06108 Nice Cedex 02, France
Blaise.Faugeras@unice.fr

** ENIT-LAMSIN,
BP 37, 1002 Tunis, Tunisie
Amel.BenAbda@enit.rnu.tn

RÉSUMÉ. Une méthode numérique de calcul du flux magnétique dans le vide entourant la plasma dans un Tokamak est étudiée. Elle est basée sur la formulation d'un problème de Cauchy qui est résolu en minimisant une fonctionnelle d'écart énergétique. Différentes expériences numériques montrent l'efficacité de la méthode.

ABSTRACT. A numerical method for the computation of the magnetic flux in the vacuum surrounding the plasma in a Tokamak is investigated. It is based on the formulation of a Cauchy problem which is solved through the minimization of an energy error functional. Several numerical experiments are conducted which show the efficiency of the method.

MOTS-CLÉS : EDP elliptique, problème inverse, écart énergétique, plasma, Tokamak

KEYWORDS : Elliptic PDE, inverse problem, energy error, plasma, Tokamak

Received,June 19, 2011,
Revised, February 2, 2012,
Accepted,April 23, 2012

ARIMA Journal, vol. 15 (2012), pp. 37-60

1. Introduction

In order to be able to control the plasma during a fusion experiment in a Tokamak it is mandatory to know its position in the vacuum vessel. This latter is deduced from the knowledge of the poloidal flux which itself relies on measurements of the magnetic field. In this paper we investigate a numerical method for the computation of the poloidal flux in the vacuum. Let us first briefly recall the equations modelizing the equilibrium of a plasma in a Tokamak [32].

Assuming an axisymmetric configuration one considers a 2D poloidal cross section of the vacuum vessel Ω_V in the (r, z) system of coordinates (Fig. 1). In this setting the poloidal flux $\psi(r, z)$ is related to the magnetic field through the relation $(B_r, B_z) = \frac{1}{r}(-\frac{\partial \psi}{\partial z}, \frac{\partial \psi}{\partial r})$ and, as there is no toroidal current density in the vacuum outside the plasma, satisfies the following equation

$$L\psi = 0 \text{ in } \Omega_X \tag{1}$$

where L denotes the elliptic operator

$$L. = -[\frac{\partial}{\partial r}(\frac{1}{r}\frac{\partial .}{\partial r}) + \frac{\partial}{\partial z}(\frac{1}{r}\frac{\partial .}{\partial z})]$$

and

$$\Omega_X = \Omega_V - \bar{\Omega}_P$$

denotes the vacuum region surrounding the domain of the plasma Ω_P of boundary Γ_P (see Fig. 2). Inside the plasma Eq. (1) is not valid anymore and the poloidal flux satisfies the Grad-Shafranov equation [30, 16] which describes the equilibrium of a plasma confined by a magnetic field

$$L\psi = \mu_0 j(r, \psi) \text{ in } \Omega_P \tag{2}$$

where μ_0 is the magnetic permeability of the vacuum and $j(r, \psi)$ is the unknown toroidal current density function inside the plasma. Since the plasma boundary Γ_P is unknown the equilibrium of a plasma in a Tokamak is a free boundary problem described by a particular non-linearity of the model. The boundary is an iso-flux line determined either as being a magnetic separatrix (hyperbolic line with an X-point as on the left hand side of Fig. 2) or by the contact with a limiter (Fig. 2 right hand side). In other words the plasma boundary is determined from the equation $\psi(r, z) = \psi_P$, ψ_P being the value of the flux at the X-point or the value of the flux for the outermost flux line inside a limiter.

In order to compute an approximation of ψ in the vacuum and to find the plasma boundary without knowing the current density j in the plasma and thus without using the Grad-Shafranov equation (2) the strategy which is routinely used in operational codes

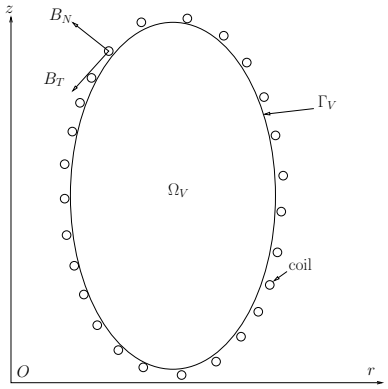

Figure 1. *Cross section of the vacuum vessel : the domain Ω_V, its boundary Γ_V. Coils providing measurements of the components of the magnetic field tangent and normal to Γ_V are represented surrounding the vacuum vessel.*

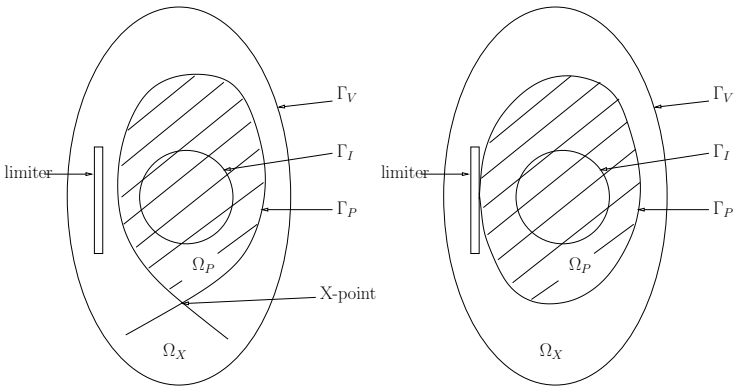

Figure 2. *The plasma domain Ω_P and the vacuum region Ω_X. The plasma boundary is determined by an X-point configuration (left) or a limiter configuration (right). The fictitious contour Γ_I is represented inside the plasma.*

mainly consists in choosing an a priori expansion method for ψ such as for example truncated Taylor and Fourier expansions for the code Apolo on the Tokamak ToreSupra [28] or piecewise polynomial expansions for the code Xloc on the Tokamak JET [26, 29]. The flux ψ can also be expanded in toroidal harmonics involving Legendre functions or expressed by using Green functions in the filament method ([23, 13], [9] and the references therein). In all cases the coefficients of the expansion are then computed through a fit to the measurements of the magnetic field. Indeed several magnetic probes and flux loops surround the boundary Γ_V of the vacuum vessel and measure the magnetic field and the flux (see Fig. 1). It should also be noted that very similar problems are studied in [18, 8, 14, 15]

In this paper we investigate a numerical method based on the resolution of a Cauchy problem introduced in ([6], Chapter 5) which we recall here below. The proposed approach uses the fact that after a preprocessing of these measurements (interpolation and possibly integration on a contour) one can have access to a complete set of Cauchy data, $f = \psi$ on Γ_V and $g = \dfrac{1}{r}\dfrac{\partial \psi}{\partial n}$ on Γ_V.

The poloidal flux satisfies

$$\begin{cases} L\psi = 0 \quad \text{in } \Omega_X \\ \psi = f \quad \text{on } \Gamma_V \\ \dfrac{1}{r}\dfrac{\partial \psi}{\partial n} = g \quad \text{on } \Gamma_V \\ \psi = \psi_P \quad \text{on } \Gamma_P \end{cases} \tag{3}$$

In this formulation the domain $\Omega_X = \Omega_X(\psi)$ is unknown since the free plasma boundary Γ_P as well as the flux ψ_P on the boundary are unknown. Moreover the problem is ill-posed in the sense of Hadamard [12] since there are two Cauchy conditions on the boundary Γ_V.

In order to compute the flux in the vacuum and to find the plasma boundary we are going to define a new problem as in [6] which is an approximation of the original one. Let us define a fictitious boundary Γ_I fixed inside the plasma (see Fig. 2). We are going to seek an approximation of the poloidal flux ψ satisfying $L\psi = 0$ in the domain contained between the fixed boundaries Γ_V and Γ_I. The problem becomes one formulated on a fixed domain Ω :

$$\begin{cases} L\psi = 0 \quad \text{in } \Omega \\ \psi = f \quad \text{on } \Gamma_V \\ \dfrac{1}{r}\dfrac{\partial \psi}{\partial n} = g \quad \text{on } \Gamma_V \end{cases} \tag{4}$$

Let us insist here on the fact that this problem is an approximation to the original one since in the domain between Γ_P and Γ_I, ψ should satisfy the Grad-Shafranov equation. The relevance of this approximating model is consolidated by the Cauchy-Kowalewska theorem [12]. For Γ_P smooth enough the function ψ can be extended in the sense of $L\psi = 0$ in a neighborhood of Γ_P inside the plasma. Hence the problem formulated on a fixed domain with a fictitious boundary Γ_I not "too far" from Γ_P is an approximation of the free boundary problem. As mentioned in [6] if Γ_I were identical with Γ_P then by the virtual shell principle [31] the quantity $w = \frac{1}{r}\frac{\partial \psi}{\partial n}|_{\Gamma_I}$ would represent the surface current density (up to a factor $\frac{1}{\mu_0}$) on Γ_P for which the magnetic field created outside the plasma by the current sheet is identical to the field created by the real current density spread throughout the plasma.

However no boundary condition is known on Γ_I. One way to deal with this second issue and to solve such a problem is to formulate it as an optimal control one. Only the Dirichlet condition on Γ_V is retained to solve the boundary value problem and a least square error functional measuring the distance between measured and computed normal derivative and depending on the unknown boundary condition on Γ_I is minimized. Due to the illposedness of the considered Cauchy problem a regularization term is needed to avoid erratic behaviour on the boundary where the data is missing. A drawback of this method developed in [6] is that Dirichlet and Neumann boundary conditions on Γ_V are not used in a symmetric way. One is used as a boundary condition for the partial differential equation, $L\psi = 0$, whereas the other is used in the functional to be minimized.

Freezing the domain to Ω by introducing the fictitious boundary Γ_I enables to remove the nonlinearity of the problem. The plasma boundary Γ_P can still be computed as an isoflux line and thus is an output of our computations. We are going to compute a function ψ such that the Dirichlet boundary condition $u = \psi$ on Γ_I is such that the Cauchy conditions on Γ_V are satisfied as nearly as possible in the sense of the error functional defined in the next Section.

The originality of the approach proposed in this paper relies on the use of an error functional having a physical meaning : an energy error functional or constitutive law error functional. Up to our knowledge this misfit functional has been introduced in [24] in the context of a posteriori estimator in the finite element method. In this context, the minimization of the constitutive law error functional allows to detect the reliability of the mesh without knowing the exact solution. Within the inverse problem community this functional has been introduced in [21, 22, 20] in the context of parameter identification. It has been widely exploited in the same context in [7]. It has also been used for Robin type boundary condition recovering [10] and in the context of geometrical flaws identification (see [4] and references therein). For lacking boundary data recovering (i.e. Cauchy problem resolution) in the context of Laplace operator, the energy error functional has been introduced in [2, 1]. A study of similar techniques can be found in [5, 3] and the analysis

found in these papers uses elements taken from the domain decomposition framework [27].

The paper is organized as follows. In Section 2 we give the formulation of the problem we are interested in and provide an analysis of its well posedness. Section 3 describes the numerical method used. Several numerical experiments are conducted to validate it. The final experiment shows the reconstruction of the poloidal flux and the localization of the plasma boundary for an ITER configuration.

2. Formulation and analysis of the method

2.1. Problem formulation

As described in the Introduction the starting point is the free boundary problem (3). We first proceed as in [6] and in a first step consider the fictitious contour Γ_I fixed in the plasma and the fixed domain Ω contained between Γ_V and Γ_I. Problem (3) is approximated by the Cauchy problem (4). The boundaries Γ_V and Γ_I are assumed to be chosen smooth enough in order not to refrain any of the developments which follow in the paper.

In a second step the problem is separated into two different ones. In the first one we retain the Dirichlet boundary condition on Γ_V only, assume v is given on Γ_I and seek the solution ψ_D of the well-posed boundary value problem :

$$\begin{cases} L\psi_D = 0 \quad \text{in } \Omega \\ \psi_D = f \quad \text{on } \Gamma_V \\ \psi_D = v \quad \text{on } \Gamma_I \end{cases} \tag{5}$$

The solution ψ_D can be decomposed in a part linearly depending on v and a part depending on f only. We have the following decomposition :

$$\psi_D = \psi_D(v, f) = \psi_D(v, 0) + \psi_D(0, f) := \psi_D(v) + \tilde{\psi}_D(f) \tag{6}$$

where $\psi_D(v)$ and $\tilde{\psi}_D(f)$ satisfy :

$$\begin{cases} L\psi_D(v) = 0 \quad \text{in } \Omega \\ \psi_D(v) = 0 \quad \text{on } \Gamma_V \\ \psi_D(v) = v \quad \text{on } \Gamma_I \end{cases} \qquad \begin{cases} L\tilde{\psi}_D(f) = 0 \quad \text{in } \Omega \\ \tilde{\psi}_D(f) = f \quad \text{on } \Gamma_V \\ \tilde{\psi}_D(f) = 0 \quad \text{on } \Gamma_I \end{cases} \tag{7}$$

In the second problem we retain the Neumann boundary condition only and look for ψ_N satisfying the well-posed boundary value problem :

$$\begin{cases} L\psi_N = 0 \quad \text{in } \Omega \\ \dfrac{1}{r}\dfrac{\partial \psi_N}{\partial n} = g \quad \text{on } \Gamma_V \\ \psi_N = v \quad \text{on } \Gamma_I \end{cases} \tag{8}$$

in which ψ_N can be decomposed in a part linearly depending on v and a part depending on g only. We have the following decomposition :

$$\psi_N = \psi_N(v,g) = \psi_N(v,0) + \psi_N(0,g) := \psi_N(v) + \tilde{\psi}_N(g) \tag{9}$$

where

$$\begin{cases} L\psi_N(v) = 0 \quad \text{in } \Omega \\ \dfrac{1}{r}\dfrac{\partial \psi_N(v)}{\partial n} = 0 \quad \text{on } \Gamma_V \\ \psi_N(v) = v \quad \text{on } \Gamma_I \end{cases} \qquad \begin{cases} L\tilde{\psi}_N(g) = 0 \quad \text{in } \Omega \\ \dfrac{1}{r}\dfrac{\partial \tilde{\psi}_N}{\partial n} = g \quad \text{on } \Gamma_V \\ \psi_N = 0 \quad \text{on } \Gamma_I \end{cases} \tag{10}$$

In order to solve problem (4), $f \in H^{1/2}(\Gamma_V)$ and $g \in H^{-1/2}(\Gamma_V)$ being given, we would like to find $u \in \mathcal{U} = H^{1/2}(\Gamma_I)$ such that $\psi = \psi_D(u,f) = \psi_N(u,g)$. To achieve this we are in fact going to seek u such that $J(u) = \inf_{v \in \mathcal{U}} J(v)$ where J is the error functional defined by

$$J(u) = \frac{1}{2}\int_\Omega \frac{1}{r}||\nabla\psi_D(u,f) - \nabla\psi_N(u,g)||^2 dx \tag{11}$$

measuring a misfit between the Dirichlet solution and the Neumann solution.

2.2. Analysis of the method

In order to minimize J one can compute its derivative and express the first order optimality condition. When doing so the two symmetric bilinear forms s_D and s_N as well as the linear form l defined below appear naturally and in a first step it is convenient to give a new expression of functional (11) using these forms.

Let $u, v \in H^{1/2}(\Gamma_I)$ and define

$$s_D(u,v) = \int_\Omega \frac{1}{r}\nabla\psi_D(u)\nabla\psi_D(v)dx \tag{12}$$

Applying Green's formula and noticing that $\psi_D(v) = v$ on Γ_I and $\psi_D(v) = 0$ on Γ_V we obtain

$$s_D(u,v) = \int_{\partial\Omega} \frac{1}{r}\partial_n\psi_D(u)\psi_D(v)d\sigma - \int_\Omega \nabla(\frac{1}{r}\nabla\psi_D(u))\psi_D(v)dx = \int_{\Gamma_I} \frac{1}{r}\partial_n\psi_D(u)vd\sigma \tag{13}$$

where the integrals on the boundary are to be understood as duality pairings. In Eq. (13) one can replace $\psi_D(v)$ by any extension $\mathcal{R}(v)$ in $H^1_0(\Omega,\Gamma_V) = \{\psi \in H^1(\Omega), \psi_{|\Gamma_V} = 0\}$ of $v \in H^{1/2}(\Gamma_I)$.
Hence s_D can be represented by

$$s_D(u,v) = \int_\Omega \frac{1}{r}\nabla\psi_D(u)\nabla\mathcal{R}(v)dx \tag{14}$$

Equivalently s_N is defined by

$$s_N(u,v) = \int_\Omega \frac{1}{r}\nabla\psi_N(u)\nabla\psi_N(v)dx \tag{15}$$

Since $\psi_N(v) = v$ on Γ_I and $\frac{1}{r}\partial_n\psi_N(u) = 0$ on Γ_V we have that

$$s_N(u,v) = \int_{\partial\Omega} \frac{1}{r}\partial_n\psi_N(u)\psi_N(v)d\sigma - \int_\Omega \nabla(\frac{1}{r}\nabla\psi_N(u))\psi_N(v)dx = \int_{\Gamma_I} \frac{1}{r}\partial_n\psi_N(u)vd\sigma \tag{16}$$

and s_N can also be represented by

$$s_N(u,v) = \int_\Omega \frac{1}{r}\nabla\psi_N(u)\nabla\mathcal{R}(v)dx \tag{17}$$

where $\mathcal{R}(v)$ is any extension in $H^1(\Omega)$ of $v \in H^{1/2}(\Gamma_I)$.

Let us now introduce

$$F(u,v) = \frac{1}{2}\int_\Omega \frac{1}{r}(\nabla\psi_D(u,f) - \nabla\psi_N(u,g))(\nabla\psi_D(v,f) - \nabla\psi_N(v,g))dx \tag{18}$$

such that $J(v) = F(v,v)$ and the linear form l defined by

$$l(v) = -\int_\Omega \frac{1}{r}(\nabla\tilde{\psi}_D(f) - \nabla\tilde{\psi}_N(g))\nabla\psi_D(v)dx \tag{19}$$

which can also be computed as

$$l(v) = -\int_\Omega \frac{1}{r}(\nabla\tilde{\psi}_D(f) - \nabla\tilde{\psi}_N(g))\nabla\mathcal{R}(v)dx \tag{20}$$

It can then be shown that

$$F(u,v) = \frac{1}{2}(s_D(u,v) - s_N(u,v) - l(u) - l(v)) + c \tag{21}$$

where the constant c is given by

$$c = \frac{1}{2}\int_\Omega \frac{1}{r}||\nabla\tilde{\psi}_D(f) - \nabla\tilde{\psi}_N(g)||^2 dx \tag{22}$$

Hence functional J can be rewritten as

$$J(v) = \frac{1}{2}(s_D(v,v) - s_N(v,v)) - l(v) + c \tag{23}$$

Following the analysis provided in [5] it can be proved that in the favorable case of compatible Cauchy data (f,g) the Cauchy problem admits a solution. There exists a unique $u \in \mathcal{U}$ such that $\psi_D(u,f) = \psi_N(u,g)$. The minimum of J is also uniquely reached at this point, $J(u) = 0$. This solution is given by the first order optimality condition which reads

$$(J'(u), v) = s_D(u,v) - s_N(u,v) - l(v) = 0 \quad \forall v \in \mathcal{U} \tag{24}$$

Equation (24) has an interpretation in terms of the normal derivative of ψ_D and ψ_N on the boundary. From Eqs. (13) and (16) and from

$$l(v) = -\int_\Omega \frac{1}{r}(\nabla\tilde{\psi}_D(f) - \nabla\tilde{\psi}_N(g))\nabla\psi_D(v)dx = -\int_{\Gamma_I} \frac{1}{r}(\partial_n\tilde{\psi}_D(f) - \partial_n\tilde{\psi}_N(g))v d\sigma \tag{25}$$

we deduce that the optimality condition can be rewritten as

$$\int_{\Gamma_I} [(\frac{1}{r}\partial_n\psi_D(u,f) - \frac{1}{r}\partial_n\psi_N(u,g))]v d\sigma = 0 \quad \forall v \in \mathcal{U} \tag{26}$$

which can be understood as the equality of the normal derivatives on Γ_I.

Hence the first optimality condition when minimizing J amounts to solve an interfacial equation

$$(S_D - S_N)(v) = \chi,$$

where S_D and S_N are the Dirichlet-to-Neumann operators associated to the bilinear forms and defined by :

$$\begin{array}{rccc} S_D & : & H^{1/2}(\Gamma_I) & \longrightarrow & H^{-1/2}(\Gamma_I) \\ & & v & \longrightarrow & \frac{1}{r}\frac{\partial\psi_D(v)}{\partial n}. \end{array} \tag{27}$$

$$\begin{array}{rcccc} S_N & : & H^{1/2}(\Gamma_I) & \longrightarrow & H^{-1/2}(\Gamma_I) \\ & & v & \longrightarrow & \dfrac{1}{r}\dfrac{\partial \psi_N(v)}{\partial n}, \end{array} \tag{28}$$

and $\chi = -\dfrac{1}{r}\dfrac{\partial \tilde{\psi}_D}{\partial n} + \dfrac{1}{r}\dfrac{\partial \tilde{\psi}_N}{\partial n}$ on Γ_I.

Since S_D and S_N have the same eigenvectors and have asymptotically the same eigenvalues, the interfacial operator $S = S_D - S_N$ is almost singular [5]. This point together with the fact that the set of incompatible Cauchy data is known to be dense in the set of compatible data (and thus numerical Cauchy data can hardly by compatible) make this inverse problem severely ill-posed.

Some regularization process has to be used. One way to regularize the problem is to directly deal with the resolution of the underlying quasi-singular linear system using for example a relaxed gradient method [2, 1]. In this paper we have chosen a regularization method of the Tikhonov type. It consists in shifting the spectrum of S by adding a term

$$(S_D - S_N) + \varepsilon S_D.$$

where ε is a small regularization parameter. This regularization method is quite natural since the ill-posedness of the inverse problem and the lack of stability in the identification of u by the minimization of J is strongly linked to the fact that J is not coercive (see [5] and below). We are thus going to minimize the regularized cost function :

$$J_\varepsilon(v) = J(v) + \varepsilon R_D(v)$$

with

$$R_D(v) = \frac{1}{2}\int_\Omega \frac{1}{r}||\nabla \psi_D(v)||^2 dx$$

This brings us to the framework described in [25]. We want to solve the following

$$\text{Problem } P_\varepsilon : \quad \text{find } u_\varepsilon \in \mathcal{U} \text{ such that } J_\varepsilon(u_\varepsilon) = \inf_{v \in \mathcal{U}} J_\varepsilon(v)$$

and the following result holds.

Proposition 1 *1) Problem P_ε admits a unique solution $u_\varepsilon \in \mathcal{U}$ characterized by the first order optimality condition*

$$(J'_\varepsilon(u_\varepsilon), v) = \varepsilon s_D(u_\varepsilon, v) + s_D(u_\varepsilon, v) - s_N(u_\varepsilon, v) - l(v) = 0 \quad \forall v \in \mathcal{U} \tag{29}$$

2) For a fixed ε the solution is stable with respect to the data f and g.
If f^1, $f^2 \in H^{1/2}(\Gamma_V)$ and g^1, $g^2 \in H^{-1/2}(\Gamma_V)$ it holds that

$$||u^1_\varepsilon - u^2_\varepsilon||_{H^{1/2}(\Gamma_I)} \leq \frac{C}{\varepsilon}(||f^1 - f^2||_{H^{1/2}(\Gamma_V)} + ||g^1 - g^2||_{H^{-1/2}(\Gamma_V)}) \tag{30}$$

3) If there exists $u \in \mathcal{U}$ such that $\psi_D(u, f) = \psi_N(u, g)$ then $u_\varepsilon \to u$ in $\mathcal{U}$ when $\varepsilon \to 0$.

Elements of the proof are given in Appendix.

3. Numerical method and experiments

3.1. Finite element discretization

The resolution of the boundary value problems (7) and (10) is based on a classical P^1 finite element method [11].

Let us consider the family of triangulation τ_h of Ω, and V_h the finite dimensional subspace of $H^1(\Omega)$ defined by

$$V_h = \{\psi_h \in H^1(\Omega), \psi_{h|T} \in P^1(T), \forall T \in \tau_h\}.$$

Let us also introduce the finite element space on Γ_I

$$D_h = \{v_h = \psi_h|_{\Gamma_I},\ \psi_h \in V_h\}.$$

Consider $(\phi_i)_{i=1,\dots N}$ a basis of V_h and assume that the first N_{Γ_I} mesh nodes (and basis functions) correspond to the ones situated on Γ_I. A function $\psi_h \in V_h$ is decomposed as $\psi_h = \sum_{i=1}^{N} a_i\phi_i$ and its trace on Γ_I as $v_h = \psi_h|_{\Gamma_I} = \sum_{i=1}^{N_{\Gamma_I}} a_i\phi_i|_{\Gamma_I}$.

Given boundary conditions v_h on Γ_I and f_h, g_h on Γ_V one can compute the approximations $\psi_{D,h}(v_h)$, $\psi_{N,h}(v_h)$, $\tilde{\psi}_{D,h}(f_h)$ and $\tilde{\psi}_{N,h}(g_h)$ with the finite element method.

In order to minimize the discrete regularized error functional, $J_{\varepsilon,h}(u_h)$ we have to solve the discrete optimality condition which reads

$$\varepsilon s_{D,h}(u_h, v_h) + s_{D,h}(u_h, v_h) - s_{N,h}(u_h, v_h) - l(v_h) = 0 \quad \forall v_h \in D_h \tag{31}$$

which is equivalent to look for the vector $\mathbf{u}$ solution to the linear system

$$\mathbf{Su} = \mathbf{l} \tag{32}$$

where the $N_{\Gamma_I} \times N_{\Gamma_I}$ matrix $\mathbf{S}$ representing the bilinear form $s_h = \varepsilon s_{D,h} + s_{D,h} - s_{N,h}$ is defined by

$$\mathbf{S}_{ij} = s_h(\phi_i, \phi_j) \tag{33}$$

and $\mathbf{l}$ is the vector $(l_h(\phi_i))_{i=1,\dots N_{\Gamma_I}}$.

In order to lighten the computations the matrices are evaluated by

$$s_{D,h}(\phi_i, \phi_j) = \int_\Omega \frac{1}{r}\nabla\psi_{D,h}(\phi_i)\nabla\mathcal{R}(\phi_j)dx \tag{34}$$

and

$$s_{N,h}(\phi_i, \phi_j) = \int_\Omega \frac{1}{r} \nabla \psi_{N,h}(\phi_i) \nabla \mathcal{R}(\phi_j) dx \tag{35}$$

where $\mathcal{R}(\phi_j)$ is the trivial extension which coincides with ϕ_j on Γ_I and vanishes elsewhere.

In the same way the right hand side $\mathbf{l}$ is evaluated by

$$l_h(\phi_i) = -\int_\Omega \frac{1}{r} (\nabla \tilde{\psi}_{D,h}(f_h) - \nabla \tilde{\psi}_{N,h}(g_h)) \nabla \mathcal{R}(\phi_i) dx \tag{36}$$

It should be noticed here that matrix $\mathbf{S}$ depends on the geometry of the problem only and not on the input Cauchy data. Therefore it can be computed once for all (as well as its LU decomposition for exemple if this is the method used to invert the system) and be used for the resolution of successive problems with varying input data as it is the case during a plasma shot in a Tokamak. Only the right hand side $\mathbf{l}$ has to be recomputed. This enables very fast computation times.

All the numerical results presented in the remaining part of this paper were obtained using the software FreeFem++ (http ://www.freefem.org/ff++/). We are concerned with the geometry of ITER and the mesh used for the computations is shown on Fig. 3. It is composed of 1804 triangles and 977 nodes 150 of which are boundary nodes divided into 120 nodes on Γ_V and 30$= N_{\Gamma_I}$ on Γ_I. The shape of Γ_I is chosen empirically.

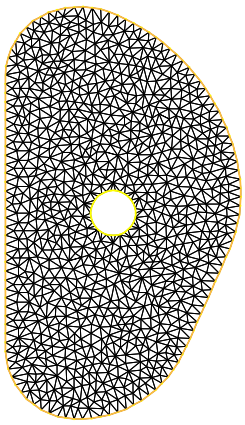

Figure 3. *The mesh used for the ITER configuration in FreeFem++*

3.2. Twin experiments

Numerical experiments with simulated input Cauchy data are conducted in order to validate the algorithm. Assume we are provided with a Neumann boundary condition function g on Γ_V. We generate the associated Dirichlet function f on Γ_V assuming a reference Dirichlet function u_{ref} is known on Γ_I. We thus solve the following boundary value problem :

$$\begin{cases} L\psi_{N,ref}(u_{ref}, g) = 0 \quad \text{in } \Omega \\ \\ \dfrac{1}{r}\partial_n \psi_{N,ref}(u_{ref}, g) = g \quad \text{on } \Gamma_V \\ \\ \psi_{N,ref}(u_{ref}, g) = u_{ref} \quad \text{on } \Gamma_I \end{cases} \tag{37}$$

and set $f = \psi_{N,ref}(u_{ref}, g)|_{\Gamma_V}$.

We have considered two test cases. In the first one (TC1)

$$u_{ref}(r, z) = 50 \sin(r)^2 + 50 \quad \text{on } \Gamma_I \tag{38}$$

and in the second one (TC2) u_{ref} is simply a constant

$$u_{ref}(r, z) = 40 \quad \text{on } \Gamma_I \tag{39}$$

The numerical experiments consist in minimizing the regularized error functional J_ε defined thanks to f and g. The obtained optimal solution u_{opt} and the associated ψ_{opt} are then compared to u_{ref} and ψ_{ref} which should ideally be recovered. Three cases are considered : the noise free case, a 1% noise on f and g and a 5% noise.

When the noise on f and g is small and the recovery of u is excellent there is very little difference between the Dirichlet solution $\psi_D(u_{opt}, f)$ and the Neumann solution $\psi_N(u_{opt}, g)$. However this is not the case any longer when the level of noise increases. The Dirichlet solution is much more sensitive to noise on f than the Neumann solution is sensitive to noise on g. Therefore the optimal solution is chosen to be $\psi_{opt} = \psi_N(u_{opt}, g)$.

The results are shown on Figs. 4 and 5 where the reference and recovered solutions are shown for the three levels of noise considered. The results are excellent for the noise free case in which the Dirichlet boundary condition u is almost perfecty recovered (Fig. 6). The differences between u_{opt} and u_{ref} increase with the level of noise (Fig. 6 and Tab. 1). As it is often the case in this type of inverse problems the most important errors on ψ_{opt} are localized close to the boundary Γ_I and vanishes as we move away from it (Fig. 7).

Tables 2 and 3 sumarize the evolution of the values of J, R_D and J_ε for the different noise level. First guess values ($u = 0$) are also provided for comparison. Please note

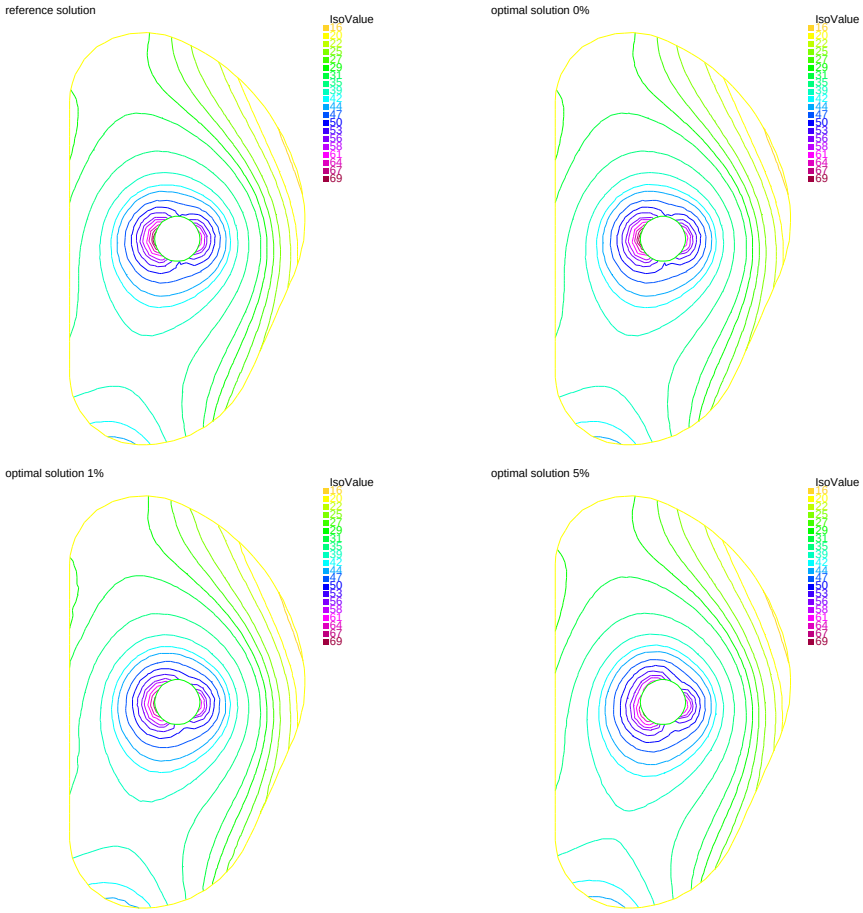

Figure 4. *First test case (TC1), u_{ref} given by Eq. (38). Top left : reference solution $\psi_{N,ref}(u_{ref}, g)$. Top right : recovered solution with no noise on the data. Bottom left : recovered solution with a 1% noise on the data. Bottom right : recovered solution with a 5% noise.*

that the regularization parameter was chosen differently from one experiment to another depending on the noise level. This was tuned by hand. In the next section we propose to use the L-curve method [19] to choose the value of ε.

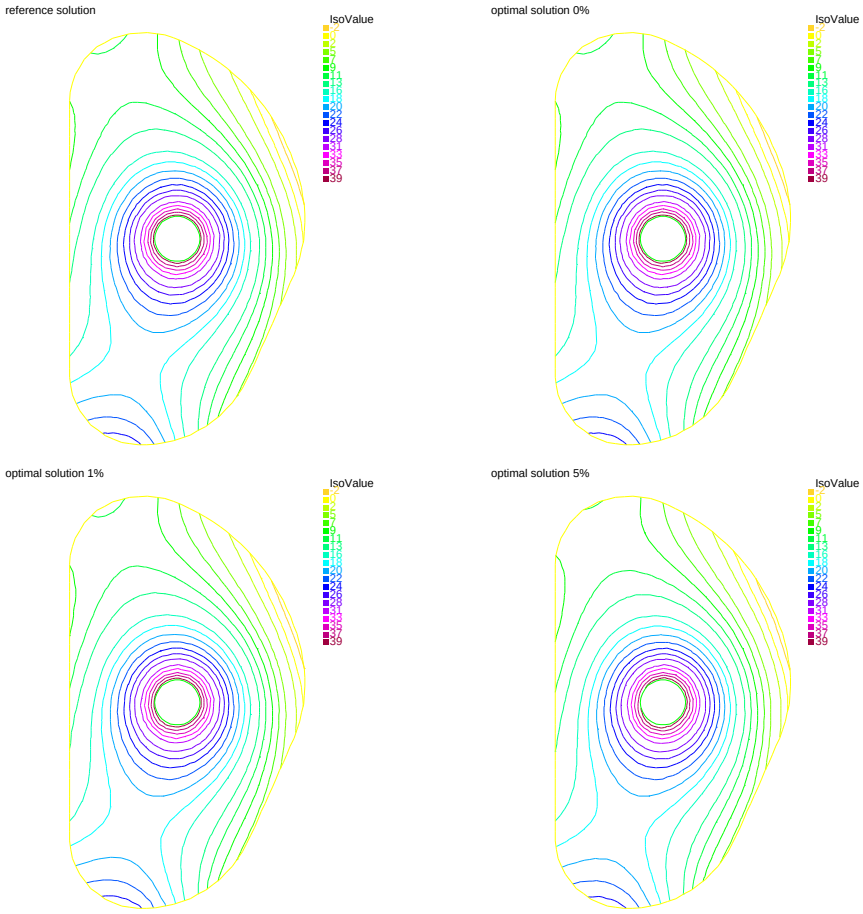

Figure 5. *Second test case (TC2), u_{ref} given by Eq. (39). Top left : reference solution $\psi_{N,ref}(u_{ref}, g)$. Top right : recovered solution with no noise on the data. Bottom left : recovered solution with a 1% noise on the data. Bottom right : recovered solution with a 5% noise.*

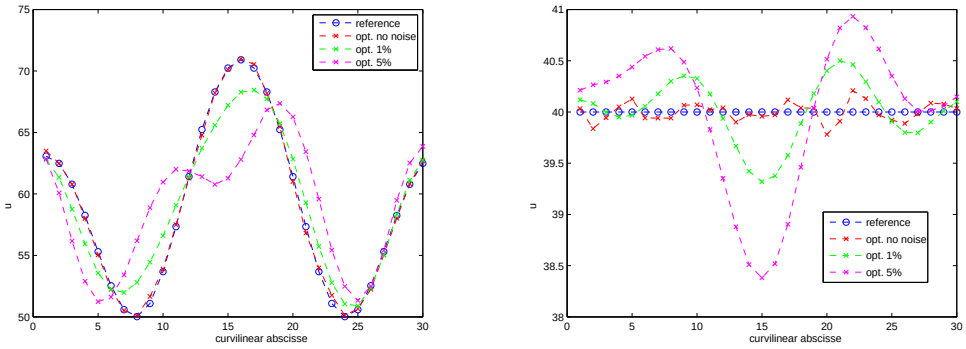

Figure 6. u_{ref} *and the recovered* u_{opt} *for the 3 levels of noise on the data. Left : TC1. Right TC2.*

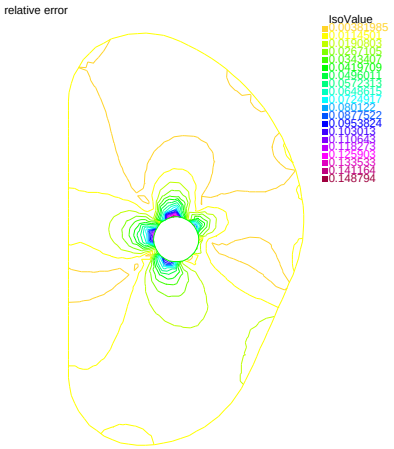

Figure 7. *Relative error* $|\psi_{opt} - \psi_{opt}|/|\psi_{ref}|$ *for TC1 with* 5% *noise.*

noise level	error TC1	error TC2
0%	0.0131	0.0055
1%	0.0659	0.0170
5%	0.1526	0.0405

Tableau 1. *Maximum relative error* $\frac{|u_{opt} - u_{ref}|}{|u_{ref}|}$ *for TC 1 and 2*

	J	R_D	J_ε	ε
$u = 0$ no noise	46.8643	0	46.8643	
u_{opt} no noise	0.0021	46.8722	0.0026	10^{-5}
u_{opt} 1% noise	1.8443	46.5553	1.8676	5×10^{-4}
u_{opt} 5% noise	9.2180	46.5575	9.2646	10^{-3}

Tableau 2. *TC1 results. Values of the error functional, the regularization term, the total cost function and the chosen regularization parameter for the initial guess (row 1), the optimal solutions for different noise levels (row 2, 3 and 4).*

	J	R_D	J_ε	ε
$u = 0$ no noise	30.7231	0	30.7231	
u_{opt} no noise	0.0003	30.7242	0.0006	10^{-5}
u_{opt} 1% noise	0.7300	30.7159	0.7607	10^{-3}
u_{opt} 5% noise	3.6516	30.6822	3.8050	5×10^{-3}

Tableau 3. *TC2 results. Values of the error functional, the regularization term, the total cost function and the chosen regularization parameter for the initial guess (row 1), the optimal solutions for different noise levels (row 2, 3 and 4).*

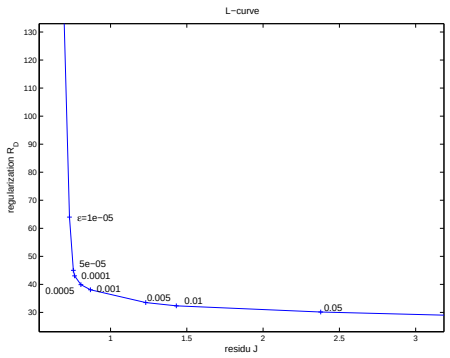

Figure 8. *L-curve computed for the ITER case. The corner is located at* $\varepsilon = 5 \times 10^{-4}$.

3.3. An ITER equilibrium

In this last numerical experiment we consider a 'real' ITER case. Measurements of the magnetic field are provided by the plasma equilibrium code CEDRES++ [17]. These mesurements are interpolated to provide f and g on Γ_V. The regularized error functional is then minimized to compute the optimal u_{opt}. The choice of the regularization parameter ε is made thanks to the computation of the L-curve shown on Fig. 8. It is a plot of $(J(u_{opt})(\varepsilon), R_D(u_{opt})(\varepsilon))$ as ε varies. The corner of the L-shaped curve provides a value of $\varepsilon = 5.10^{-4}$.

The computed u_{opt} is shown on Fig. 9 and numerical values are given in Tab. 4. The recovered poloidal flux ψ is shown on Fig. 10. The boundary of the plasma is found to be the isoflux $\psi = 16.3$ which shows an X-point configuration.

	J	R_D	J_ε	ε
$u = 0$	31.1026	0	31.1026	
u_{opt}	0.8053	39.9169	0.8253	5×10^{-4}

Tableau 4. *ITER case results. Values of the error functional, the regularization term, the total cost function and the chosen regularization parameter for the initial guess (row 1) and the optimal solution (row 2)*

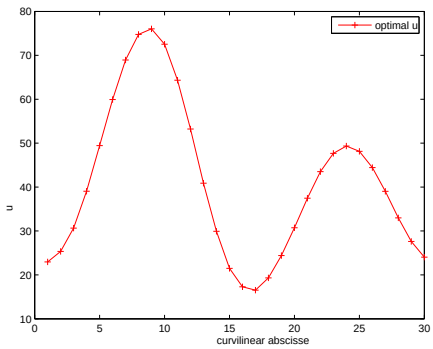

Figure 9. *Optimal u_{opt} for the ITER case.*

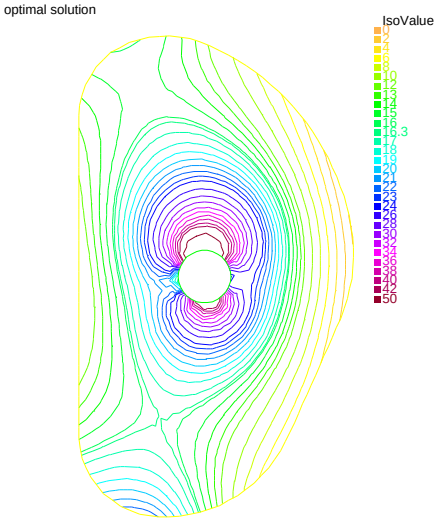

Figure 10. *Optimal solution for the ITER case. The plasma is in an X-point configuration*

4. Conclusion

We have presented a numerical method for the computation of the poloidal flux in the vacuum region surrounding the plasma in a Tokamak. The algorithm is based on the optimization of a regularized error functional. This computation enables in a second step the identification of the plasma boundary.

Numerical experiments have been conducted. They show that the method is precise and robust to noise on the Cauchy input data. It is fast since the optimization reduces to the resolution of a linear system of very reasonable dimension. Successive equilibrium reconstructions can be conducted very rapidly since the matrix of this linear system can be completely precomputed and only the right hand side has to be updated. The L-curve method proved to be efficient to specify the regularization parameter.

Appendix. Proof of Proposition 1

1. We need to prove the continuity and the coercivity of J_ε.
Continuity.
The maps $v \mapsto \psi_D(v)$ and $v \mapsto \psi_N(v)$ are continuous and linear from $H^{1/2}(\Gamma_I)$ to $H^1(\Omega)$. Moreover since $\tilde{\psi}_D(f)$ and $\tilde{\psi}_N(g)$ are in $H^1(\Omega)$ and $r_M \geq r \geq r_m > 0$ in Ω it is shown with Cauchy Schwarz that the bilinear forms s_D and s_N, the linear form l and thus J_ε are continuous on $H^{1/2}(\Gamma_I)$.

Coercivity.
The bilinear form s_D is coercive on $H^{1/2}(\Gamma_I)$. One obtains this from the fact that $\psi_D(v) \in H^1_0(\Omega, \Gamma_V)$ and the Poincaré inequality holds, and from the continuity of the application $\psi_D(v) \in H^1(\Omega) \to \psi_D(v)|_{\Gamma_I} = v \in H^{1/2}(\Gamma_I)$.

On the contrary, since for $\psi_N(v) \in H^1(\Omega)$ the seminorm does not bound the L^2 norm, the bilinear form s_N is not coercive and because of the minus sign in $s = s_D - s_N$ we need to prove that $s(v, v) \geq 0$ to obtain the coercivity of the bilinear part of functional J_ε. One can use the same type of argument as in [5] to de so.

Eventually it holds that

$$\frac{1}{2}s(v,v) + \frac{\varepsilon}{2}s_D(v,v) \geq C\varepsilon||v||^2_{H^{1/2}(\Gamma_I)}$$

Using the continuity and the coercivity of J_ε it results from [25] that problem P_ε admits a unique solution $u_\varepsilon \in \mathcal{U}$.

The solution u_ε is characterized by the first order optimality condition which is written as the following well-posed variational problem

$$(J'_\varepsilon(u_\varepsilon), v) = \varepsilon s_D(u_\varepsilon, v) + s_D(u_\varepsilon, v) - s_N(u_\varepsilon, v) - l(v) = 0 \quad \forall v \in \mathcal{U} \qquad (40)$$

which as in Eq. (26) can be understood as an equality on Γ_I.

2. The stability result is deduced from the optimality condition (40).

Let u^1_ε (resp. u^2_ε) be the solution associated to (f^1, g^1) (resp. (f^2, g^2)). Substracting the two optimality conditions, choosing $v = u^1_\varepsilon - u^2_\varepsilon$ and using the coercivity leads to

$$C\varepsilon||u^1_\varepsilon - u^2_\varepsilon||^2_{H^{1/2}(\Gamma_I)} \leq |(l_1 - l_2)(u^1_\varepsilon - u^2_\varepsilon)|$$

The map $f \mapsto \tilde{\psi}_D(f)$ is linear and continuous from $H^{1/2}(\Gamma_V)$ to $H^1(\Omega)$, and so is the map $g \mapsto \tilde{\psi}_N(g)$ from $H^{-1/2}(\Gamma_V)$ to $H^1(\Omega)$. Using these facts and Cauchy Schwarz it follows that

$$||u_\varepsilon^1 - u_\varepsilon^2||_{H^{1/2}(\Gamma_I)} \leq \frac{C'}{r_m C}\frac{1}{\varepsilon}(||f^1 - f^2||_{H^{1/2}(\Gamma_V)} + ||g^1 - g^2||_{H^{-1/2}(\Gamma_V)})$$

3. For this point the proof of Proposition 3.2 in [3] can be adpated. A sketch of the proof is as follows. Let us suppose that there exists $u \in \mathcal{U}$ such that $\psi_D(u, f) = \psi_N(u, g)$. A key point is to show that $s_D(u_\varepsilon, u_\varepsilon) \to s_D(u, u)$ when $\varepsilon \to 0$. Then in a second step using the optimality conditions for u and u_ε it is shown that

$$s_D(u_\varepsilon - u, u_\varepsilon - u) \leq s_D(u, u) - s_D(u_\varepsilon, u_\varepsilon)$$

which gives the result thanks to the coercivity of s_D in $H^{1/2}(\Gamma_I)$.

5. Bibliographie

[1] S. Andrieux, T.N. Baranger, and A. Ben Abda. Solving cauchy problems by minimizing an energy-like functional. *Inverse Problems*, 22 :115–133, 2006.

[2] S. Andrieux, A. Ben Abda, and T.N. Baranger. Data completion via an energy error functional. *C.R. Mecanique*, 333 :171–177, 2005.

[3] M. Azaiez, F. Ben Belgacem, and H. El Fekih. On Cauchy's problem : II. Completion, regularization and approximation. *Inverse Problems*, 22 :1307–1336, 2006.

[4] A Ben Abda, M. Hassine, M. Jaoua, and M. Masmoudi. Topological sensitivity analysis for the location of small cavities in stokes flows. *SIAM J. Cont. Opt.*, 2009.

[5] F. Ben Belgacem and H. El Fekih. On Cauchy's problem : I. A variational Steklov-Poincaré theory. *Inverse Problems*, 21 :1915–1936, 2005.

[6] J. Blum. *Numerical Simulation and Optimal Control in Plasma Physics with Applications to Tokamaks*. Series in Modern Applied Mathematics. Wiley Gauthier-Villars, Paris, 1989.

[7] M. Bonnet and A. Constantinescu. Inverse problems in elasticity. *Inverse Problems*, 21(2), 2005.

[8] L. Bourgeois and J. Dardé. A quasi-reversibility approach to solve the inverse obstacle problem. *Inverse Problems and Imaging*, 4/3 :351–377, 2010.

[9] B.J. Braams. The interpretation of tokamak magnetic diagnostics. *Nuc. Fus.*, 33(7) :715–748, 1991.

[10] S. Chaabane and M. Jaoua. Identification of robin coefficients by means of boundary measurements. *Inverse Problems*, 15(6) :1425, 1999.

[11] P.G. Ciarlet. *The Finite Element Method For Elliptic Problems*. North-Holland, 1980.

[12] R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 1-2. Interscience, 1962.

[13] W. Feneberg, K. Lackner, and P. Martin. Fast control of the plasma surface. *Computer Physics Communications*, 31(2) :143–148, 1984.

[14] Y. Fischer. *Approximation dans des classes de fonctions analytiques généralisées et résolution de problèmes inverses pour les tokamaks*. Phd thesis, Université de Nice Sophia Antipolis, France, 2011.

[15] Y. Fischer, B. Marteau, and Y. Privat. Some inverse problems around the tokamak tore supra. *Comm. Pure and Applied Analysis*, to appear.

[16] H. Grad and H. Rubin. Hydromagnetic equilibria and force-free fields. In *2nd U.N. Conference on the Peaceful uses of Atomic Energy*, volume 31, pages 190–197, Geneva, 1958.

[17] V. Grandgirard. *Modélisation de l'équilibre d'un plasma de tokamak - Tokamak plasma equilibrium modelling*. Phd thesis, Université de Besançon, France, 1999.

[18] H. Haddar and R. Kress. Conformal mappings and inverse boundary value problem. *Inverse Problems*, 21 :935–953, 2005.

[19] C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems : Numerical Aspects of Linear Inversion*. SIAM, Philadelphia, 1998.

[20] R.V. Kohn and A. McKenney. Numerical implementation of a variational method for electrical impedance tomography. *Inverse Problems*, 6(3) :389, 1990.

[21] R.V. Kohn and M.S. Vogelius. Determining conductivity by boundary measurements : II. Interior results. *Commun. Pure Appl. Math.*, 31 :643–667, 1985.

[22] R.V. Kohn and M.S. Vogelius. Relaxation of a variational method for impedance computed tomography. *Commun. Pure Appl. Math.*, 11 :745–777, 1987.

[23] K. Lackner. Computation of ideal MHD equilibria. *Computer Physics Communications*, 12 :33–44, 1976.

[24] P. Ladeveze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Num. Anal.*, 20(3) :485–509, 1983.

[25] J.L. Lions. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles (Optimal control of systems governed by partial differential equations)*. Dunod, Paris, 1968.

[26] D.P. O'Brien, J.J Ellis, and J. Lingertat. Local expansion method for fast plasma boundary identification in JET. *Nuc. Fus.*, 33(3) :467–474, 1993.

[27] A. Quarteroni and V Alberto. *Domain Decomposition Methods for Partial Differential Equations*. Oxford University Press, 1999.

[28] F. Saint-Laurent and G. Martin. Real time determination and control of the plasma localisation and internal inductance in Tore Supra. *Fusion Engineering and Design*, 56-57 :761–765, 2001.

[29] F. Sartori, A. Cenedese, and F. Milani. JET real-time object-oriented code for plasma boundary reconstruction. *Fus. Engin. Des.*, 66-68 :735–739, 2003.

[30] V.D. Shafranov. On magnetohydrodynamical equilibrium configurations. *Soviet Physics JETP*, 6(3) :1013, 1958.

[31] V.D. Shafranov and L.E. Zakharov. Use of the virtual-casing principle in calculating the containing magnetic field in toroidal plasma systems. *Nuc. Fus.*, 12 :599–601, 1972.

[32] J. Wesson. *Tokamaks*, volume 118 of *International Series of Monographs on Physics*. Oxford University Press Inc., New York, Third Edition, 2004.

Article E : [9] H. Heumann, J. Blum, C. Boulbe, B. Faugeras, G. Selig, J.-M. Ané, S. Brémond, V. Grangirard, P. Hertout et E. Nardon. Quasi-static free-boundary equilibrium of toroidal plasma with CEDRES++ : computational methods and applications. *J. Plasma. Phys.* (2015). doi : 10.1017/S0022377814001251

J. Plasma Physics: page 1 of 35.
doi:10.1017/S0022377814001251

Quasi-static free-boundary equilibrium of toroidal plasma with CEDRES++: Computational methods and applications

H. Heumann[1]†, J. Blum[1], C. Boulbe[1], B. Faugeras[1], G. Selig[1], J.-M. Ané[2], S. Brémond[2], V. Grandgirard[2], P. Hertout[2] and E. Nardon[2]

[1]TEAM CASTOR, INRIA, 2004 Route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France and Laboratoire J.A. Dieudonné, UMR 7351, Université de Nice Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 02, France

[2]CEA, IRFM, F-13108 Saint-Paul-lez-Durance, France

(Received 10 October 2014; revised 21 November 2014; accepted 25 November 2014)

We present a comprehensive survey of various computational methods in CEDRES++ (Couplage Equilibre Diffusion Résistive pour l'Etude des Scénarios) for finding equilibria of toroidal plasma. Our focus is on free-boundary plasma equilibria, where either poloidal field coil currents or the temporal evolution of voltages in poloidal field circuit systems are given data. Centered around a piecewise linear finite element representation of the poloidal flux map, our approach allows in large parts the use of established numerical schemes. The coupling of a finite element method and a boundary element method gives consistent numerical solutions for equilibrium problems in unbounded domains. We formulate a new Newton method for the discretized nonlinear problem to tackle the various nonlinearities, including the free plasma boundary. The Newton method guarantees fast convergence and is the main building block for the inverse equilibrium problems that we can handle in CEDRES++ as well. The inverse problems aim at finding either poloidal field coil currents that ensure a desired shape and position of the plasma or at finding the evolution of the voltages in the poloidal field circuit systems that ensure a prescribed evolution of the plasma shape and position. We provide equilibrium simulations for the tokamaks ITER and WEST to illustrate the performance of CEDRES++ and its application areas.

1. Introduction

Computer codes that address the equilibrium of toroidal plasmas are central tools in tokamak fusion science. They are essential, both for detailed simulations with sophisticated magnetohydrodynamic (MHD) models as well as for experimenters that need to control real tokamak reactors. Detailed MHD simulations, which model the plasma on very short timescales, are used to study the various effects of turbulence and instability. They rely on a given plasma equilibrium as initial condition. Experimenters use equilibrium codes to set up discharge scenarios, to study breakdowns and disruptions, or to design the layout of new machines. They also use such codes, in connection with transport codes (Hinton and Hazeltine 1976; Hirshman and Jardin 1979; Artaud et al. 2010; Coster et al. 2010; Parail et al. 2013), to design

† Email address for correspondence: holger.heumann@inria.fr

and validate plasma feedback controller for real tokamak machines and to verify the feasibility of scenarios in terms of operational limits (e.g. coil currents or forces).

Hence, equilibrium codes are essential tools for tokamak scientists, and applicants expect a certain degree of maturity and robustness. In the design of discharge scenarios or in the validation of feedback controller, for example, a robust, fast, and automated computation of equilibria allows to shift the focus of research towards the difficulties of coupling with complex physics or improved control algorithms. CEDRES++ deals with equilibrium problems that are related to a quasi-static description of plasma evolution, which asserts balance of forces at each instant of time. A code that treats such *quasi-static free-boundary equilibrium problems* needs to solve nonlinear elliptic or parabolic problems with nonlinear source terms representing the current density profile, that vanishes outside the unknown free boundary of the plasma. The computational challenges in the design of free-boundary equilibrium codes are a problem setting in an unbounded domain with a nonlinearity due the current density profile in the unknown plasma domain and the nonlinear magnetic permeability if the reactor has ferromagnetic structures.

The simulation on the unbounded domain can be reduced to computations on a finite domain thanks to analytical Green's functions (Lackner 1976). The numerical solution on the finite interior domain is coupled through boundary conditions to the Green's function representation of the solution in the exterior domain. This approach is today fairly standard in many other application areas such as electromagnetics (Hiptmair 2003; Zhao et al. 2006) or elasticity (Costabel and Stephan 1990; Bielak and MacCamy 1991; Stephan 1992). The *boundary element method* (see the review article (Costabel 1987) or the text books (Chen and Zhou 1992; Nédélec 2001)) is the name of this general framework. The boundary element method reduces problems on unbounded domains to problems on boundaries that can then be coupled to any numerical method for the interior of a bounded domain.

The nonlinearity due to the current profile in the unknown plasma domain poses the major difficulties according to our experience. It is a peculiarity of plasma equilibrium problems, that the domain of the plasma is an unknown. Speaking differently, the boundary of the plasma is a free boundary, defined either by a contact with a limiter which prevents the plasma from touching the vacuum vessel, or defined as being a separatrix in the case of a poloidal divertor configuration. On the top of this fairly unusual kind of nonlinearity, also the current profile in the plasma itself is a nonlinear function. Moreover, in the so-called iron transformer tokamaks, a third type of nonlinearity appears due to the nonlinear magnetic permeability. All these nonlinearities will require some iterations towards the numerical solution. Simple fixed-point iterations usually suffer from very slow convergence or even fail to converge, which made researchers move towards Newton-type methods. The latter use the information of gradients, sometimes also referred to as sensitivities, to speed up the convergence, and they can converge in cases where fixed-point iterations do not converge – a very important example is vertically unstable plasmas.

There are basically two different families of solution methods for axisymmetric plasma equilibrium problems. The first family are the so-called flux or Lagrangian coordinate methods, determining the localization of level lines that have equidistant flux-values (DeLucia et al. 1980; Lao et al. 1981; Degtyarev and Drozdov 1985; Lao et al. 1985; Ling and Jardin 1985; Jardin et al. 1986; Gruber et al. 1987; Degtyarev and Drozdov 1991; Turkington et al. 1993) (see also Jardin (2010, Sec. 5.5)). A second family of methods uses standard finite difference methods on rectangular grids (Feneberg and Lackner 1973; Lackner 1976; Helton and Wang 1978; Johnson

et al. 1979) or finite element methods on triangular grids (Blum et al. 1981). The main difference between most methods of both of these families is the treatment of the so-called fixed boundary equilibrium problem, i.e. a problem where the plasma domain is known. The computational issues related to the unknown boundary have received less attention.

The CEDRES++ code uses a finite element formulation for the axisymmetric free-boundary equilibrium problem in the interior domain. This allows first, for standard, well-established coupling methods to the boundary element formulation on the exterior domain (Albanese et al. 1986). Second, we can derive a perfect Newton method, that uses the information about all nonlinearities, e.g. also those related to the free-boundary setting. We consider this to be the most distinctive feature of CEDRES++ among many other equilibrium codes. Up to our knowledge there is no other equilibrium solver that uses this information to speed up the convergence. Furthermore, accurate derivatives are vital for inverse free-boundary equilibrium problems, which aim at finding the values of control parameters that ensure that the plasma attains a certain desired state, i.e. shape or position. Inverse free-boundary equilibrium problems are formulated as constrained optimization problems and only accurately computed derivatives can guarantee that the optimization algorithms find indeed the optimum. For the moment, CEDRES++ uses linear Lagrangian elements, which due to the low regularity of the solution, seem to be the obvious choice. We would like to refer to Sec. 5 for a general discussion on this topic.

CEDRES++ inherits the basic ideas of the free-boundary equilibrium codes SCED (Blum et al. 1981) and Proteus (Albanese et al. 1987) but relies on object oriented and modular programming principles. CEDRES++ uses well established and tested external modules for e.g. mesh generation (Shewchuk 1996), linear algebra (Renard and Pommier 2014) and algebraic solver (Davis 2011). The very first conception of CEDRES++, that used the same methods as SCED and Proteus, was developed in (Grandgirard 1999). Various simulations with this old version of CEDRES++ are reported in Grandgirard (1999) and Hertout et al. (2011).

The current version of CEDRES++ contains a new module that, when coupled to a transport code, simulates a quasi-statically evolving equilibrium: the classical Grad–Shafranov equation, a nonlinear elliptic partial differential equation, is satisfied at each instant of time. This mode assumes that the evolution of voltages in poloidal field circuits and the nonlinearities in plasma current profile are known. The new mode is referred to as the *evolution* mode as opposed to the *static* mode that takes poloidal field coil currents and the current density as input. Within the new evolution mode, we solve the full parabolic partial differential equation system. We do not have to estimate the nonlinear mutual inductance of the plasma with the electromagnetic reactor components as the approach in Albanese and Villone (1998) and (Ariola and Pironti 2008, Chapter 2) would require. All the dynamics of the plasma core related to resistive diffusion of magnetic flux and transport of particle density and temperatures, are supposed to be treated by external tools and are not subject of this report. We refer to Falchetto et al. (2014) for the coupling of CEDRES++ (Couplage Equilibre Diffusion Résistive pour l'Etude des Scénarios i.e. Coupling of Equilibrium and Resistive Diffusion for the Evaluation of Scenarios) with the transport code ETS (Coster et al. 2010). CEDRES++ is also coupled to the transport code CRONOS (Artaud et al. 2010). The evolution mode used with prescribed evolution of the current profile is also a good practical approach for vertical stability studies, where the timescale of interest is much shorter than the current diffusion timescale of the plasma.

Further, CEDRES++ can solve inverse free-boundary equilibrium problems. The inverse problem in the static mode aims at finding poloidal field coil currents that ensure a desired shape and position of the plasma. The inverse problem in the evolution mode aims at finding the evolution of the voltages in the poloidal field circuits that ensure a prescribed evolution of the plasma shape and position. We use standard algorithms for constrained optimization to solve the inverse problems. Therefore it will be straightforward to add in the near future further constraints, such as constraints on the flux consumption or the currents in the coils. In Table 1 we summarize the basic CEDRES++ modes and their areas of application.

Previous implementations of the Newton method in SCED (Blum et al. 1981) and Proteus (Albanese et al. 1987) relied on the discretization of a Newton method formulated on a continuous level. It is not clear, whether this formulation remains valid for equilibria with plasma boundaries in the case of a poloidal divertor configuration. The distinctive new feature of CEDRES++ is a Newton method, that solves the discretized nonlinear equations. Our new approach has more rigorous mathematical foundations and is supposed to have slightly faster convergence. Moreover, it is only this new approach, which guarantees that the optimization algorithms for solving the inverse problems converge to the correct solution. Section 3 gives more explanations on that.

The users of CEDRES++ do not need to know about details of the algorithms and the parameters. CEDRES++ is a robust, fast, and accurate and an easily usable tool. CEDRES++ focuses for the moment on the solution of the so-called *axisymmetric free-boundary plasma equilibrium with isotropic pressure and without flow*. The assumption of perfect axial symmetry is a common model reduction in many equilibrium applications and the treatment of *3D plasma equilibria* (Hirshman and Betancourt 1991; Park et al. 1999) requires still a lot of computational power. We are planning to include in the near future numerical methods for *plasma equilibria with flow* and *plasma equilibria with anisotropic pressure* (Grad 1967; Maschke and Perrin 1984; Goedbloed and Lifschitz 1997; Zwingmann et al. 2001; Guazzotto et al. 2004; Cooper et al. 2009; Pustovitov 2010; Fitzgerald et al. 2013). Toroidal equilibria with anisotropic pressure and flow are an active area of research that will benefit from our contribution to the computation of free-boundary equilibria. CEDRES++ is not considered to be used as a so-called *equilibrium reconstruction code* (Hofmann and Tonetti 1988; Lao et al. 1990; Mc Carthy et al. 1999; Blum et al. 2012), which relies on measurements during the discharge to compute the magnetic fields and estimates of current profiles and other characteristics of plasma equilibria.

The outline of the article is the following: In the first section, we recall briefly the basic equations that describe the free-boundary plasma equilibrium in a tokamak and state the four main problems that can be solved with CEDRES++. The subsequent section contains detailed descriptions of the various numerical methods that are implemented in CEDRES++. This is followed by a short section containing tests for the numerical validation and various application examples.

2. Quasi-static free-boundary equilibrium of toroidal plasma

The essential equations for describing plasma equilibrium in a tokamak are force balance, the solenoidal condition and Ampère's law

$$\operatorname{grad} p = \mathbf{J} \times \mathbf{B}\,, \quad \operatorname{div} \mathbf{B} = 0\,, \quad \operatorname{curl} \frac{1}{\mu}\mathbf{B} = \mathbf{J}\,, \tag{2.1}$$

CEDRES++ mode	Functionality	Application areas
Static, direct	Simulates plasma equilibria for given poloidal field coil currents	Reference equilibria, initial conditions for short timescale plasma models
Static, inverse	Finds the poloidal field coil currents which allow the best match with a given plasma shape at a fixed time	Preparation of scenarios
Evolution, direct	Simulates the quasi-static evolution of plasma equilibrium for given poloidal field circuit voltages	Coupling to transport codes, design and test environment for feedback controller, study of breakdowns and disruptions.
Evolution, inverse	Finds optimal poloidal field circuit voltages and desired evolution of the plasma	Feedforward control for discharge scenario optimization.

TABLE 1. The functionality and the application areas of the four different modes of CEDRES++.

where p is the plasma kinetic pressure, $\mathbf{B}$ is the magnetic field, $\mathbf{J}$ is the current density and μ the magnetic permeability. In the quasi-static approximation these static equations are augmented by Faraday's law

$$-\partial_t \mathbf{B} = \operatorname{curl} \mathbf{E}\,, \tag{2.2}$$

with $\mathbf{E}$ the electric field, and by Ohm's laws in plasma, coils, and passive structures.

For the calculations in CEDRES++ we will differentiate between static problems and evolution problems, where the keyword static indicates that the equations do not give a time-varying solution. The static problems and the evolution problems are treated by CEDRES++ *static* mode and *evolution* mode, respectively. Force balance, solenoidal condition and Ampère's law in (2.1) yield the static problem we will introduce in detail in Sec. 2.1, while the evolution problems introduced afterwards in Sec. 2.3, take also into account the Faraday's and Ohm's law in the poloidal field coils and in the passive structures. All the dynamics due to Faraday's and Ohm's laws in the plasma, as well as the dynamics related to transport of heat and particles are supposed to be treated by external tools.

Under the common assumption of perfect axial symmetry, it is convenient to put (2.1) and the quasi-static approximation of Maxwell's equations in a cylindrical coordinate system (r, φ, z) and to consider only a meridian section of the tokamak. Then, Faraday's and Ampère's laws decouple into corresponding laws in the toroidal direction and the poloidal plane and give rise to the toroidal and poloidal Poynting theorems:

$$\operatorname{div} \frac{\mathbf{B}_{tor} \times \mathbf{E}_{pol}}{\mu} = \mathbf{J}_{pol} \cdot \mathbf{E}_{pol} + \frac{1}{2\mu} \partial_t \mathbf{B}_{tor}^2 \tag{2.3}$$

and

$$\operatorname{div} \frac{\mathbf{B}_{pol} \times \mathbf{E}_{tor}}{\mu} = \mathbf{J}_{tor} \cdot \mathbf{E}_{tor} + \frac{1}{2\mu} \partial_t \mathbf{B}_{pol}^2, \tag{2.4}$$

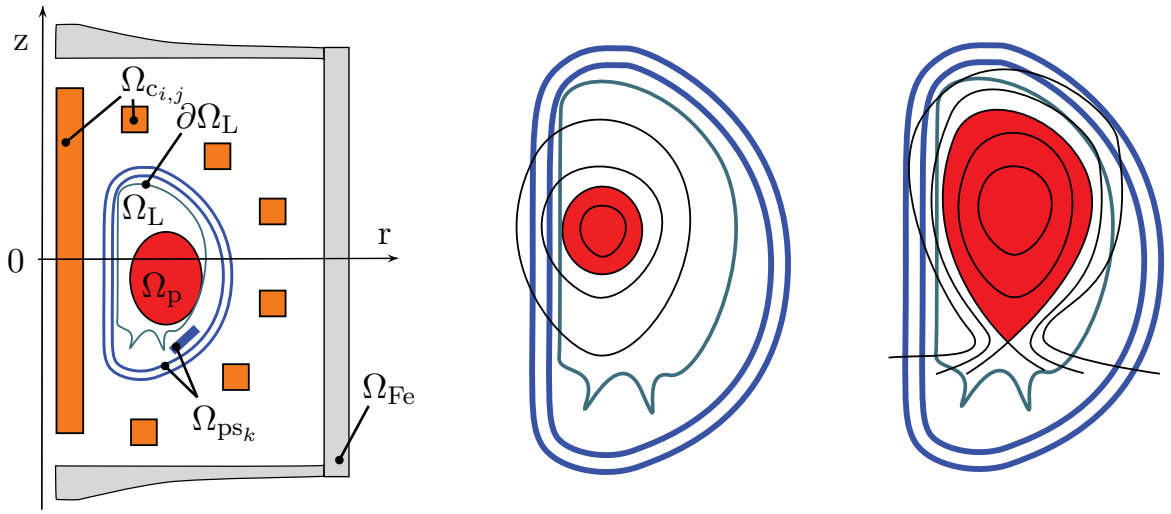

FIGURE 1. Left: Geometric description of the tokamak in the poloidal plane. Middle and right: Sketch for characteristic plasma shapes. The plasma boundary touches the limiter (middle) or the plasma is enclosed by a flux line that goes through an X-point (right).

where the subscripts $_{tor}$ and $_{pol}$ indicate the toroidal and poloidal components of the fields. Poynting theorems can be used to check the conservation of energy in the simulation of quasi-static plasma equilibria, thus providing a global control of the accuracy.

We introduce $\Omega_\infty = [0, \infty] \times [-\infty, \infty]$, the positive half plane, to denote the meridian plane that contains the tokamak centered at the origin. The geometry of the tokamak determines the various subdomains (see Fig. 1):

- the domain $\Omega_{\mathrm{Fe}} \subset \Omega_\infty$ corresponds to those parts that are made of iron; for an air-transformer tokamak $\Omega_{\mathrm{Fe}} = \emptyset$;
- the domains $\Omega_{c_{i,j}} \subset \Omega_\infty$, $1 \leqslant i \leqslant L$, $1 \leqslant j \leqslant N_i$, correspond to the $\sum_{i=1}^{L} N_i = N$ poloidal field coils. The coils are grouped into L poloidal field circuits and the ith circuit contains N_i coils. The intersection of the jth coil in the ith circuit with the poloidal plane is $\Omega_{c_{i,j}}$, and it has $n_{i,j}$ wire turns, total resistance $R_{i,j}$ and cross-section area $S_{i,j}$;
- the domains $\Omega_{ps_k} \subset \Omega_\infty$, $k = 1, \ldots, N_{ps}$ corresponding to N_{ps} passive structures with conductivity σ_k;
- the domain $\Omega_L \subset \Omega_\infty$, bounded by the limiter, corresponds to the domain which is accessible by the plasma;
- the domain $\Omega_p \subset \Omega_L$, is the domain covered by the plasma.

The classical primal unknowns for toroidal plasma equilibria described by (2.1) are the *poloidal magnetic flux* $\psi = \psi(r, z)$, the pressure p and the *diamagnetic function* f. The poloidal magnetic flux $\psi := r\mathbf{A} \cdot \mathbf{e}_\varphi$ is the scaled toroidal component of the vector potential $\mathbf{A}$, i.e. $\mathbf{B} = \operatorname{curl} \mathbf{A}$ and $\mathbf{e}_\varphi$ the unit vector for φ. The diamagnetic function $f = r\mathbf{B} \cdot \mathbf{e}_\varphi$ is the scaled toroidal component of the magnetic field. It can be shown that both the pressure p and the diamagnetic function f are constant on ψ -isolines, i.e. $p = p(\psi)$ and $f = f(\psi)$. We refer to standard text books, e.g. Freidberg (1987), Blum (1989), Wesson (2004), Goedbloed and Poedts (2004), Goedbloed et al. (2010) and Jardin (2010) for the details and state in the following paragraphs only the final equations describing the static and evolution problems solved in CEDRES++.

2.1. *Direct static problem*

Force balance, solenoidal condition and Ampère's law in (2.1) yield in axisymmetric configuration the following set of equations:

$$-\nabla\cdot\left(\frac{1}{\mu r}\nabla\psi\right)=\begin{cases} rp'(\psi)+\frac{1}{\mu_0 r}ff'(\psi) & \text{in } \Omega_{\mathrm{p}}(\psi)\,;\\ \frac{I_{i,j}}{S_{i,j}} & \text{in } \Omega_{\mathrm{c}_{i,j}}\,;\\ 0 & \text{elsewhere}\,,\end{cases} \qquad \psi(0,z)=0\,;\quad \lim_{\|(r,z)\|\to+\infty}\psi(r,z)=0\,; \tag{2.5}$$

where ∇ is the gradient in the two dimensions (r, z), $I_{i,j}$ is the total current (in Ampère turns) in the j-th coil of the ith circuit and μ is a functional of ψ

$$\mu=\begin{cases}\mu_{\mathrm{Fe}}(|\nabla\psi|^2 r^{-2}) & \text{in } \Omega_{\mathrm{Fe}}\\ \mu_0 & \text{elsewhere.}\end{cases} \tag{2.6}$$

with μ_0 the constant magnetic permeability of vacuum and μ_{Fe} the nonlinear magnetic permeability of iron. Here again, we would like to stress that the plasma domain $\Omega_{\mathrm{p}}(\psi)$ is an unknown, which depends nonlinearly on the magnetic flux ψ: the plasma domain $\Omega_{\mathrm{p}}(\psi)$ is a functional of the poloidal flux ψ. The different characteristic shapes of $\Omega_{\mathrm{p}}(\psi)$ are illustrated in Fig. 1: the boundary of $\Omega_{\mathrm{p}}(\psi)$ either touches the boundary of Ω_{L} (limiter case) or the boundary contains one or more saddle points of ψ (divertor configuration). The saddle points of ψ, denoted by $(r_{\mathrm{X}}, z_{\mathrm{X}})=(r_{\mathrm{X}}(\psi), z_{\mathrm{X}}(\psi))$, are called X-points of ψ. The plasma domain $\Omega_{\mathrm{p}}(\psi)$ is the largest subdomain of Ω_{L} bounded by a closed ψ-isoline in Ω_{L} and containing the magnetic axis $(r_{\mathrm{ax}}, z_{\mathrm{ax}})$. The magnetic axis is the point $(r_{\mathrm{ax}}, z_{\mathrm{ax}}) = (r_{\mathrm{ax}}(\psi), z_{\mathrm{ax}}(\psi))$, where ψ has its global maximum in Ω_{L}. For convenience, we introduce also the coordinates $(r_{\mathrm{bnd}}, z_{\mathrm{bnd}}) = (r_{\mathrm{bnd}}(\psi), z_{\mathrm{bnd}}(\psi))$ of the point that determines the plasma boundary. $(r_{\mathrm{bnd}}, z_{\mathrm{bnd}})$ is either an X-point of ψ or the contact point with the limiter $\partial\Omega_{\mathrm{L}}$.

The (2.5) in the plasma domain, i.e.

$$-\frac{\partial}{\partial r}\left(\frac{1}{\mu_0 r}\frac{\partial\psi}{\partial r}\right)-\frac{\partial}{\partial z}\left(\frac{1}{\mu_0 r}\frac{\partial\psi}{\partial z}\right)=rp'(\psi)+\frac{1}{\mu_0 r}ff'(\psi), \tag{2.7}$$

is the celebrated *Grad–Shafranov–Schlüter* equation (Lüst and Schlüter 1957; Grad and Rubin 1958; Shafranov 1958). The domain of p' and $f\,f'$ is the interval $[\psi_{\mathrm{bnd}}, \psi_{\mathrm{ax}}]$ with the scalar values ψ_{ax} and ψ_{bnd} being the flux values at the *magnetic axis* and at the boundary of the plasma:

$$\begin{aligned}\psi_{\mathrm{ax}}(\psi) &:= \psi(r_{\mathrm{ax}}(\psi), z_{\mathrm{ax}}(\psi))\,,\\ \psi_{\mathrm{bnd}}(\psi) &:= \psi(r_{\mathrm{bnd}}(\psi), z_{\mathrm{bnd}}(\psi))\,.\end{aligned} \tag{2.8}$$

The two functions p' and $f\,f'$ and the currents $I_{i,j}$ in the coils are not determined by the model (2.5) and have to be supplied as data. Since the domain of p' and $f\,f'$ depends on the poloidal flux itself, it is more practical to supply those profiles as functions of the normalized poloidal flux $\psi_{\mathrm{N}}(r, z)$:

$$\psi_{\mathrm{N}}(r,z)=\frac{\psi(r,z)-\psi_{\mathrm{ax}}(\psi)}{\psi_{\mathrm{bnd}}(\psi)-\psi_{\mathrm{ax}}(\psi)}. \tag{2.9}$$

These two functions, subsequently termed $S_{p'}$ and $S_{ff'}$, have, independently of ψ, a fixed domain $[0, 1]$.

Further, in many applications, one assumes that the current profile, i.e. the function $rS_{p'} + \frac{1}{\mu_0 r} S_{ff'}$, is only known up to some scaling constant λ. In those cases the set of equations in (2.5) has to be augmented by an additional equation that matches the scaling with the given total plasma current I_P.

Let us state the two problems that we will consider in the following.

PROBLEM 1 (DIRECT STATIC). *Let $S_{p'} : [0,1] \to \mathbb{R}$ and $S_{ff'} : [0,1] \to \mathbb{R}$ be two known functions and let the currents $I_{i,j}$ in the coils be given. We want to find the ψ such that* (2.5) *holds with $p'(\psi) = S_{p'}(\psi_\mathrm{N})$ and $ff'(\psi) = S_{ff'}(\psi_\mathrm{N})$.*

PROBLEM 2 (DIRECT STATIC, WITH GIVEN PLASMA CURRENT I_P). *Let $S_{p'} : [0,1] \to \mathbb{R}$ and $S_{ff'} : [0,1] \to \mathbb{R}$ be two known functions and let the currents $I_{i,j}$ in the coils be given. Additionally we assume that the total plasma current I_P is given. We want to find ψ and λ such that* (2.5) *holds with $p'(\psi) = \lambda S_{p'}(\psi_\mathrm{N})$ and $ff'(\psi) = \lambda S_{ff'}(\psi_\mathrm{N})$ together with*

$$I_\mathrm{P} = \lambda \int_{\Omega_\mathrm{p}(\psi)} \left(r S_{p'}(\psi_\mathrm{N}(r,z)) + \frac{1}{\mu_0 r} S_{ff'}(\psi_\mathrm{N}(r,z)) \right) dr dz. \tag{2.10}$$

The functions $S_{p'}$ and $S_{ff'}$ are usually given as piecewise polynomial functions. Another frequent a priori model is

$$S_{p'}(\psi_\mathrm{N}) = \frac{\beta}{r_0}(1 - \psi_\mathrm{N}^\alpha)^\gamma\,, \quad S_{ff'}(\psi_\mathrm{N}) \;=\; (1-\beta)\mu_0 r_0 (1 - \psi_\mathrm{N}^\alpha)^\gamma\,, \tag{2.11}$$

with r_0 the major radius of the vacuum chamber and $\alpha, \beta, \gamma \in \mathbb{R}$ given parameters. We refer to Luxon and Brown (1982) for a physical interpretation of these parameters. The parameter β is related to the poloidal beta, whereas α and γ describe the peakage of the current profile.

2.2. *Inverse static problem*

The direct problem in the previous section computes a free-boundary equilibrium for given coil currents. In many applications, in particular in the area of tokamak operation, the inverse problem is equally relevant: *What are the currents that give a certain desired shape to the plasma?* A popular approach to answer such a question is its formulation as an optimal control problem. The currents I_i in the poloidal field coils are the *control variables* and the magnetic flux map ψ describing the equilibrium is the controlled variable. Then we introduce a cost function for the magnetic flux ψ and the coil currents I_i penalizing the deviation from a desired plasma shape and position, and we minimize this cost function under the constraint that the magnetic flux ψ and the currents in the coils solve the equilibrium problem (2.5). A regularization term ensures well posedness of the inverse problem. Here again, the current profile in the plasma is supposed to be known data.

In CEDRES++, we prescribe a plasma state by a desired plasma boundary Γ_desi. Let $\Gamma_\mathrm{desi} \subset \Omega_\mathrm{L}$ denote a closed line, contained in the domain Ω_L that is either smooth and touches the limiter at one point or has at least one corner. The former case prescribes a desired plasma boundary that touches the limiter. The latter case aims at a plasma with X-point lying entirely in the interior of Ω_L. Further let $(r_\mathrm{desi}, z_\mathrm{desi}) \in \Gamma_\mathrm{desi}$ and $(r_1, z_1), \ldots (r_{N_\mathrm{desi}}, z_{N_\mathrm{desi}}) \in \Gamma_\mathrm{desi}$ be $N_\mathrm{desi}+1$ points on that line. We define a *quadratic* cost functional $K(\psi)$ that evaluates to zero if Γ_desi is a ψ-isoline, i.e. if ψ is constant

on Γ_{desi}:

$$K(\psi) := \frac{1}{2}\sum_{i=1}^{N_{\text{desi}}} \big(\psi(r_i, z_i) - \psi(r_{\text{desi}}, z_{\text{desi}})\big)^2. \tag{2.12}$$

Another quadratic cost functional, that will serve as *regularization*, is

$$R(I_{1,1}, \ldots, I_{L,N_L}) := \sum_{i=1}^{L}\sum_{j=1}^{N_i} \frac{w_{i,j}}{2} I_{i,j}^2. \tag{2.13}$$

The coefficients $w_{i,j} \geqslant 0$ are called *regularization weights.*

Let us state the two inverse problems that we will consider in the following.

PROBLEM 3 (INVERSE STATIC). *Let* $S_{p'} : [0, 1] \to \mathbb{R}$ *and* $S_{ff'} : [0, 1] \to \mathbb{R}$ *be two known functions. We solve the following minimization problem:*

$$\min_{\psi, I_{1,1}, \ldots I_{L,N_L}} K(\psi) + R(I_{1,1}, \ldots I_{L,N_L}) \quad \textit{subject to } (2.5) \tag{2.14}$$

with $p'(\psi) = S_{p'}(\psi_{\text{N}})$ *and* $ff'(\psi) = S_{ff'}(\psi_{\text{N}})$.

PROBLEM 4 (INVERSE STATIC, WITH GIVEN PLASMA CURRENT I_{P}). *Let* $S_{p'} : [0, 1] \to \mathbb{R}$ *and* $S_{ff'} : [0, 1] \to \mathbb{R}$ *be two known functions and assume additionally that the total plasma current* I_{P} *is given. We solve the following minimization problem:*

$$\min_{\lambda, \psi, I_{1,1}, \ldots I_{L,N_L}} K(\psi) + R(I_{1,1}, \ldots I_{L,N_L}) \quad \textit{subject to } (2.5) \textit{ and } (2.10) \tag{2.15}$$

with $p'(\psi) = \lambda S_{p'}(\psi_{\text{N}})$ *and* $ff'(\psi) = \lambda S_{ff'}(\psi_{\text{N}})$.

Clearly, it is also possible to define other cost functions forcing the plasma to have other characteristics. CEDRES++ can be easily extended in this direction. Furthermore it is possible to add both equality and inequality constraints. We are planning to include for example upper and lower bounds on the currents and the forces in the coils.

Another class of inverse problems related to static equilibrium, appears in real time tokamak control. There, it is important to reconstruct both the plasma boundary as well as the current profile functions p' and ff' in the plasma from external measurements. Frequent and fast prediction of the current state of the plasma in the tokamak machine are essential information for feedback control system. Hence, the computational challenges in solving these inverse problems are much different, and lead to the development of a separate class of equilibrium codes (Hofmann and Tonetti 1988; Lao et al. 1990; Mc Carthy et al. 1999; Blum et al. 2012).

2.3. *Direct evolution problem*

In contrast to the static problems, the evolution problems in CEDRES++ take also into account the Faraday's and Ohm's laws in the poloidal field coils and in the passive structures. The dynamics due to Faraday's and Ohm's law in the plasma, as well as the dynamics related to transport of heat and particles are supposed to be treated by external tools. Alternatively, one can prescribe the profiles $S_{p'}$ and $S_{ff'}$ as functions of time. The Poynting Theorems (2.4) and (2.3) could provide a global mean to check whether the coupling between CEDRES++ and such external tools is accurate. However, due to discretization, one needs to resort to integrated versions of the Poynting theorems for the accuracy check. Later, in Sec. 3.6, we will present detailed formulas of such integrated Poynting theorems.

The N poloidal field coils are gathered into L *poloidal field circuits* which contain in total M supplies. Each of the L poloidal field circuits contains a subset of the N coils and a subset of the M supplies. We denote by $\vec{I}_i$ the vector of size $M_i + N_i$ which contains the currents at the M_i supplies and in the N_i coils of the circuit with index i, $1 \leqslant i \leqslant L$. The circuit equations in the ith circuit can be written in the form:

$$\vec{I}_i = \boldsymbol{S}_i \vec{V}_i + \boldsymbol{R}_i \vec{\Psi}_i(\partial_t \psi)\,, \tag{2.16}$$

where the matrices $\boldsymbol{S}_i \in \mathbb{R}^{(M_i+N_i)\times M_i}$ and $\boldsymbol{R}_i \in \mathbb{R}^{(M_i+N_i)\times N_i}$ depend on the wire turns $n_{i,\cdot}$, the total resistances $R_{i,\cdot}$ and the cross sections $S_{i,\cdot}$ of the poloidal field coils in the circuit i and on the topology of the circuit. Details on the computation of matrices $\boldsymbol{S}_i$ and $\boldsymbol{R}_i$ are given in Appendix (A). The vectors $\vec{V}_i \in \mathbb{R}^{M_i}$ contain the voltages applied to the supplies, and the vectors $\vec{\Psi}_i(\psi) \in \mathbb{R}^{N_i}$ are $\vec{\Psi}_i(\psi) = (\int_{\Omega_{\mathrm{c}_{i,1}}} \psi dr dz, \ldots, \int_{\Omega_{\mathrm{c}_{i,N_i}}} \psi dr dz)^T$. In the case of a simple circuit composed of one supply connected to the coil $\Omega_{\mathrm{c}_{i,1}}$ the (2.16) writes

$$I_{i,1} = \frac{n_{i,1} V_{i,1}(t)}{R_{i,1}} - 2\pi \frac{n_{i,1}^2}{R_{i,1} S_{i,1}} \int_{\Omega_{\mathrm{c}_{i,1}}} \frac{\partial \psi}{\partial t}\, dr dz.$$

The free-boundary equilibrium problem on the time interval $[0, T]$ for the time dependent poloidal flux $\psi = \psi(t) = \psi(r, z, t)$ is:

$$\begin{aligned}
-\nabla \cdot \left(\frac{1}{\mu r} \nabla \psi\right) &= \begin{cases} r p'(\psi, t) + \frac{1}{\mu_0 r} f f'(\psi, t) & \text{in } \Omega_{\mathrm{p}}(\psi)\,; \\ S_{i,j}^{-1} (\boldsymbol{S}_i \vec{V}_i + \boldsymbol{R}_i \vec{\Psi}_i(\partial_t \psi))_j & \text{in } \Omega_{\mathrm{c}_{i,j}},\ 1 \leqslant i \leqslant L,\ 1 \leqslant j \leqslant N_i\,; \\ -\frac{\sigma_k}{r} \frac{\partial \psi}{\partial t} & \text{in } \Omega_{\mathrm{ps}_k}\,; \\ 0 & \text{elsewhere}\,, \end{cases} \\
\psi(0, z, t) &= 0\,; \qquad \lim_{\|(r,z)\| \to +\infty} \psi(r, z, t) = 0\,; \\
\psi(r, z, 0) &= \psi_0(r, z)\,,
\end{aligned} \tag{2.17}$$

The equation in the passive structures Ω_{ps_k} is deduced from Ohm's law and Faraday's law, σ_k being the equivalent axi-symmetric conductivity.

Problem 5 (Evolution, direct). *Let $S_{p'} : [0, 1] \times [0, T] \to \mathbb{R}$ and $S_{ff'} : [0, 1] \times [0, T] \to \mathbb{R}$ be two known functions. Let the evolution of the voltages $\vec{V}_1(t), \ldots, \vec{V}_L(t)$ in the poloidal field circuits and the initial data ψ_0 be given. We want to find the evolution of $\psi(t)$ such that* (2.17) *holds with $p'(\psi(t), t) = \lambda S_{p'}(\psi_{\mathrm{N}}(t), t)$ and $f f'(\psi(t), t) = \lambda S_{ff'}(\psi_{\mathrm{N}}(t), t)$.*

Problem 6 (evolution, with given plasma current $I_{\mathrm{P}}(t)$, direct). *Let $S_{p'} : [0, 1] \times [0, T] \to \mathbb{R}$ and $S_{ff'} : [0, 1] \times [0, T] \to \mathbb{R}$ be two known functions. Let the evolution of the voltages $\vec{V}_1(t), \ldots, \vec{V}_L(t)$ in the poloidal field circuits and the initial data ψ_0 be given. Additionally we assume that the evolution of the total plasma current $I_{\mathrm{P}}(t)$ is given. We want to find the evolution of $\psi(t)$ and $\lambda(t)$ such that* (2.17) *holds with $p'(\psi(t), t) = \lambda S_{p'}(\psi_{\mathrm{N}}(t), t)$ and $f f'(\psi(t), t) = \lambda S_{ff'}(\psi_{\mathrm{N}}(t), t)$ together with*

$$I_{\mathrm{P}}(t) = \lambda(t) \int_{\Omega_{\mathrm{p}}(\psi(t))} \left(r S_{p'}(\psi_{\mathrm{N}}(r, z, t), t) + \frac{1}{\mu_0 r} S_{ff'}(\psi_{\mathrm{N}}(r, z, t), t) \right) dr dz. \tag{2.18}$$

To model a consistent quasi-static evolution of plasma equilibrium the equations in (2.17) have to be augmented by the diffusion of density, temperature and magnetic flux. In that case, both functions $S_{p'}$ and $S_{ff'}$ in the Problems 5 and 6 appear as unknowns of the full system of equations.

2.4. *Inverse evolution problem*

The inverse evolution problem is the problem of determining external voltages such that the evolution of the plasma has certain prescribed properties. We will state this problem again as an *optimal control problem.*

Let $\Gamma_{\text{desi}}(t) \subset \Omega_{\text{L}}$ denote the evolution of a closed line, contained in the domain Ω_{L} that is either smooth and touches the limiter at one point or has at least one corner. The former case prescribes a desired plasma boundary that touches the limiter. The latter case aims at a plasma with X-point that is entirely in the interior of Ω_{L}. Further let $(r_{\text{desi}}(t), z_{\text{desi}}(t)) \in \Gamma_{\text{desi}}(t)$ and $(r_1(t), z_1(t)), \ldots, (r_{N_{\text{desi}}}(t), z_{N_{\text{desi}}}(t)) \in \Gamma_{\text{desi}}(t)$ be $N_{\text{desi}} + 1$ points on that line. We define a *quadratic* functional $K(\psi)$ that evaluates to zero if $\Gamma_{\text{desi}}(t)$ is an $\psi(t)$-isoline, i.e. if $\psi(t)$ is constant on $\Gamma_{\text{desi}}(t)$:

$$K(\psi(t)) := \frac{1}{2} \int_0^T \left(\sum_{i=1}^{N_{\text{desi}}} \left(\psi(r_i(t), z_i(t), t) - \psi(r_{\text{desi}}(t), z_{\text{desi}}(t), t) \right)^2 \right) dt. \tag{2.19}$$

Another functional, that will serve as *regularization*, is

$$R(\vec{V}_1(t), \ldots, \vec{V}_L(t)) := \sum_{i=1}^{L} \frac{w_i}{2} \int_0^T \vec{V}_i(t) \cdot \vec{V}_i(t) dt. \tag{2.20}$$

It penalizes the strength of the voltages $\vec{V}_i$ and represents the energetic cost in the coil system. The coefficients $w_i \geqslant 0$ are called *regularization weights.*

Problem 7 (Evolution, inverse). *Let* $S_{p'} : [0, 1] \times [0, T] \to \mathbb{R}$ *and* $S_{ff'} : [0, 1] \times \mathbb{R} \to \mathbb{R}$ *be two known functions. We solve the following minimization problem:*

$$\min_{\psi(t), \vec{V}_1(t), \ldots \vec{V}_L(t)} K(\psi(t)) + R(\vec{V}_1(t), \ldots \vec{V}_L(t)) \quad \textit{subject to } (2.17) \tag{2.21}$$

with $p'(\psi(t), t) = S_{p'}(\psi_{\text{N}}(t), t)$ *and* $ff'(\psi(t)) = S_{ff'}(\psi_{\text{N}}(t), t)$.

Problem 8 (Evolution, with given plasma current $I_{\text{P}}(t)$, inverse). *Let* $S_{p'} : [0, 1] \times [0, T] \to \mathbb{R}$ *and* $S_{ff'} : [0, 1] \times \mathbb{R} \to \mathbb{R}$ *be two known functions. Additionally we assume that the evolution of the total plasma current* $I_{\text{P}}(t)$ *is given. We solve the following minimization problem:*

$$\min_{\lambda(t), \psi(t), \vec{V}_1(t), \ldots \vec{V}_L(t)} K(\psi(t)) + R(\vec{V}_1(t), \ldots \vec{V}_L(t)) \quad \textit{subject to } (2.17) \textit{ and } (2.18) \tag{2.22}$$

with $p'(\psi(t), t) = \lambda(t) S_{p'}(\psi_{\text{N}}(t), t)$ *and* $ff'(\psi(t)) = \lambda(t) S_{ff'}(\psi_{\text{N}}(t), t)$.

3. Computational methods and applications

The main challenges for solving the Problems 1–8 numerically are their formulation on an infinite domain, the nonlinear right-hand side, the nonlinear permeability in iron and the nonlinearity due to the free plasma boundary. In the following, we will use finite element methods (Ciarlet 1978) to discretize the Problems 1–8, and we will see that this approach is flexible enough to tackle all those challenges at once. First, finite element methods are favored approximation methods due to their

flexibility on domains with complex geometry. Second, they allow for a straightforward implementation of Newton methods to handle the strong nonlinearities related to the free boundary setting. The convergence speed of such Newton methods is superior to the convergence speed of fixed-point approaches that are otherwise applied for such kind of problems. As a variational formulation is the starting point for any finite element method the section starts with the variational formulations of Problems 1, 2, 5, and 6. The subsequent paragraph on the spatial discretization, a standard finite element method with linear Lagrangian basis functions on triangles, focuses mainly on the special treatment of the free plasma boundary. It gives the important formulas, required to derive the new Newton method afterwards. Having these Newton methods at hand, it is straightforward to tackle the inverse problems. The overview on the computational methods finishes with two paragraphs describing the interfaces of CEDRES++ for the coupling with transport codes and presenting volume integrated Poynting theorems.

3.1. *Variational formulation on the truncated domain*

We chose a semi-circle Γ of radius ρ_Γ surrounding the iron domain Ω_{Fe}, the coil domains $\Omega_{\mathrm{c}_{i,j}}$ and the passive structures domain Ω_{ps_k}. The truncated domain, we use for our computations, is the domain $\Omega \subset \Omega_\infty$ having the boundary $\partial\Omega = \Gamma \cup \Gamma_{r=0}$, where $\Gamma_{r=0} := \{(r,z)\,,\ r=0\}$. The variational formulations of Problems 1, 2, 5, and 6 use the following Sobolev space:

$$V := \left\{ \psi : \Omega \to \mathbb{R}, \int_\Omega \psi^2 r\, dr dz < \infty, \int_\Omega (\nabla\psi)^2 r^{-1}\, dr dz < \infty, \psi_{|\Gamma_{r=0}} = 0 \right\} \cap C^0(\bar{\Omega}) \tag{3.1}$$

and are obtained by multiplying equations in 1, 2, 5, and 6 by test functions $\xi \in V$ and integrating by parts over Ω. They are called the variational formulations since they are the Euler equations of the minimization of the energy. Then we define

- two mappings $\mathsf{A} : V \times V \to \mathbb{R}$ and $\mathsf{J}_\mathrm{p} : V \times V \to \mathbb{R}$ that are linear in the last argument:

$$\begin{aligned} \mathsf{A}(\psi,\xi) &:= \int_\Omega \frac{1}{\mu(\psi) r} \nabla\psi \cdot \nabla\xi\, dr dz \\ \mathsf{J}_\mathrm{p}(\psi,\xi) &:= \int_{\Omega_\mathrm{p}(\psi)} \left(r p'(\psi) + \frac{1}{\mu_0 r} f f'(\psi) \right) \xi\, dr dz \end{aligned} \tag{3.2}$$

- two bilinear forms $\mathsf{j}^{\mathrm{ps}}, \mathsf{j}^{\mathrm{c}} : V \times V \to \mathbb{R}$

$$\begin{aligned} \mathsf{j}^{\mathrm{ps}}(\psi,\xi) &:= -\sum_{i=1}^{N_{\mathrm{ps}}} \int_{\Omega_{\mathrm{ps}_k}} \frac{\sigma_k}{r} \psi\xi\, dr dz \\ \mathsf{j}^{\mathrm{c}}(\psi,\xi) &:= \sum_{i=1}^{L} \sum_{j=1}^{N_i} S^{-1}_{i,j} \left(\boldsymbol{R}_i \vec{\Psi}_i(\psi) \right)_j \int_{\Omega_{\mathrm{c}_{i,j}}} \xi\, dr dz \end{aligned} \tag{3.3}$$

- N bilinear mappings $\ell_{i,j} : \mathbb{R} \times V \to \mathbb{R}$:

$$\ell_{i,j}(I,\xi) := S^{-1}_{i,j} I \int_{\Omega_{\mathrm{c}_{i,j}}} \xi\, dr dz \tag{3.4}$$

• a bilinear form $\mathsf{c} : V \times V \to \mathbb{R}$, accounting for the boundary conditions at infinity (Albanese et al. 1986):

$$\begin{aligned}\mathsf{c}(\psi,\xi) := &\frac{1}{\mu_0}\int_\Gamma \psi(\mathbf{P}_1)N(\mathbf{P}_1)\xi(\mathbf{P}_1)dS_1 \\ &+\frac{1}{2\mu_0}\int_\Gamma\int_\Gamma(\psi(\mathbf{P}_1)-\psi(\mathbf{P}_2))M(\mathbf{P}_1,\mathbf{P}_2)(\xi(\mathbf{P}_1)-\xi(\mathbf{P}_2))dS_1dS_2. \quad (3.5)\end{aligned}$$

with

$$\begin{aligned}M(\mathbf{P}_1,\mathbf{P}_2) &= \frac{k_{\mathbf{P}_1,\mathbf{P}_2}}{2\pi(r_1r_2)^{\frac{3}{2}}}\left(\frac{2-k^2_{\mathbf{P}_1,\mathbf{P}_2}}{2-2k^2_{\mathbf{P}_1,\mathbf{P}_2}}E(k_{\mathbf{P}_1,\mathbf{P}_2})-K(k_{\mathbf{P}_1,\mathbf{P}_2})\right)\\ N(\mathbf{P}_1) &= \frac{1}{r_1}\left(\frac{1}{\delta_+}+\frac{1}{\delta_-}-\frac{1}{\rho_\Gamma}\right) \quad \text{and} \quad \delta_\pm = \sqrt{r_1^2+(\rho_\Gamma \pm z_1)^2},\end{aligned}$$

where $\mathbf{P}_i = (r_i, z_i)$ and K and E the complete elliptic integrals of first and second kind, respectively and

$$k_{\mathbf{P}_j,\mathbf{P}_k} = \sqrt{\frac{4r_jr_k}{(r_j+r_k)^2+(z_j-z_k)^2}}.$$

We refer to (Grandgirard 1999, Chapter 2.4) for the details of the derivation. The bilinear form $\mathsf{c}(\cdot,\cdot)$ follows basically from the so-called *uncoupling procedure* in (Gatica and Hsiao 1995) for the usual coupling of boundary integral and finite element methods. In our case, it can be shown that for all $\mathbf{P}_1$, $\mathbf{P}_2$ the integral term $(\psi(\mathbf{P}_1)-\psi(\mathbf{P}_2))M(\mathbf{P}_1,\mathbf{P}_2)(\xi(\mathbf{P}_1)-\xi(\mathbf{P}_2))$ remains bounded. The Green's function that is used in the derivation of the boundary integral method for our problem was used earlier in finite difference methods for the Grad–Shafranov–Schlüter equations (Lackner 1976). We derive the following variational formulations of the direct Problems 1 and 2.

VARIATIONAL FORMULATION 9 (STATIC). *Let $S_{p'} : [0,1] \to \mathbb{R}$ and $S_{ff'} : [0,1] \to \mathbb{R}$ be two known functions and let the currents $I_{i,j}$ in the coils be given. We set $p'(\psi) = S_{p'}(\psi_{\mathrm{N}})$ and $ff'(\psi) = S_{ff'}(\psi_{\mathrm{N}})$ in* (3.2). *We want to find $\psi \in V$ such that*

$$\mathsf{A}(\psi,\xi) - \mathsf{J}_{\mathrm{p}}(\psi,\xi) + \mathsf{c}(\psi,\xi) = \sum_{i=1}^{L}\sum_{j=1}^{N_i}\ell_{i,j}(I_{i,j},\xi) \quad (3.6)$$

holds for all $\xi \in V$.

VARIATIONAL FORMULATION 10 (STATIC, WITH GIVEN PLASMA CURRENT I_{P}). *Let $S_{p'} : [0,1] \to \mathbb{R}$ and $S_{ff'} : [0,1] \to \mathbb{R}$ be two known functions and let the currents $I_{i,j}$ in the coils be given. We set $p'(\psi) = S_{p'}(\psi_{\mathrm{N}})$ and $ff'(\psi) = S_{ff'}(\psi_{\mathrm{N}})$ in* (3.2). *Additionally we assume that the total plasma current I_{P} is given. We want to find $\psi \in V$ and $\lambda \in \mathbb{R}$ such that*

$$\begin{aligned}\mathsf{A}(\psi,\xi) - \lambda\,\mathsf{J}_{\mathrm{p}}(\psi,\xi) + \mathsf{c}(\psi,\xi) &= \sum_{i=1}^{L}\sum_{j=1}^{N_i}\ell_{i,j}(I_{i,j},\xi),\\ I_{\mathrm{p}} - \lambda\,\mathsf{J}_{\mathrm{p}}(\psi,1) &= 0,\end{aligned} \quad (3.7)$$

holds for all $\xi \in V$.

The variational formulation of the evolution Problems 5 and 6 is based on an implicit Euler time-stepping scheme $0 := t_0 < t_0 + \Delta t_1 = t_1 < \ldots t_{n-1} + \Delta t_n = t_n = T$.

Other choices are possible. Since we will anyway employ only low order spatial discretization, the implicit Euler is the obvious choice.

VARIATIONAL FORMULATION 11 (EVOLUTION). *Let $S_{p'} : [0,1] \times [0,T] \to \mathbb{R}$ and $S_{ff'} : [0,1] \times [0,T] \to \mathbb{R}$ be two known functions. Let the evolution of the voltages $\vec{V}_1(t), \ldots, \vec{V}_L(t)$ in the poloidal field circuits and the initial data ψ_0 be given. We set $p'(\psi) = S_{p'}(\psi_N(t), t)$ and $ff'(\psi) = S_{ff'}(\psi_N(t), t)$ in (3.2). We want to find $\psi^k \in V$ approximating $\psi(t_k)$ such that*

$$\begin{aligned} &\Delta t_k \mathsf{A}(\psi^k, \xi) - \Delta t_k \mathsf{J}_p^k(\psi^k, \xi) - \mathsf{j}^{ps}(\psi^k, \xi) - \mathsf{j}^c(\psi^k, \xi) + \Delta t_k \mathsf{c}(\psi^k, \xi) \\ &\quad = \Delta t_k \sum_{i=1}^{L} \sum_{j=1}^{N_i} \ell_{i,j}((\boldsymbol{S}_i \vec{V}_i(t_k))_j, \xi) - \mathsf{j}^{ps}(\psi^{k-1}, \xi) - \mathsf{j}^c(\psi^{k-1}, \xi), \qquad (3.8) \\ &\psi^0 = \psi_0, \end{aligned}$$

holds for all $\xi \in V$ with $\mathsf{J}_p^k(\cdot,\cdot) = \mathsf{J}_p(\cdot,\cdot)_{|t=t_k}$.

VARIATIONAL FORMULATION 12 (EVOLUTION, WITH GIVEN PLASMA CURRENT $I_P(t)$). *Let $S_{p'} : [0,1] \times [0,T] \to \mathbb{R}$ and $S_{ff'} : [0,1] \times [0,T] \to \mathbb{R}$ be two known functions. Let the evolution of the voltages $\vec{V}_1(t), \ldots, \vec{V}_L(t)$ in the poloidal field circuits and the initial data ψ_0 be given. We set $p'(\psi) = S_{p'}(\psi_N(t), t)$ and $ff'(\psi) = S_{ff'}(\psi_N(t), t)$ in (3.2). Additionally we assume that the evolution of the total plasma current $I_P(t)$ is given. We want to find $\psi^k \in V$ and $\lambda^k \in \mathbb{R}$ approximating $\psi(t_k)$ and $\lambda(t_k)$ such that*

$$\begin{aligned} &\Delta t_k \mathsf{A}(\psi^k, \xi) - \Delta t_k \lambda^k \mathsf{J}_p^k(\psi^k, \xi) - \mathsf{j}^{ps}(\psi^k, \xi) - \mathsf{j}^c(\psi^k, \xi) + \Delta t_k \mathsf{c}(\psi^k, \xi) \\ &\quad = \Delta t_k \sum_{i=1}^{L} \sum_{j=1}^{N_i} \ell_{i,j}((\boldsymbol{S}_i \vec{V}_i(t_k))_j, \xi) - \mathsf{j}^{ps}(\psi^{k-1}, \xi) - \mathsf{j}^c(\psi^{k-1}, \xi), \qquad (3.9) \\ &I_p(t_k) - \lambda^k \mathsf{J}_p^k(\psi^k, 1) = 0, \quad \psi^0 = \psi_0. \end{aligned}$$

holds for all $\xi \in V$ with $\mathsf{J}_p^k(\cdot,\cdot) = \mathsf{J}_p(\cdot,\cdot)_{|t=t_k}$.

3.2. *A Galerkin discretization*

We use a standard linear Lagrangian finite element to discretize the nonlinear operators in the previous section. Finite element methods are particularly well suited to treat complex geometries, such as the one of the tokamak (plasma, passive structures, poloidal field coils.) We refer to Sec. 5 for a general discussion on the choice of the order of the finite element method. For this we introduce a triangulation Ω_h of the domain Ω that resolves the subdomains $\Omega_L, \Omega_{Fe}, \Omega_{c_{i,j}}, \Omega_{ps_k}$. The finite element approximation ψ_h of ψ in the finite element space V_h is an expansion in basis functions λ_i:

$$\psi_h(r,z) = \sum_i \psi_i \lambda_i(r,z) \text{ with } \psi_i \in \mathbb{R}. \qquad (3.10)$$

Each Lagrangian basis function $\lambda_i(r,z)$ is piecewise linear and vanishes at all vertices except one. The *domain of the plasma* $\Omega_p(\psi_h)$ of a finite element function ψ_h is bounded by a continuous, polygonal, closed line. The critical points $(r_{bnd}(\psi_h), z_{bnd}(\psi_h))$ and $(r_{ax}(\psi_h), z_{ax}(\psi_h))$ are the coordinates of certain vertices of the mesh. The saddle point of a piecewise linear function ψ_h is some vertex (r_0, z_0) with the following property: if $(r_1, z_1), (r_2, z_2) \ldots (r_n, z_n)$, denote the counterclockwise ordered neighboring vertices

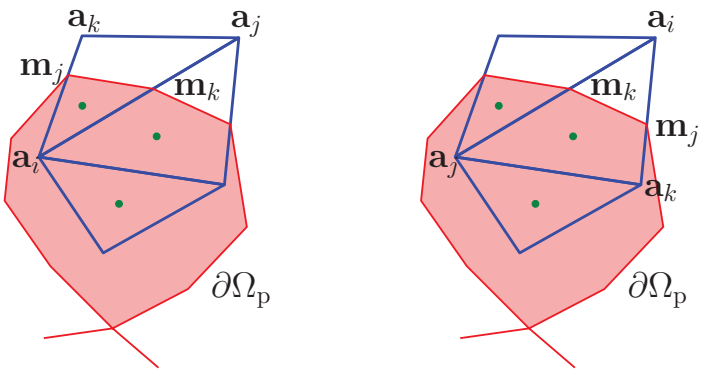

FIGURE 2. Integration over $T \cap \Omega_\mathrm{p}(\psi_h)$. The green dots indicate the location of the quadrature point. The integration domain $T \cap \Omega_\mathrm{p}(\psi_h)$ is either (a) empty, (b) the whole element T, c) a triangular domain or quadrilateral domain.

the sequence of discrete gradients $\psi_0 - \psi_1$, $\psi_0 - \psi_2 \ldots \psi_0 - \psi_n$ changes at least four times the sign.

It remains to specify the quadrature rule that is used to approximate integrals over triangles T and integrals over intersection $T \cap \Omega_\mathrm{p}(\psi_h)$ of triangles with the plasma domain

$$\int_T f(r, \psi_h)\lambda_i dr dz \quad \text{and} \quad \int_{T\cap\Omega_\mathrm{p}(\psi_h)} g(r, \psi_h)\lambda_i dr dz. \tag{3.11}$$

The second type of integrals appears in $\mathrm{J_P}$ due to the fact that the mesh does not resolve the boundary of the plasma domain Ω_p. In any case we will use the centers of gravity

$$\mathbf{b}_T := (r_T, z_T) \quad \text{and} \quad \mathbf{b}_T(\psi_h) := (r_T(\psi_h), z_T(\psi_h)) := (r_{T\cap\Omega_\mathrm{p}(\psi_h)}, z_{T\cap\Omega_\mathrm{p}(\psi_h)}) \tag{3.12}$$

of the integration domains T or $T \cap \Omega_\mathrm{p}(\psi_h)$ as quadrature points. The corresponding quadrature weights are the size of the corresponding domain $|T|$ and $|T \cap \Omega_\mathrm{p}(\psi_h)|$. The barycenter for the second type of integrals depends itself on ψ_h. Our choice of quadrature rule introduces a consistency error of order $O(h^2)$, where h is the diameter of the triangle, i.e. the quadrature is exact for linear integrands.

For a triangle T with vertex coordinates $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k \in \mathbb{R}^2$ the center of gravity corresponds to the barycenter:

$$(r_T, z_T) = \frac{1}{3}(\mathbf{a}_i + \mathbf{a}_j + \mathbf{a}_k). \tag{3.13}$$

If the domain of integration is $T \cap \Omega_\mathrm{p}(\psi_h)$, we have to distinguish the two cases, where $T \cap \Omega_\mathrm{p}(\psi_h)$ is either a triangle or a quadrilateral. Without loss of generality we assume that $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k$ is a counterclockwise ordering of the vertex coordinates of T and that $\partial\Omega_\mathrm{p}(\psi)$ intersects ∂T at two points $\mathbf{m}_k$ and $\mathbf{m}_j$ at the edges opposite to the vertices $\mathbf{a}_k$ and $\mathbf{a}_j$ (See Fig. 2). The barycentric coordinates of the intersecting points $\mathbf{m}_k$ and $\mathbf{m}_j$ are functions of ψ_h:

$$\lambda_j(\mathbf{m}_k(\psi_h)) = \frac{\psi_\mathrm{bnd}(\psi_h) - \psi_i}{\psi_j - \psi_i}, \qquad \lambda_k(\mathbf{m}_j(\psi_h)) = \frac{\psi_\mathrm{bnd}(\psi_h) - \psi_i}{\psi_k - \psi_i}, \tag{3.14}$$

and, clearly, we have $\lambda_k(\mathbf{m}_k(\psi_h)) = \lambda_j(\mathbf{m}_j(\psi_h)) = 0$.

If $T \cap \Omega_{\mathrm{p}}(\psi_h)$ is a *triangle* and $\mathbf{a}_i$ that vertex of T that is contained in $T \cap \Omega_{\mathrm{p}}(\psi_h)$ (See Fig. 2, left) we find:

$$(r_T(\psi_h), z_T(\psi_h)) = \mathbf{a}_i + \frac{1}{3}\lambda_j(\mathbf{m}_k(\psi_h))(\mathbf{a}_j - \mathbf{a}_i) + \frac{1}{3}\lambda_k(\mathbf{m}_j(\psi_h))(\mathbf{a}_k - \mathbf{a}_i) \tag{3.15}$$

and

$$|T \cap \Omega_{\mathrm{p}}(\psi_h)| = |T|\lambda_j(\mathbf{m}_k(\psi_h))\lambda_k(\mathbf{m}_j(\psi_h)). \tag{3.16}$$

If $T \cap \Omega_{\mathrm{p}}(\psi_h)$ is a *quadrilateral* and $\mathbf{a}_i$ that vertex of T that is *not* contained in $T \cap \Omega_{\mathrm{p}}(\psi_h)$ (See Fig. 2, right) we find:

$$\begin{aligned}(r_T(\psi_h), z_T(\psi_h)) = \mathbf{a}_i &+ \frac{1}{3}\frac{1-\lambda_j^2(\mathbf{m}_k(\psi_h))\lambda_k(\mathbf{m}_j(\psi_h))}{1-\lambda_j(\mathbf{m}_k(\psi_h))\lambda_k(\mathbf{m}_j(\psi_h))}(\mathbf{a}_j - \mathbf{a}_i) \\ &+ \frac{1}{3}\frac{1-\lambda_j(\mathbf{m}_k(\psi_h))\lambda_k^2(\mathbf{m}_j(\psi_h))}{1-\lambda_j(\mathbf{m}_k(\psi_h))\lambda_k(\mathbf{m}_j(\psi_h))}(\mathbf{a}_k - \mathbf{a}_i)\end{aligned} \tag{3.17}$$

and

$$|T \cap \Omega_{\mathrm{p}}(\psi_h)| = |T|\left(1 - \lambda_j(\mathbf{m}_k(\psi_h))\lambda_k(\mathbf{m}_j(\psi_h))\right). \tag{3.18}$$

In the next paragraph we will present a Newton method for the discretized nonlinear problems, and it is important to work out accurately all the nonlinear dependencies on the finite element solution ψ_h. Only then we can compute the correct derivatives.

For the sake of brevity we do not write down explicitly the discrete versions of the operators from the previous paragraph, but introduce the subcript h to denote the discretized nonlinear operators. A_h for example is the discretized version of A. We get fully discrete nonlinear formulations.

Galerkin Formulation 13 (Static). *Let $S_{p'} : [0,1] \to \mathbb{R}$ and $S_{ff'} : [0,1] \to \mathbb{R}$ be two known functions and let the currents I_i in the coils be given. We set $p'(\psi) = S_{p'}(\psi_{\mathrm{N}})$ and $ff'(\psi) = S_{ff'}(\psi_{\mathrm{N}})$ in* (3.2). *We want to find $\psi_h \in V_h$ such that*

$$\mathsf{A}_h(\psi_h, \xi_h) - \mathsf{J}_{\mathrm{p},h}(\psi_h, \xi_h) + \mathsf{c}_h(\psi_h, \xi_h) = \sum_{i=1}^{L}\sum_{j=1}^{N_i} \ell_{i,j,h}(I_{i,j}, \xi_h) \tag{3.19}$$

holds for all $\xi_h \in V_h$.

Galerkin Formulation 14 (Static, with fixed plasma current I_{P}). *Let $S_{p'} : [0,1] \to \mathbb{R}$ and $S_{ff'} : [0,1] \to \mathbb{R}$ be two known functions and let the currents I_i in the coils be given. We set $p'(\psi) = S_{p'}(\psi_{\mathrm{N}})$ and $ff'(\psi) = S_{ff'}(\psi_{\mathrm{N}})$ in* (3.2). *Additionally we assume that the total plasma current I_{P} is given. We want to find $\psi_h \in V_h$ and $\lambda \in \mathbb{R}$ such that*

$$\begin{aligned}\mathsf{A}_h(\psi_h, \xi_h) - \lambda\,\mathsf{J}_{\mathrm{p},h}(\psi_h, \xi_h) + \mathsf{c}_h(\psi_h, \xi_h) &= \sum_{i=1}^{L}\sum_{j=1}^{N_i} \ell_{i,j,h}(I_{i,j}, \xi_h), \\ I_{\mathrm{p}} - \lambda\,\mathsf{J}_{\mathrm{p},h}(\psi_h, 1) &= 0,\end{aligned} \tag{3.20}$$

holds for all $\xi_h \in V_h$.

Galerkin Formulation 15 (Evolution). *Let $S_{p'} : [0,1] \times [0,T] \to \mathbb{R}$ and $S_{ff'} : [0,1] \times [0,T] \to \mathbb{R}$ be two known functions. Let the evolution of the voltages $\vec{V}_i(t)$ in the poloidal field circuits and the initial data ψ_0 be given. We set $p'(\psi) = S_{p'}(\psi_{\mathrm{N}}(t), t)$*

and $ff'(\psi) = S_{ff'}(\psi_{\mathrm{N}}(t), t)$ *in* (3.2). *We want to find* $\psi_h^k \in V_h$ *approximating* $\psi(t_k)$ *such that*

$$\begin{aligned}
&\Delta t_k \mathsf{A}_h(\psi_h^k, \xi_h) - \Delta t_k \mathsf{J}_{\mathrm{p},h}^k(\psi_h^k, \xi_h) - \mathsf{j}_h^{\mathrm{ps}}(\psi_h^k, \xi_h) - \mathsf{j}_h^{\mathrm{c}}(\psi_h^k, \xi_h) + \Delta t_k \mathsf{c}(\psi_h^k, \xi_h) \\
&\quad = \Delta t_k \sum_{i=1}^{L} \sum_{j=1}^{N_i} \ell_{i,j,h}((\boldsymbol{S}_i \vec{V}_i(t_k))_j, \xi_h) - \mathsf{j}_h^{\mathrm{ps}}(\psi_h^{k-1}, \xi_h) - \mathsf{j}_h^{\mathrm{c}}(\psi_h^{k-1}, \xi_h), \qquad (3.21) \\
&\psi_h^0 = \psi_0
\end{aligned}$$

holds for all $\xi_h \in V_h$ *with* $\mathsf{J}_{\mathrm{p},h}^k(\cdot,\cdot) = \mathsf{J}_{\mathrm{p},h}(\cdot,\cdot)_{|t=t_k}$.

Galerkin Formulation 16 (Evolution, with given plasma current $I_{\mathrm{P}}(t)$). *Let* $S_{p'} : [0,1] \times [0,T] \to \mathbb{R}$ *and* $S_{ff'} : [0,1] \times [0,T] \to \mathbb{R}$ *be two known functions. Let the evolution of the voltages* $\vec{V}_i(t)$ *in the poloidal field circuits and the initial data* ψ_0 *be given. We set* $p'(\psi) = S_{p'}(\psi_{\mathrm{N}}(t), t)$ *and* $ff'(\psi) = S_{ff'}(\psi_{\mathrm{N}}(t), t)$ *in* (3.2). *Additionally we assume that the evolution of the total plasma current* $I_{\mathrm{P}}(t)$ *is given. We want to find* $\psi_h^k \in V_h$ *and* $\lambda^k \in \mathbb{R}$ *approximating* $\psi(t_k)$ *and* $\lambda(t_k)$ *such that*

$$\begin{aligned}
&\Delta t_k \mathsf{A}_h(\psi_h^k, \xi_h) - \Delta t_k \mathsf{J}_{\mathrm{p},h}^k(\psi_h^k, \xi_h) - \mathsf{j}_h^{\mathrm{ps}}(\psi_h^k, \xi_h) - \mathsf{j}_{,h}^{\mathrm{c}}(\psi_h^k, \xi_h) + \Delta t_k \mathsf{c}(\psi_h^k, \xi_h) \\
&\quad = \Delta t_k \sum_{i=1}^{L} \sum_{j=1}^{N_i} \ell_{i,j,h}^{\mathrm{c}}((\boldsymbol{S}_i \vec{V}_i(t_k))_j, \xi_h) - \mathsf{j}_h^{\mathrm{ps}}(\psi_h^{k-1}, \xi_h) - \mathsf{j}_{,h}^{\mathrm{c}}(\psi_h^{k-1}, \xi_h), \qquad (3.22) \\
&I_{\mathrm{p}}(t_k) - \lambda^k \mathsf{J}_{\mathrm{p},h}^k(\psi_h^k, 1) = 0, \quad \psi_h^0 = \psi_0\,,
\end{aligned}$$

holds for all $\xi \in V_h$ *with* $\mathsf{J}_{\mathrm{p},h}^k(\cdot,\cdot) = \mathsf{J}_{\mathrm{p},h}(\cdot,\cdot)_{|t=t_k}$.

The Galerkin formulations assume that the function μ_{Fe} is known. In practical applications μ_{Fe} needs to be estimated from experimental data. We refer to Glowinski and Marrocco (1974) and, more recently to Pechstein and Jüttler (2006), for details.

3.3. *Newton's method and the free plasma boundary*

Newton's methods for solving a nonlinear problem $F(x) = 0$ for x is the following iterative scheme:

$$F'(x_i)(x_{i+1} - x_i) = -F(x_i) \quad \Leftrightarrow \quad F'(x_i)x_{i+1} = F'(x_i)x_i - F(x_i). \qquad (3.23)$$

If F is sufficiently smooth, standard theory for Newton methods asserts that this iteration converges quadratically fast to the solution x. In our case the magnetic flux ψ or its finite element approximation ψ_h plays the role of the unknown x. If we want to apply this method to either our continuous nonlinear variational formulations 9, 10, 11, and 12 or the discretized versions, namely the Galerkin Formulations 13, 14, 15, and 16, we need to compute derivatives of the nonlinear operators.

For the continuous formulations we need to calculate all the directional derivatives $D_\psi \mathsf{A}(\psi,\xi)(\widetilde{\psi})$, $D_\psi \mathsf{J}_{\mathrm{p}}(\psi,\xi)(\widetilde{\psi})$, $D_\psi \mathsf{j}^{\mathrm{ps}}(\psi,\xi)(\tilde{\psi})$, $D_\psi \mathsf{j}^{\mathrm{c}}(\psi,\xi)(\tilde{\psi})$ and $D_\psi \mathsf{c}(\psi,\xi)(\tilde{\psi})$. This calculation is simple for the bilinear mappings j^{c}, j^{ps}, e.g.,

$$D_\psi \mathsf{j}^{\mathrm{ps}}(\psi,\xi)(\tilde{\psi}) = \mathsf{j}^{\mathrm{ps}}(\tilde{\psi},\xi), \quad D_\psi \mathsf{j}^{\mathrm{c}}(\psi,\xi)(\tilde{\psi}) = \mathsf{j}^{\mathrm{c}}(\tilde{\psi},\xi), \qquad (3.24)$$

and the nonlinear mapping A (see (2.6)):

$$\begin{aligned}
D_\psi \mathsf{A}(\psi,\xi)(\tilde{\psi}) = &\int_\Omega \frac{1}{\mu(\psi) r} \nabla\tilde{\psi} \cdot \nabla\xi \, dr dz \\
&- 2 \int_{\Omega_{\mathrm{Fe}}} \frac{\mu'_{\mathrm{Fe}}(|\operatorname{grad}\psi|^2 r^{-2})}{\mu^2_{\mathrm{Fe}}(|\operatorname{grad}\psi|^2 r^{-2}) r^3} \nabla\tilde{\psi} \cdot \nabla\psi \nabla\psi \cdot \nabla\xi \, dr dz.
\end{aligned}$$

The remaining derivative of J_p was given in (Blum 1989, Lemma I.4):

$$\begin{aligned} D_\psi \mathsf{J}_\mathrm{p}(\psi,\xi)(\widetilde{\psi}) = &\int_{\Omega_\mathrm{p}(\psi)} \frac{\partial j_\mathrm{p}(r,\psi_\mathrm{N}(\psi))}{\partial \psi_\mathrm{N}} \frac{\partial \psi_\mathrm{N}(\psi)}{\partial \psi} \widetilde{\psi}\,\xi\,dr dz, \\ &- \int_{\Gamma_\mathrm{p}(\psi)} j_\mathrm{p}(r,1)|\nabla\psi|^{-1}(\widetilde{\psi} - \widetilde{\psi}(r_\mathrm{bnd}(\psi), z_\mathrm{bnd}(\psi)))\xi\,d\Gamma \\ &+ \int_{\Omega_\mathrm{p}(\psi)} \frac{\partial j_\mathrm{p}(r,\psi_\mathrm{N}(\psi))}{\partial \psi_\mathrm{N}} \frac{\partial \psi_\mathrm{N}(\psi)}{\partial \psi_\mathrm{ax}} \widetilde{\psi}(r_\mathrm{ax}(\psi), z_\mathrm{ax}(\psi))\xi\,dr dz, \\ &+ \int_{\Omega_\mathrm{p}(\psi)} \frac{\partial j_\mathrm{p}(r,\psi_\mathrm{N}(\psi))}{\partial \psi_\mathrm{N}} \frac{\partial \psi_\mathrm{N}(\psi)}{\partial \psi_\mathrm{bnd}} \widetilde{\psi}(r_\mathrm{bnd}(\psi), z_\mathrm{bnd}(\psi))\xi\,dr dz, \end{aligned} \tag{3.25}$$

where Γ_p is the plasma boundary $\partial\Omega_\mathrm{p}$ and

$$j_\mathrm{p}(r,\psi_\mathrm{N}(\psi)) = r S_{p'}(\psi_\mathrm{N}(\psi)) + \frac{1}{\mu_0 r} S_{ff'}(\psi_\mathrm{N}(\psi)). \tag{3.26}$$

The derivation involves shape calculus (Murat and Simon 1976; Delfour and Zolésio 2011) and the non-trival derivatives:

$$D_\psi \psi_\mathrm{ax}(\psi)(\tilde{\psi}) = \tilde{\psi}(r_\mathrm{ax}(\psi), z_\mathrm{ax}(\psi)) \quad \text{and} \quad D_\psi \psi_\mathrm{bnd}(\psi)(\tilde{\psi}) = \tilde{\psi}(r_\mathrm{bnd}(\psi), z_\mathrm{bnd}(\psi)).$$

The formula of the derivative relies on certain smoothness assumptions on ψ. Up to our knowledge, there is no theoretical evidence that this formula holds also for plasma equilibria with boundaries that contain X-points. In particular the second term on the right-hand side seems to blow up if ψ reaches a critical point.

Also in Blum (1989), it is shown that the derivative of $\mathsf{J}_\mathrm{p}(\psi,\xi)$ in the direction ψ vanishes: $D_\psi \mathsf{J}_\mathrm{p}(\psi,\xi)(\psi) = 0$. Then the Newton scheme for solving Problem 10 is the following iteration: Let (ψ^n, λ^n) be the solution at the nth iteration. For given (ψ^n, λ^n) we introduce the linear form:

$$\begin{aligned} F^n(\xi) &:= -\mathsf{A}(\psi^n,\xi) + D_\psi \mathsf{A}(\psi^n,\xi)(\psi^n) + \sum_{i=1}^{L}\sum_{j=1}^{N_i} \ell^\mathrm{c}_{i,j}(I_{i,j},\xi) \\ &= -2\int_{\Omega_\mathrm{Fe}} \frac{\mu'_\mathrm{Fe}(|\nabla\psi|^2 r^{-2})}{\mu^2_\mathrm{Fe}(|\nabla\psi|^2 r^{-2}) r^3} |\nabla\psi^n|^2 \nabla\psi^n \cdot \nabla\xi\,dr dz + \sum_{i=1}^{L}\sum_{j=1}^{N_i} \ell^\mathrm{c}_{i,j}(I_{i,j},\xi). \end{aligned} \tag{3.27}$$

and the Newton update $(\psi^{n+1}, \lambda^{n+1})$ is the solution of the infinite dimensional linear system

$$\begin{aligned} D_\psi \mathsf{A}(\psi^n,\xi)(\psi^{n+1}) - \lambda^n D_\psi \mathsf{J}_\mathrm{p}(\psi^n,\xi)(\psi^{n+1}) + \mathsf{c}(\psi^{n+1},\xi) - \mathsf{J}_\mathrm{p}(\psi^n,\xi)\lambda^{n+1} &= F^n(\xi), \\ \lambda^n D_\psi \mathsf{J}_\mathrm{p}(\psi^n,1)(\psi^{n+1}) \qquad\qquad + \mathsf{J}_\mathrm{p}(\psi^n,1)\lambda^{n+1} &= I_\mathrm{p} \end{aligned}$$

with $\xi \in V$. After each iteration we need to recompute $\psi_\mathrm{ax}(\psi^n) = \psi^n(r_\mathrm{ax}(\psi^n), z_\mathrm{ax}(\psi^n))$, $\psi_\mathrm{bnd}(\psi^n) = \psi(r_\mathrm{bnd}(\psi^n), z_\mathrm{bnd}(\psi^n))$ and $\Omega_\mathrm{p}(\psi^n)$. For the computation of the initial flux function ψ^0 we choose a constant permeability in iron and replace the nonlinear form $\mathsf{J}_\mathrm{p}(\psi,\xi)$ with some linear form $\int_{\Omega_\mathrm{p}} j_{init}\xi\,dr dz$, where Ω_p is a given ellipse and j_{init} a given constant current density. Hence ψ^0 is the solution to a linear problem and determines the plasma axis and the plasma boundary in the first Newton iteration.

The Newton iterations for the Problems 9, 11, and 12 follow likewise. The equilibrium codes SCED (Blum et al. 1981) and Proteus (Albanese et al. 1987) are based on discretizations of such Newton iterations. The flux functions ψ^n and ψ^{n+1} are approximated by finite weighted sums of finite element basis functions and

the test functions ξ cycle over all test functions. In each Newton iteration, one has to invert an algebraic system whose size is equal to the number of finite element basis functions. But since it is not clear, whether the formula for the derivative of $\mathsf{J}_\mathrm{p}(\psi,\xi)(\widetilde{\psi})$ remains valid for plasma boundaries with X-points, these approaches are not very trustworthy.

In CEDRES++, we prefer to use Newton methods for the Galerkin Formulations 13, 14, 15, and 16. Such Newton methods need the directional derivatives $D_{\psi_h}\mathsf{A}_h(\psi_h,\xi_h)(\widetilde{\psi}_h)$, $D_{\psi_h}\mathsf{J}_{\mathrm{p},h}(\psi_h,\xi_h)(\widetilde{\psi}_h)$, $D_{\psi_h}\mathsf{j}_h^{\mathrm{ps}}(\psi_h,\xi_h)(\tilde{\psi}_h)$, $D_{\psi_h}\mathsf{j}_h^{\mathrm{c}}(\psi_h,\xi_h)(\tilde{\psi}_h)$ and $D_{\psi_h}\mathsf{c}_h(\psi_h,\xi_h)(\tilde{\psi}_h)$.

Here again, this is a straightforward and simple calculation for all mappings except one: the mapping $\mathsf{J}_{\mathrm{P},h}$ that is related to the nonlinear current profile in the plasma domain. The mapping $\mathsf{J}_{\mathrm{P},h}$ is given by

$$\mathsf{J}_{\mathrm{P},h}(\psi_h,\lambda_m)=\sum\nolimits_T \mathsf{J}_{\mathrm{P},h}^T(\psi_h,\lambda_m) := \sum\nolimits_T |T\cap\Omega_\mathrm{p}(\psi_h)|\; j_\mathrm{p}(\mathbf{b}_T(\psi_h))\,\lambda_m(\mathbf{b}_T(\psi_h)),$$

where $j_\mathrm{p}(\mathbf{b}_T(\psi_h)) = j_\mathrm{p}(r_T(\psi_h),\psi_\mathrm{N}(\psi_h(\mathbf{b}_T(\psi_h)),\psi_\mathrm{ax}(\psi_h),\psi_\mathrm{bnd}(\psi_h)))$. The directional derivative of $\mathsf{J}_{\mathrm{P},h}(\psi_h,\lambda_m)$ in direction λ_n is the partial derivative with respect to the expansion coefficient ψ_n:

$$D_\psi \mathsf{J}_{\mathrm{P},h}(\psi_h,\lambda_m)(\lambda_n)=\frac{\partial}{\partial\psi_n}\mathsf{J}_{\mathrm{P},h}(\psi_h,\lambda_m)=\frac{\partial}{\partial\psi_n}\mathsf{J}_{\mathrm{P},h}\left(\sum\nolimits_i \psi_i\lambda_i,\lambda_m\right).$$

Computing the derivative of each terms of $\mathsf{J}_{\mathrm{P},h}(\psi_h,\lambda_m)$ is a tedious application of chain and product rules. We distinguish three different cases: $T\cap\Omega_\mathrm{p}(\psi_h)=0$, $T\cap\Omega_\mathrm{p}(\psi_h)=T$ and $T\cap\Omega_\mathrm{p}(\psi_h)\subset T$ (see Fig. 2). With a slight abuse of notation we identify ψ_bnd and ψ_ax with the corresponding finite element expansion coefficient and use the Kronecker deltas $\delta_{n,\mathrm{bnd}}$ and $\delta_{n,\mathrm{ax}}$.

(*a*) $T\cap\Omega_\mathrm{p}(\psi_h)=0$:

$$\frac{\partial}{\partial\psi_n}\mathsf{J}_{\mathrm{P},h}^T(\psi_h,\lambda_m)=0\,;$$

(*b*) $T\cap\Omega_\mathrm{p}(\psi_h)=T$:

$$\begin{aligned}\frac{\partial}{\partial\psi_n}\mathsf{J}_{\mathrm{P},h}^T(\psi_h,\lambda_m)=|T|\frac{\partial j_\mathrm{p}(r_T,\psi_\mathrm{N}(\mathbf{b}_T))}{\partial\psi_\mathrm{N}}\Bigg(&\frac{\partial\psi_\mathrm{N}(\mathbf{b}_T)}{\partial\psi_h}\lambda_n(\mathbf{b}_T)\\&+\frac{\partial\psi_\mathrm{N}(\mathbf{b}_T)}{\partial\psi_\mathrm{bnd}}\delta_{n,\mathrm{bnd}}+\frac{\partial\psi_\mathrm{N}(\mathbf{b}_T)}{\partial\psi_\mathrm{ax}}\delta_{n,\mathrm{ax}}\Bigg)\lambda_m(\mathbf{b}_T);\end{aligned}$$

(*c*) $T\cap\Omega_\mathrm{p}(\psi_h)\subset T$: Without loss of generality we adopt the notation from section 3.2, introduce $\lambda_j^k=\lambda_j(\mathbf{m}_k)$ and $\lambda_k^j=\lambda_k(\mathbf{m}_j)$ use $\mathbf{b}_T$ to denote $\mathbf{b}_T(\psi_h)$. We define $AR=|T|\lambda_j^k\lambda_k^j$ if $T\cap\Omega_\mathrm{p}(\psi_h)$ is a triangle and $AR=|T|(1-\lambda_j^k\lambda_k^j)$ if $T\cap\Omega_\mathrm{p}(\psi_h)$ is a quadrilateral. We find

$$\frac{\partial}{\partial\psi_n}\mathsf{J}_{\mathrm{P},h}^T(\psi_h,\lambda_m)=A_n^T(\psi_h,\lambda_m)+C_n^T(\psi_h,\lambda_m)+T_n^T(\psi_h,\lambda_m)$$

with

- the derivative related to the area $|T \cap \Omega_{\mathrm{p}}(\psi_h)|$:

$$A_n^T(\psi_h, \lambda_m) = s|T| \left(\left(\frac{\partial \lambda_j^k}{\partial \psi_n} \lambda_k^j + \lambda_k^j \frac{\partial \lambda_j^k}{\partial \psi_n} \right) + \left(\frac{\partial \lambda_j^k}{\partial \psi_{\mathrm{bnd}}} \lambda_k^j + \lambda_k^j \frac{\partial \lambda_j^k}{\partial \psi_{\mathrm{bnd}}} \right) \delta_{n,\mathrm{bnd}} \right) j_{\mathrm{p}}(\mathbf{b}_T)\, \lambda_m(\mathbf{b}_T),$$

  where $s = 1$ if $|T \cap \Omega_{\mathrm{p}}(\psi_h)|$ is a triangle and $s = -1$ else.
- the derivative related to the current $j_{\mathrm{p}}(r_T, \psi_{\mathrm{N}}(\mathbf{b}_T))$:

$$\begin{aligned} C_n^T(\psi_h, \lambda_m) = {} & AR \frac{\partial j_{\mathrm{p}}(r_T, \psi_{\mathrm{N}}(\mathbf{b}_T))}{\partial r} \left(\frac{\partial r_T}{\partial \psi_n} \lambda_n(\mathbf{b}_T) + \frac{\partial r_T}{\partial \psi_{\mathrm{bnd}}} \delta_{n,\mathrm{bnd}} \right) \lambda_m(\mathbf{b}_T) \\ & + AR \frac{\partial j_{\mathrm{p}}(r_T, \psi_{\mathrm{N}}(\mathbf{b}_T))}{\partial \psi_{\mathrm{N}}} \left(\frac{\partial \psi_{\mathrm{N}}(\mathbf{b}_T)}{\partial \psi_{\mathrm{bnd}}} \delta_{n,\mathrm{bnd}} + \frac{\partial \psi_{\mathrm{N}}(\mathbf{b}_T)}{\partial \psi_{\mathrm{ax}}} \delta_{n,\mathrm{ax}} \right) \lambda_m(\mathbf{b}_T) \\ & + AR \frac{\partial j_{\mathrm{p}}(r_T, \psi_{\mathrm{N}}(\mathbf{b}_T))}{\partial \psi_{\mathrm{N}}} \left(\frac{\partial \psi_{\mathrm{N}}(\mathbf{b}_T)}{\partial \psi_h} \lambda_n(\mathbf{b}_T) + \frac{\partial \psi_{\mathrm{N}}(\mathbf{b}_T)}{\partial \psi_h} \nabla \psi_h(\mathbf{b}_T) \cdot \frac{\partial \mathbf{b}_T}{\partial \psi_n} \right. \\ & \left. + \frac{\partial \psi_{\mathrm{N}}(\mathbf{b}_T)}{\partial \psi_h} \nabla \psi_h(\mathbf{b}_T) \cdot \frac{\partial \mathbf{b}_T}{\partial \psi_{\mathrm{bnd}}} \delta_{n,\mathrm{bnd}} \right) \lambda_m(\mathbf{b}_T). \end{aligned}$$

- the derivative related to the test function $\lambda_m(\mathbf{b}_T)$:

$$T_n^T(\psi_h, \lambda_m) = AR\, j_{\mathrm{p}}(\mathbf{b}_T) \left(\nabla \lambda_m(\mathbf{b}_T) \cdot \frac{\partial \mathbf{b}_T}{\partial \psi_n} + \nabla \lambda_m(\mathbf{b}_T) \cdot \frac{\partial \mathbf{b}_T}{\partial \psi_{\mathrm{bnd}}} \delta_{n,\mathrm{bnd}} \right).$$

The derivatives of ψ_{N} follow easily from the Definition (2.9). We would like to stress that the Galerkin matrix $D_\psi \mathrm{J}_{\mathrm{P},h}^T(\psi_h, \lambda_m)(\lambda_n)$ can be assembled in a fairly standard, i.e. element wise, fashion, provided we compute in a preprocessing step the following information for each element T belonging to the last case: We need to know the barycententric coordinates of the intersection points $\lambda_k(\mathbf{m}_j)$ and $\lambda_j(\mathbf{m}_k)$, the barycenter $\mathbf{b}_T(\psi_h)$ and the derivatives $\frac{\partial \lambda_k(\mathbf{m}_j)}{\partial \psi_i}$, $\frac{\partial \lambda_k(\mathbf{m}_j)}{\partial \psi_j}$, $\frac{\partial \lambda_k(\mathbf{m}_j)}{\partial \psi_k}$, $\frac{\partial \lambda_k(\mathbf{m}_j)}{\partial \psi_{\mathrm{bnd}}}$, $\frac{\partial \lambda_j(\mathbf{m}_k)}{\partial \psi_i}$, $\frac{\partial \lambda_j(\mathbf{m}_k)}{\partial \psi_j}$, $\frac{\partial \lambda_j(\mathbf{m}_k)}{\partial \psi_k}$, $\frac{\partial \lambda_j(\mathbf{m}_k)}{\partial \psi_{\mathrm{bnd}}}$ and $\frac{\partial \mathbf{b}_T}{\partial \psi_i}$, $\frac{\partial \mathbf{b}_T}{\partial \psi_j}$, $\frac{\partial \mathbf{b}_T}{\partial \psi_k}$, $\frac{\partial \mathbf{b}_T}{\partial \psi_{\mathrm{bnd}}}$. All this information can be easily computed for given ψ_h, ψ_{bnd} and ψ_{ax} using the Formulas (3.14), (3.15), and (3.17). All the terms that contain the Kronecker deltas $\delta_{n,\mathrm{bnd}}$ or $\delta_{n,\mathrm{ax}}$ lead to non-local entries in the stiffness matrix. They connect the coefficients $\psi_{i_1} = \psi_{\mathrm{bnd}}$ and $\psi_{i_2} = \psi_{\mathrm{ax}}$ with all coefficients ψ_j that are associated to vertices of elements that are intersected by the plasma domain $\Omega_{\mathrm{P}}(\psi_h)$.

The size of the algebraic systems that we need to solve in each iteration corresponds to the number of vertices of the triangulation. Even for very fine discretizations it is today possible to use *direct linear solvers* such as UMFPACK (Davis 2011). As long as the storage amount for the algebraic system does not exceed the memory, modern direct solvers will outperform in most cases an iterative solver.

3.4. *Sequential quadratic programming for the inverse problems*

In CEDRES++, we use the following fully discrete reformulation of the inverse Problems 3 and 4, to find optimal currents in the poloidal field coils.

Inverse Problem 17 (Static). *Let* $S_{p'} : [0,1] \to \mathbb{R}$ *and* $S_{ff'} : [0,1] \to \mathbb{R}$ *be two known functions. We set* $p'(\psi) = S_{p'}(\psi_{\mathrm{N}})$ *and* $ff'(\psi) = S_{ff'}(\psi_{\mathrm{N}})$ *in* (3.2). *We solve the*

following minimization problem:

$$\min_{\psi_h, I_{1,1}, \ldots I_{L,N_L}} K(\psi_h) + R(I_{1,1}, \ldots I_{L,N_L}) \quad \textit{subject to } (3.19). \tag{3.28}$$

INVERSE PROBLEM 18 (STATIC, WITH GIVEN PLASMA CURRENT I_P). *Let $S_{p'} : [0, 1] \to \mathbb{R}$ and $S_{ff'} : [0, 1] \to \mathbb{R}$ be two known functions and assume additionally that the total plasma current I_P is given. We set $p'(\psi) = S_{p'}(\psi_N)$ and $ff'(\psi) = S_{ff'}(\psi_N)$ in (3.2). We solve the following minimization problem:*

$$\min_{\lambda, \psi_h, I_{1,1}, \ldots I_{L,N_L}} K(\psi_h) + R(I_{1,1}, \ldots I_{L,N_L}) \quad \textit{subject to } (3.20). \tag{3.29}$$

The inverse Problems 17 and 18 are *finite dimensional constrained optimization problems.* The *sequential quadratic programming (SQP)* method is the fastest method for finite dimensional constrained optimization problems. We refer to the text book (Nocedal and Wright 2006, Chapter 18) for the details and explain here only the basic idea.

Both inverse Problems 17 and 18 are optimization problems of the following type

$$\min_{\mathbf{u},\mathbf{y}} \frac{1}{2}\mathbf{y}^T \boldsymbol{K}\mathbf{y} + \frac{1}{2}\mathbf{u}^T \boldsymbol{H}\mathbf{u} \quad \text{s.t} \qquad \boldsymbol{B}(\mathbf{y}) = \boldsymbol{F}(\mathbf{u}), \tag{3.30}$$

where the quadratic matrices $\boldsymbol{H}$ and $\boldsymbol{K}$ are the discretization of the cost functions K and R, the state variable $\mathbf{y}$ is the vector of the finite element coefficients ψ_i and the scaling factor λ, the control variable $\mathbf{u}$ is the vector of the N currents I_i in the poloidal field coils and $\boldsymbol{B}$ and $\boldsymbol{F}$ the Galerkin discretizations of (3.19) or (3.20). The Lagrange function formalism in combination with Newton-type iterations is one approach to derive the SQP-methods: the Lagrangian for (3.30) is

$$L(\mathbf{y}, \mathbf{u}, \mathbf{p}) = \frac{1}{2}\mathbf{y}^T \boldsymbol{K}\mathbf{y} + \frac{1}{2}\mathbf{u}^T \boldsymbol{H}\mathbf{u} + \mathbf{p}^T(\boldsymbol{B}(\mathbf{y}) - \boldsymbol{F}(\mathbf{u})) \tag{3.31}$$

and the solution of (3.30) is a stationary point of this Lagrangian:

$$\begin{aligned} \boldsymbol{K}\mathbf{y} + D_{\mathbf{y}}\boldsymbol{B}^T(\mathbf{y})\mathbf{p} &= 0, \\ \boldsymbol{H}\mathbf{u} - D_{\mathbf{u}}\boldsymbol{F}^T(\mathbf{u})\mathbf{p} &= 0, \\ \boldsymbol{B}(\mathbf{y}) - \boldsymbol{F}(\mathbf{u}) &= 0. \end{aligned} \tag{3.32}$$

The superscript T indicates the adjoint operator, which corresponds to matrix transposition in the finite dimensional case. The second line in (3.32) corresponds to the optimality condition for the gradient of the reduced cost functional $\frac{1}{2}\mathbf{u}^T \boldsymbol{H}\mathbf{u} + \frac{1}{2}\mathbf{y}^T(\mathbf{u})\boldsymbol{K}\mathbf{y}(\mathbf{u})$, where $\mathbf{y}(\mathbf{u})$ is implicitly defined by $\boldsymbol{B}(\mathbf{y}(\mathbf{u})) = \boldsymbol{F}(\mathbf{u})$. This is the main reason for which gradient type methods for a corresponding unconstrained optimization problem for the reduced cost function are too expensive: one evaluation of the gradient requires the very expensive solution of the nonlinear problem in the third line of (3.32). For the SQP-methods on the other hand, the overall computing time in practical examples has about the same magnitude as the computing time for solving the constraint for given control parameters.

A quasi-Newton method for solving (3.32) are iterations of the type

$$\begin{pmatrix} \boldsymbol{K} & 0 & D_{\mathbf{y}}\boldsymbol{B}^T(\mathbf{y}^i) \\ 0 & \boldsymbol{H} & -D_{\mathbf{u}}\boldsymbol{F}^T(\mathbf{u}^i) \\ D_{\mathbf{y}}\boldsymbol{B}(\mathbf{y}^i) & -D_{\mathbf{u}}\boldsymbol{F}(\mathbf{u}^i) & 0 \end{pmatrix} \begin{pmatrix} \mathbf{y}^{i+1} - \mathbf{y}^i \\ \mathbf{u}^{i+1} - \mathbf{u}^i \\ \mathbf{p}^{i+1} - \mathbf{p}^i \end{pmatrix} = - \begin{pmatrix} \boldsymbol{K}\mathbf{y}^i + D_{\mathbf{y}}\boldsymbol{B}^T(\mathbf{y}^i)\mathbf{p}^i \\ \boldsymbol{H}\mathbf{u}^i - D_{\mathbf{u}}\boldsymbol{F}^T(\mathbf{u}^i)\mathbf{p}^i \\ \boldsymbol{B}(\mathbf{y}^i) - \boldsymbol{F}(\mathbf{u}^i) \end{pmatrix} \tag{3.33}$$

We call the iterative scheme (3.33) a quasi-Newton method since we omit the second order derivatives of $\boldsymbol{B}$ and $\boldsymbol{F}$. The quadratic convergence of Newton's method deteriorates to super-linear convergence. The number of control parameters is much smaller than the number of state coefficients. Therefore the algebraic system in (3.33) is roughly twice as large as the algebraic system of a Newton iteration of the direct problem. Hence, as in the direct case, there is today no need to use iterative linear solver.

This will be different for the inverse problems of evolving free-boundary equilibria. There the size of the algebraic system increases by a factor that corresponds to the number of time steps. We refer for Blum and Heumann (2014) for details and state here only the finite dimensional inverse problems that are addressed in CEDRES++. They are based on a discrete cost function $K_h(\{\psi_h^k\}_{k=1}^n)$:

$$K_h(\{\psi_h^k\}_{k=1}^n) = \sum_{k=1}^{n} \left(\frac{\Delta t_k}{2} \sum_{i=1}^{N_{\text{desi}}} \left(\psi_h^k(r_i, z_i) - \psi_h^k(r_{\text{desi}}(t_k), z_{\text{desi}}(t_k)) \right)^2 \right) \tag{3.34}$$

for the finite element approximation $\{\psi_h^k\}_{k=1}^n$ at t_k and a discrete regularization function:

$$R_h(\{\vec{V}_1(t_k)\}_{k=1}^n, \ldots, \{\vec{V}_L(t_k)\}_{k=1}^n) = \sum_{i=1}^{L} \frac{w_i}{2} \sum_{k=1}^{n} \Delta t_k \vec{V}_i(t_k) \cdot \vec{V}_i(t_k). \tag{3.35}$$

for the coil voltages $\{\vec{V}_i(t_k)\}_{k=1}^n$.

Inverse Problem 19 (Evolution). *Let $S_{p'} : [0,1] \times \mathbb{R} \to \mathbb{R}$ and $S_{ff'} : [0,1] \times \mathbb{R} \to \mathbb{R}$ be two known functions. We set $p'(\psi) = S_{p'}(\psi_N(t), t)$ and $ff'(\psi) = S_{ff'}(\psi_N(t), t)$ in* (3.2). *We solve the following minimization problem:*

$$\min_{\{\psi_h^k, \vec{V}_i(t_k)\}_{k=1}^n} K_h(\{\psi_h^k\}_{k=1}^n) + R_h(\{\vec{V}_1(t_k)\}_{k=1}^n, \ldots, \{\vec{V}_L(t_k)\}_{k=1}^n) \quad \textit{subject to } (3.21).$$

Inverse Problem 20 (Evolution, with given plasma current I_P). *Let $S_{p'} : [0,1] \times \mathbb{R} \to \mathbb{R}$ and $S_{ff'} : [0,1] \times \mathbb{R} \to \mathbb{R}$ be two known functions and assume additionally that the total plasma current I_P is given. We set $p'(\psi) = S_{p'}(\psi_N(t), t)$ and $ff'(\psi) = S_{ff'}(\psi_N(t), t)$ in* (3.2). *We solve the following minimization problem:*

$$\min_{\{\lambda^k, \psi_h^k, \vec{V}_i(t_k)\}_{k=1}^n} K_h(\{\psi_h^k\}_{k=1}^n) + R_h(\{\vec{V}_1(t_k)\}_{k=1}^n, \ldots, \{\vec{V}_L(t_k)\}_{k=1}^n) \quad \textit{subject to } (3.22).$$

We would like to highlight that the SQP-method relies on proper derivatives of the nonlinear operators $\boldsymbol{B}$ and $\boldsymbol{F}$. In our case $\boldsymbol{F}$ is affine, hence the derivative of $\boldsymbol{B}$ remains the most difficult part. On the other hand these derivatives are exactly the same derivatives that we used for the new Newton methods. Hence the implementation of a SQP-method for the inverse problem uses the same main building blocks.

3.5. *Flux surface averages and geometric coefficients*

As for any equilibrium code numerous outputs can be extracted from the poloidal flux map computed. These include purely geometric information on the plasma shape (plasma boundary, geometric axis, elongation ...), global parameters (such as total plasma current I_P, poloidal beta β_p, internal inductance li, ...), 1D profiles of quantities constant on flux isolines in the plasma and 2D maps (ψ itself but also B_r,

B_z, j_p, ...). All these outputs are standardized and follow the conventions of the *European Integrated Tokamak Modelling Project* (Falchetto et al. 2014; ITM 2013). We are not going to detail all of them in this paper. Let us however give some details on the computation of some of the important 1D profiles in the plasma. For $\psi_N \in [0, 1]$, $S_f(\psi_N) = f(\psi)$ is computed by integration of $S_{ff'}$

$$S_f(\psi_N) = [(r_0 B_0)^2 - 2(\psi_{bnd} - \psi_{ax}) \int_{\psi_N}^{1} S_{ff'}(x)dx]^{1/2}, \tag{3.36}$$

where B_0 is the vacuum toroidal field at $r = r_0$. Let us define a discretization of the unit interval [0, 1] by $S + 1$ values $\psi_N^0 = 0, \ldots, \psi_N^S = 1$. These points are taken as abscissa for all computed 1D profiles. For each ψ_N^s the contour line $\Gamma_{\psi_N^s}$ is extracted from the finite element representation of the solution as a list of N_s segments between $\mathbf{m}_{s,1}^l = (r_{s,1}^l, z_{s,1}^l)$ and $\mathbf{m}_{s,2}^l = (r_{s,2}^l, z_{s,2}^l)$ with length $|L_s^l|$, for $l = 1$ to N_s.

The toroidal flux coordinate is defined as $\rho(\psi_N) = \sqrt{\phi(\psi_N)/\pi B_0}$ where $\phi(\psi_N) = \int_{\Omega_{\psi_N}} \frac{f(\psi(r,z))}{r} dr dz$ and Ω_{ψ_N} is the domain bounded by the line of flux Γ_{ψ_N}. The quantities ϕ_s and ρ_s are computed from the discrete ψ_h for all ψ_N^s using a barycentric quadrature rule (cf. Sec. 3.2):

$$\phi_s = \sum_T \frac{S_f(\psi_N(\mathbf{b}_T(\psi_h)))}{r_T(\psi_h)} |T \cap \Omega_{\psi_N^s}|. \tag{3.37}$$

The profiles ψ_s and ρ_s being known one can compute $(\frac{\partial \psi}{\partial \rho})_s = \psi_s'$ using finite differences.

In the same way the volume profile is computed as

$$Vol_s = 2\pi \sum_T r_T(\psi_h) |T \cap \Omega_{\psi_N^s}| \tag{3.38}$$

and $(\frac{\partial Vol}{\partial \rho})_s = Vol_s'$ using finite differences.

Following (Blum 1989) the average of a quantity A over magnetic surfaces can be computed as

$$\langle A \rangle_s = \left(\int_{\Gamma_{\psi_N^s}} \frac{A\, r}{|\nabla \psi_h|} dl \right) \Big/ \left(\int_{\Gamma_{\psi_N^s}} \frac{r}{|\nabla \psi_h|} dl \right). \tag{3.39}$$

A number of 1D profiles, also called geometric coefficients, are computed as such averages, e.g. $\langle 1/r^2 \rangle$ or $\langle |\nabla \rho|^2 / r^2 \rangle$. The integrals over flux contour lines involved are approximated as follows:

$$\int_{\Gamma_{\psi_N^s}} \frac{A\, r}{|\nabla \psi_h|} dl \approx \sum_{l=1}^{N_s} \frac{1}{2} \left(\frac{r_{s,1}^l A(\mathbf{m}_{s,1}^l)}{|\nabla \psi_{h|T_s^l}|} + \frac{r_{s,2}^l A(\mathbf{m}_{s,2}^l)}{|\nabla \psi_{h|T_s^l}|} \right) |L_s^l|, \tag{3.40}$$

where T_s^l is the triangle which is intersected by the segment between $\mathbf{m}_{s,1}^l$ and $\mathbf{m}_{s,2}^l$ and $\mathbf{m}_{s,\cdot}^l = (r_{s,\cdot}^l, z_{s,\cdot}^l)$. $|\nabla \psi_{h|T_s^l}|$ is constant in the triangle and computed from the 3 values at the nodes of T_s^l.

3.6. *Volume integrated Poynting theorems*

The subset of (2.1) and (2.2) we used in Sec. 2 to derive the evolution Problems 5 and 6 that are solved in CEDRES++, involve the poloidal Faraday and the toroidal Ampère law. Hence, the poloidal Poynting Theorem (2.4) can be used to check the accuracy of the solution independently of an additional treatment of the transport equations.

We integrate the poloidal Poynting Theorem (2.4) over a volume Vol that contains all the plasma of a given scenario and get, by toroidal symmetry:

$$\int_{\partial S} \frac{\nabla\psi \cdot \mathbf{n}}{\mu_0 r}\, \partial_t \psi ds = -\int_{S\cap\Omega_{\mathrm{P}}(\psi)} \left(rp'(\psi) + \frac{1}{\mu_0 r} ff'(\psi) \right) \partial_t \psi dr dz + \int_S \frac{\nabla\psi \cdot \nabla \partial_t \psi}{\mu_0 r} dr dz, \tag{3.41}$$

where S denotes the intersection of Vol with the poloidal plane.

If we choose S to be the domain of the plasma $\Omega_{\mathrm{P}}(\psi)$ then the left-hand side of (3.41) is $V_{\mathrm{loop}} I_{\mathrm{P}}$, where V_{loop} is the plasma loop voltage. The first integral of the right-hand side is related to the time rate of variation of the toroidal magnetic energy and to the work done against the plasma pressure gradient. The last term of the right-hand side is the time rate of variation of the poloidal magnetic field energy.

We would like to stress that the integrated poloidal Poynting theorem corresponds to a variational formulation of the Grad–Shafranov–Schlüter equations on S. Using the notation (3.2) of our variational formulation from Sec. 3, we remark that the two integrals on the right-hand side correspond to $\mathsf{J}_{\mathrm{p}}(\psi, \chi_S \partial_t \psi)$ and $\mathsf{A}(\psi, \chi_S \partial_t \psi)$, where χ_S is the characteristic function of S. Hence, it can be shown that the solutions ψ^k of the evolution Problems 15 and 16 fulfill the volume integrated Poynting theorem up to first order accuracy in the mesh size.

The volume integrated version of the toroidal Poynting Theorem (2.3) together with the static inverse mode was used in Ané et al. (2000) for the optimization of ITER scenarios.

4. Tests and examples

4.1. *Validation and performance*

From the best of our knowledge, there does not exist analytical solutions for the free boundary equilibrium problem considered in this paper. To provide nevertheless some evidence for convergence of the method, we follow a common approach in engineering and study the convergence towards a numerical solution that is computed on a very fine mesh.

We consider a static equilibrium with a given plasma current (Problem 2) in ITER geometry. The plasma current is $I_p = 15.10 \times 10^6 A$ and the current density profile is prescribed using the model (2.11) with $r_0 = 6.2$ m $\alpha = 0.5978$, $\beta = 0.5978$, $\gamma = 1.395$. With these data, we solve the Galerkin Formulation (16) on a sequence of five meshes with increasing number of elements. The solution obtained on the mesh with the largest number of triangular elements, is used as a reference solution and is noted ψ_{ref}. In our case, the reference solution ψ_{ref} has 577 415 number of unknowns and the mesh consists of 1153 174 triangular elements. The reference solution is depicted in Fig. 3. For each of the other four meshes, the numerical solution $\psi_{N_{ukwn}}$ is evaluated at $N_{points} = 812$ different points of the computational domain which are located independently of the mesh (see Fig. 4). Then the relative error

$$E_{points}(N_{ukwn}) = \frac{(\sum_{i=1}^{N_{points}} |\psi_{N_{ukwn}}(M_i) - \psi_{ref}(M_i)|^2)^{1/2}}{(\sum_{i=1}^{N_{points}} |\psi_{ref}(M_i)|^2)^{1/2}} \tag{4.1}$$

is used to quantify the convergence. The values in Fig. 4 demonstrate the expected linear convergence. Similarly, in Fig. 5, we monitor convergence of the plasma quantities total plasma volume Vol_S, the numerical derivative Vol'_s (see (3.38)) and the geometric coefficient $G_s := \langle |\nabla\rho|^2 r^{-2} \rangle_s$ (see (3.39)) that are computed in a post-processing step from the numerical solution ψ_h with the methods from Sec. 3.5. Here

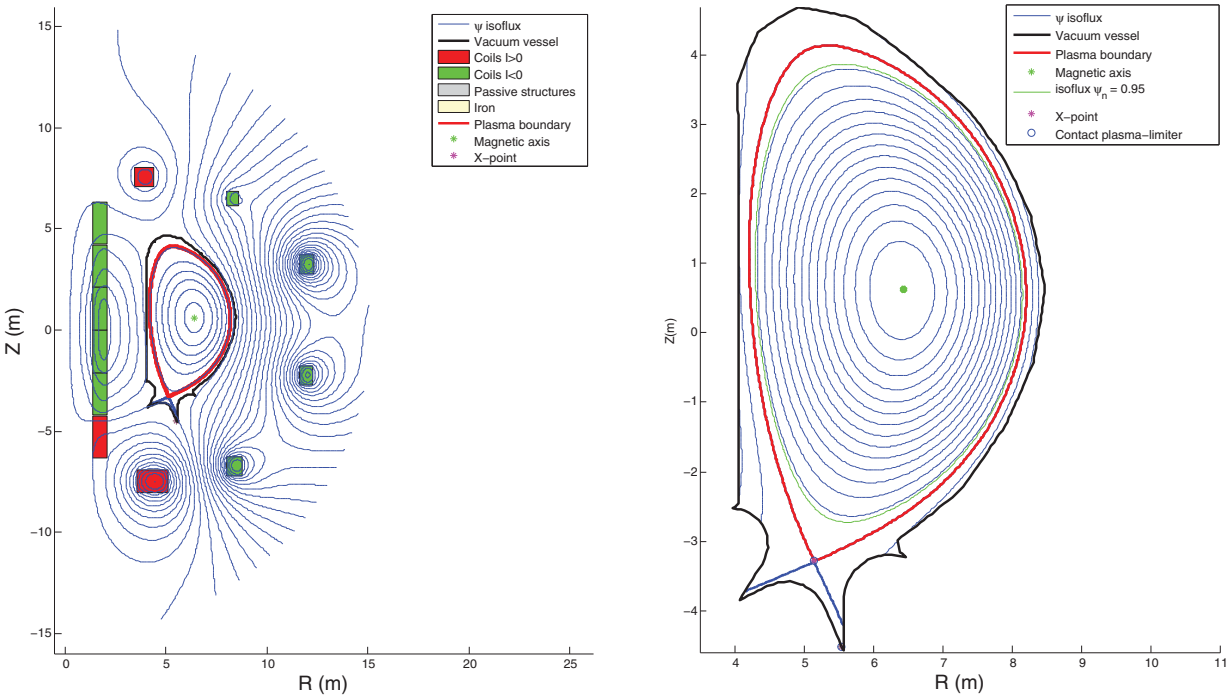

FIGURE 3. The reference solution for the ITER test problem.

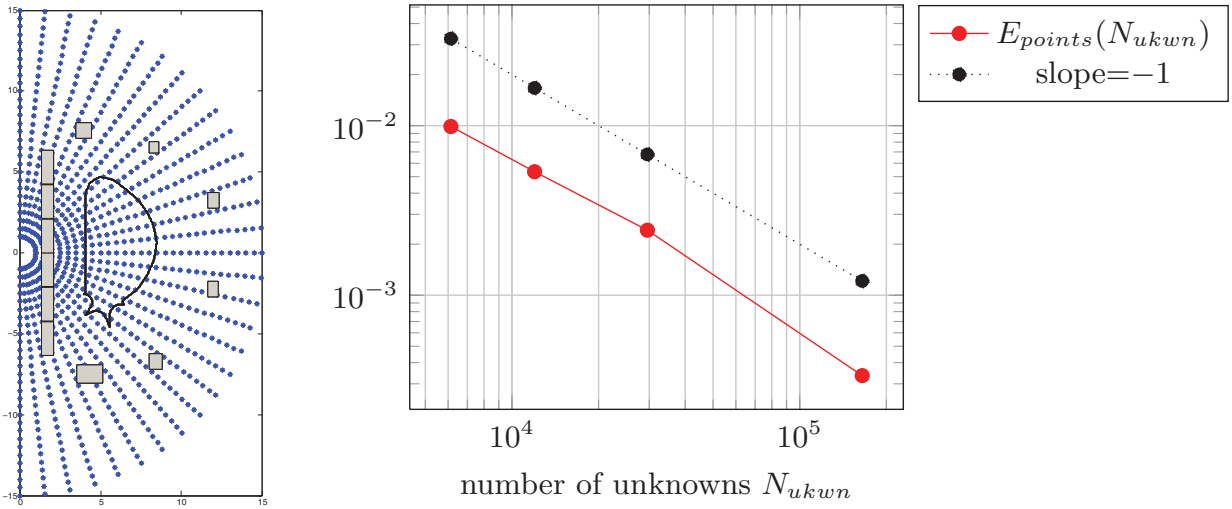

FIGURE 4. Left: the location of the points in the definition of E_{points}. Right: convergence of E_{point}.

again (see Fig. 5) we observe approximately linear convergence in the number of unknowns.

In Table 2 we give the overall computing time for the previous five computations on an Intel Sandy Bridge 2.6 GHz. The computing time scales linearly with the number of unknowns. Given the fact that we solve nonlinear problems, the computation time is reasonably small. Application engineers can easily solve a huge amount of different scenarios in short time to do parameter studies for example. CEDRES++ is perfectly suitable in larger work-flow environments, such as the *European Integrated Tokamak Modelling Project* (ITM 2013; Falchetto et al. 2014). One reason for such short running times is the Newton method. The convergence history of the residuum in Table 2 shows perfect quadratic convergence: we need only very few iterations to find a numerical solution solving the discrete nonlinear problem within the limits of machine precision.

Number of triangles	Number of unknowns	Computing time (in s)
12 099	6134	2
23 723	11 985	5
58 744	29 556	11
328 693	164 887	88
1 153 174	577 415	368

Iteration	Relative residual
1	$2.667\,473 \times 10^{+00}$
2	$9.157\,459 \times 10^{-02}$
3	$1.781\,645 \times 10^{-03}$
4	$0.525\,234 \times 10^{-06}$
5	$3.935\,226 \times 10^{-12}$

TABLE 2. Left: calculation time. Right: convergence history of the Newton iteration in the calculation of ψ_{ref}.

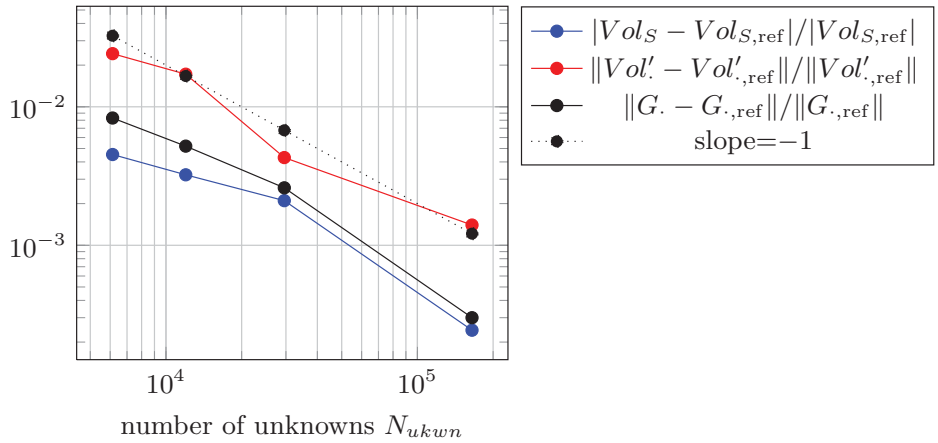

FIGURE 5. Convergence of plasma volume Vol_S, the numerical derivative $Vol'_s = (\frac{\partial Vol}{\partial \rho})_s, s = 0, \ldots S$ and the geometric coefficient $G_s = \langle |\nabla \rho|^2 r^{-2} \rangle_s, s = 0, \ldots S$. The reference quantities $Vol_{S,\mathrm{ref}}$, $Vol'_{s,\mathrm{ref}}$ and $G_{s,\mathrm{ref}}$ correspond to the quantities computed for ψ_{ref}.

4.2. *Quasi-static plasma equilibrium simulations for WEST*

The tungsten (W) environment in steady-state tokamak (WEST) project (Bucalossi et al. 2011) aims at equipping Tore Supra with an actively cooled tungsten divertor. This represents a major change in the magnetic configuration of Tore Supra, moving from a circular limited configuration to a diverted (or X-point) configuration. CEDRES++ is one of the main modeling tools used for the preparation of WEST. It has been employed in particular for the definition of reference equilibria, the dimensioning of the plasma vertical position feedback system, the design of the plasma shape controller, breakdown studies, disruption simulations, etc. We give below a few examples of CEDRES++ simulations for WEST. Note that Tore Supra is an iron core tokamak and that the iron is taken into account in all of these simulations. The six return arms of the iron core are represented in CEDRES++ by an axisymmetric equivalent model, which gives the $1/R$ shape of the return arms visible in Fig. 6. We are using the experimental data for the poloidal magnetic field $\mathbf{B}_{pol}$ and the poloidal magnetizing field $\mathbf{H}_{pol}$ from Table 3, do piecewise linear interpolation of these data and reconstruct the permeability for arbitrary magnetic field values via $\mu_{\mathrm{FE}}(\mathbf{B}^2_{pol}) = |\mathbf{B}_{pol}||\mathbf{H}_{pol}|^{-1}$.

We present in the following sections three different examples from research for WEST that use the static direct, static evolution, and inverse static modes of CEDRES++. First simulations with the inverse evolution mode are presented in Blum and Heumann (2014). The inverse static mode of CEDRES++ is also extremely

	$\lvert\mathbf{B}_{pol}\rvert$	$\lvert\mathbf{H}_{pol}\rvert$		$\lvert\mathbf{B}_{pol}\rvert$	$\lvert\mathbf{H}_{pol}\rvert$
1	0.00	0	9	1.76	7.968×10^3
2	0.50	3.833×10^2	10	2.06	4.821×10^4
3	0.70	3.982×10^2	11	2.25	1.628×10^5
4	0.80	4.102×10^2	12	3.05	8.090×10^5
5	0.88	4.270×10^2	13	4.05	1.588×10^6
6	1.00	4.703×10^2	14	6.05	3.178×10^6
7	1.20	6.274×10^2	15	98.20	7.651×10^7
8	1.52	2.474×10^3	16	10^5	7.957×10^{10}

TABLE 3. The data for the poloidal magnetic field $|\mathbf{B}_{pol}|$ and the poloidal magnetizing field $|\mathbf{H}_{pol}|$ that is used to reconstruct the magnetic permeability.

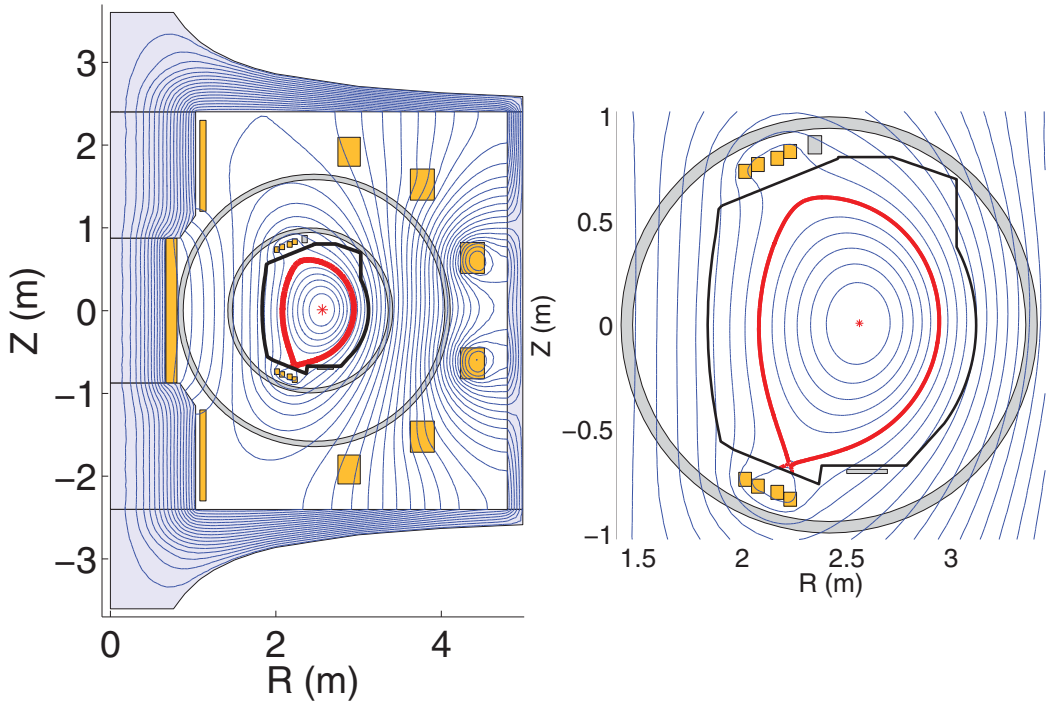

FIGURE 6. Poloidal cross section showing ψ-isolines for a WEST typical equilibrium. Left: global view; right: zoom on the inner vacuum vessel region. The iron is displayed in blue, the poloidal field coils in orange and the passive structures (vacuum vessel and vertical stabilization components) in gray. The black curve is the limiter curve, i.e. the domain accessible to the plasma. The red curve is the plasma boundary and the red star the magnetic axis.

useful in order to define and optimize reference equilibria. We will give details for WEST in a forthcoming publication.

4.2.1. *The current-focused case: direct static mode.*

Figure 6 shows a typical WEST poloidal flux map calculated by CEDRES++ in current-focused mode. The X-point is visible at the bottom of the plasma. Here, CEDRES++ solves the direct static Problem 2 with prescribed total plasma current $I_P = 700$ kA and with the parametrized current profiles $S_{p'}$ and $S_{ff'}$ in (2.11), using $\alpha = 1$, $\beta = 1.5$, $\gamma = 0.9$, $r_0 = 2.6$ m. The vacuum toroidal field is $B_0 = 3.524$ T at $r = 2.6$ m. A few output parameters are: $\beta_p = 1.70$, $li = 0.93$, $q_{95} = 3.33$, $q_0 = 1.17$.

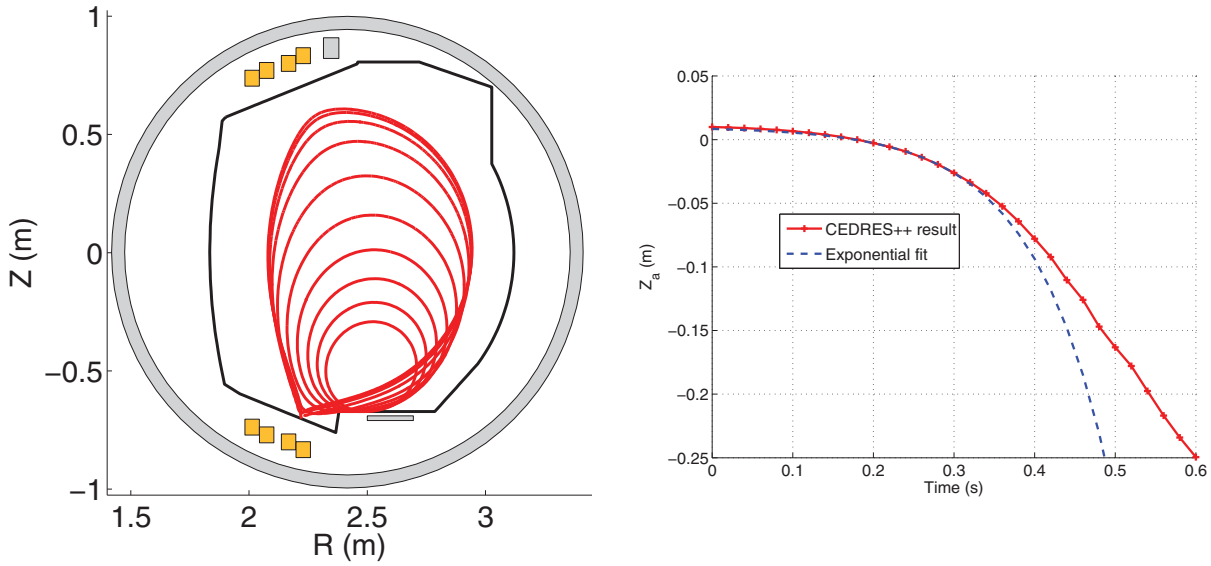

FIGURE 7. Left: plasma boundary at intervals of 100 ms in a vertical instability simulation for WEST. Right: time evolution of the vertical position of the magnetic axis z_{ax} in a vertical instability simulation for WEST.

4.2.2. *The voltage-evolution-focused case, direct evolution mode.*

Starting from the equilibrium in the previous section, we run CEDRES++ in direct evolution mode to solve Problem 6. We keep all the input parameters fixed and we apply a constant voltage to the coils, equal to the resistive voltage being the product of coil resistance and current, except for the divertor coils, where we perturb the resistive voltage with $\Delta V = +0.1$ V in the lower coil and $\Delta V = -0.1$ V in the upper coil. This is in order to trigger a vertical instability (otherwise the plasma would stay in place). The simulation is run with a time step of 20 ms. Figure 7 shows the plasma boundary at intervals of 100 ms. The plasma moves down and the diverted configuration is lost after a few 100 ms when the plasma comes in contact with the baffle of the pumping system. Figure 7 shows the vertical position of the magnetic axis z_{ax} as a function of time. The early evolution is exponential (as one expects) with a time constant $\tau_z = 95$ ms, while the later evolution is rather linear. The error of the integrated Poynting theorem from Sec. 3.6 is approximately 5% and decreases if the time step size and the triangle size are refined.

4.2.3. *The current-focused case, inverse static mode.*

We present here an example that requires to solve an inverse free-boundary equilibrium problem: the post-processing of CRONOS (Artaud et al. 2010) simulations of WEST scenarios. Indeed, CRONOS simulations are typically run in fixed-boundary equilibrium mode (using the HELENA equilibrium code (Huysmans et al. 1991) with a prescribed boundary geometry). One may however need to know the magnetic field outside the plasma boundary (for example in order to prepare JOREK (Huysmans and Czarny 2007) nonlinear MHD simulations), or to assess whether the scenario is feasible in terms of current limits in the coils for example. These questions may be addressed by solving the inverse Problems 17 and 18, using as desired boundary, the boundary used in the CRONOS simulation, and as current profile those profiles $S_{p'}$ and $S_{ff'}$ that are calculated by CRONOS. These profiles are shown in Fig. 8, where one can notice peaks at the edge of the plasma which are

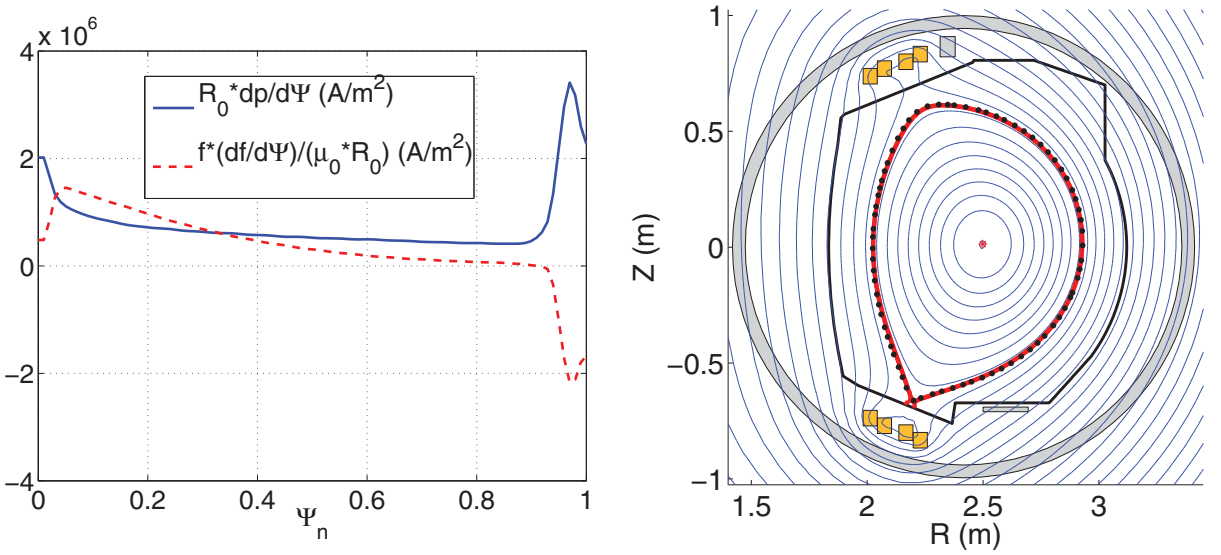

FIGURE 8. Left: The profile functions $S_{p'}$ and $S_{ff'}$ (normalized so as to fit in the same figure) from a CRONOS WEST scenario simulation. Right: Lines: ψ-isolines calculated by CEDRES++ in inverse snapshot mode. Black dots: CRONOS plasma boundary, used as the desired plasma boundary in CEDRES++.

characteristic for H-mode profiles, while Fig. 8 shows the result of the CEDRES++ calculation. It can be seen that the CRONOS boundary (black dots) is well matched.

5. Conclusions and perspectives

We have presented in detail the computational methods of CEDRES++. It enables to compute quasi-static equilibrium configurations, the currents in the poloidal field coils or the voltages applied in the circuits of the poloidal field system being prescribed. In its inverse mode the code computes these currents and voltages that ensure a certain prescribed plasma shape that might evolve in time.

Due to its stability and robustness, CEDRES++ is a perfect tool to be coupled with transport codes (Hinton and Hazeltine 1976; Hirshman and Jardin 1979), so that the evolution of plasma equilibrium is simulated at the resistive timescale consistently with transport processes. Reciprocally the transport codes take into account the precise geometry of the magnetic flux lines. The numerical stability of such a coupling is challenging and subject of ongoing research. This is particularly important for the simulation and optimization of scenarios in new generation tokamaks. CEDRES++, when coupled to the transport codes CRONOS (Artaud et al. 2010) and ETS (Coster et al. 2010), is in use for simulating such self-consistent plasma evolution. The evolution mode itself, when plasma current profiles are given, is a good practical approach for vertical stability studies where the timescale of interest is much shorter than the current diffusion timescale of the plasma.

Furthermore, the modular and clear structure of CEDRES++ and the emphasis on accurate Newton methods, will make CEDRES++ very useful to implement fully automatic approaches to the optimization of scenarios. It will be easy to study and predict operational limits, and to devise control strategies that circumvent such limits.

It is possible to extend the methods presented in this work to higher order finite elements. Nevertheless, there are a couple of obstacles in order to obtain entirely higher order accurate methods:

- We are solving here a nonlinear elliptic problem with discontinuous coefficients (in the case of iron-transformer tokamaks) and discontinuous right-hand side. The

standard convergence theory for finite elements and elliptic regularity theory does not yield improved approximation results for polynomials of degree higher than 1. Nevertheless practical experience and the theoretical results in Feistauer and Sobotikova (1990) for low-order approximation of nonlinear problems and in Li et al. (2010) for high-order approximation of linear problems suggest improved accuracy. From the regularity theory for magneto-statics, we know in the case of iron-transformer tokamaks (such as JET or Tore Supra) that the solution lacks of regularity in the vicinity of the iron parts, in particular close to the interfaces air/iron and the corners of subdomains with iron. In such non-regular settings, it is required to switch to the so-called hp-version of the finite element method (Schwab 2004), that uses small triangles and low order polynomials in regions with non-regular solution and large triangles and high order polynomials elsewhere.

- Moreover, the general setting suffers from a fairly large modeling error due to the experimental permeability curve and to the axisymmetric representation of the ferromagnetic circuit that in reality consists of an iron core and a certain number of non-axisymmetric return limbs. It is not clear whether such modeling errors might surpass the discretization error.
- Higher order accuracy requires also sufficiently accurate quadrature rules in the definition of the Galerkin methods. While such higher order quadrature rules are standard for (iso-parametric) finite elements, we foresee technical difficulties alongside with a considerable increase of computational complexity for the integrals over the intersection of the plasma with triangles. We need to implement sufficiently accurate quadrature rules for polygonal domains with non-straight boundaries. On top of this, we need to implement for the Newton method the derivatives of such quadrature rules.

We are planning to investigate such topics in the near future and to compare with alternative approaches. One promising alternative might be to switch after a couple of Newton iterations to a different discretization scheme that uses separate meshes and separate polynomial degrees for the representation of the flux in the plasma domain and its exterior. The nonlinear coupling of the plasma and its exterior will lead to iteration schemes that induce small variations of the two meshes from one iteration to the next. With this, both the location of the magnetic axis and X-point are *a priori* not limited to a finite set of points, as it is the case for linear Lagrangian elements with fixed global mesh.

Any higher order method will involve more sophisticated algorithms and it is hard to predict if the accuracy at a fixed computing time will drastically improve. If high accuracy is required at the moment, this can still be achieved with reasonable effort by simple mesh refinement. The running time of the current version of the code is not yet optimized and still it is possible to do calculations with over half a million unknowns in less than 7 min on a workstation.

Acknowledgments

This work, supported by the European Communities under the contract of Association between EURATOM-CEA was carried out within the framework of the Task Force on Integrated Tokamak Modelling of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

The authors would like to thank the reviewers for their comments that help improve the manuscript.

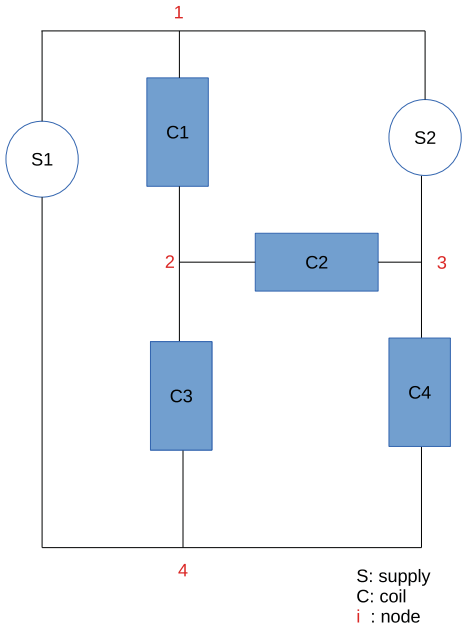

FIGURE A.1. Example of PF-circuit.

Appendix A. Circuit equations

In a tokamak, the poloidal field system is made of a number of circuits each comprising a number of coils and power supplies (see Fig. A.1). In the following we will present the derivation of the circuit equations for one single circuit.

Let N_N be the number of nodes, N_C the number of coils, and N_S the number of supplies of one circuit. To get a model for this circuit taking into account all those connections, the idea is to write the potential difference for each supply and each coil of the circuit. For a supply S, we have

$$U_k - U_l = V, \tag{A 1}$$

where V is the applied voltage and U_k and U_l the potential at the nodes that enclose the supply. For a coil C, the potential difference writes

$$U_i - U_j = R\frac{I}{n} + \frac{2\pi n}{S}\int_C \frac{\partial \psi}{\partial t} dr dz, \tag{A 2}$$

where U_i and U_j are the potentials at the nodes that enclose the coil, R is the total resistance of the coil C, S its cross section, n its number of wire turns and I the total current (in Ampère-turns). The average of $\frac{\partial \psi}{\partial t}$ over the coil C is an approximation of the discrete sum of inductive terms seen by the various turns of the coil. This approximation is perfectly valid, if one has a homogeneous distribution of turns in the coil.

We also consider Kirchhoff's current law at each node of the circuit adding N_N equations to the system. To fix the potential, we suppose that $U_1 = 0$. Thus, we get a set of $N_{eq} = N_S + N_C + N_N + 1$ equations which can be written in the form

$$\boldsymbol{A}\vec{U} = \boldsymbol{B}\vec{V} + \boldsymbol{C}\vec{I} + \boldsymbol{D}\vec{\Psi}(\partial_t \psi). \tag{A 3}$$

where the N_S first equations represent (A 1) and the following N_C equations are (A 2). The last equation of the system fixes the potential U_1. The matrices $\boldsymbol{A} \in \mathbb{R}^{N_{eq} \times N_N}$, $\boldsymbol{B} \in \mathbb{R}^{N_{eq} \times N_S}$, $\boldsymbol{C} \in \mathbb{R}^{N_{eq} \times (N_S + N_C)}$ and $\boldsymbol{D} \in \mathbb{R}^{N_{eq} \times N_C}$ are called *potential matrix*, *voltage matrix*, *current matrix* and *induction matrix*, respectively. The vectors $\vec{U} \in \mathbb{R}^{N_N}$, $\vec{V} \in \mathbb{R}^{N_S}$ and $\vec{I} \in \mathbb{R}^{N_S + N_C}$ contain the electric potential at the circuit nodes, the

voltages applied at the supplies and currents at the coils and the supplies. The components of the vector $\vec{\Psi}(\partial_t \psi) \in \mathbb{R}^{N_C}$ are the integrals $\int_C \partial_t \psi dr dz$ over the domain that is occupied by a coil C of the circuit.

For given $\vec{V}$, $\vec{I}$ and $\vec{\Psi}(\partial_t \psi)$, there is a unique $\vec{U}$ which satisfies (A 3), hence $\boldsymbol{A}^T \boldsymbol{A}$ is regular. We find

$$\vec{U} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{B} \vec{V} + (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{C} \vec{I} + (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{D} \vec{\Psi}(\partial_t \psi), \tag{A 4}$$

plug this into (A 3) and get

$$\boldsymbol{E} \vec{I} + \boldsymbol{F} \vec{V} + \boldsymbol{G} \vec{\Psi}(\partial_t \psi) = 0\,, \tag{A 5}$$

with

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{C} - \boldsymbol{C} \\ \boldsymbol{F} &= \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{B} - \boldsymbol{B} \\ \boldsymbol{G} &= \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{D} - \boldsymbol{D}. \end{aligned}$$

The system (A 5) of N_{eq} equations is over determined and $\vec{I}$ can be computed using the normal equation

$$\vec{I} = \boldsymbol{S} \vec{V} + \boldsymbol{R} \vec{\Psi}(\partial_t \psi) \tag{A 6}$$

with $\boldsymbol{S} = -(\boldsymbol{E}^T \boldsymbol{E})^{-1} \boldsymbol{E}^T \boldsymbol{F}$ and $\boldsymbol{R} = -(\boldsymbol{E}^T \boldsymbol{E})^{-1} \boldsymbol{E}^T \boldsymbol{G}$.

REFERENCES

Albanese, R., Blum, J. and De Barbieri, O. 1986 On the solution of the magnetic flux equation in an infinite domain. In: *EPS. 8th Europhysics Conf. on Computing in Plasma Physics*, European Physical Society, Mulhouse, France, pp. 41–44.

Albanese, R., Blum, J. and De Barbieri, O. 1987 Numerical studies of the Next European Torus via the PROTEUS code. In: *12th Conf. on Numerical Simulation of Plasmas, San Francisco*.

Albanese, R. and Villone, F. 1998 The linearized CREATE-L plasma response model for the control of current, position and shape in tokamaks. *Nucl. Fusion* **38**(5), 723.

Ané, J. M., Grandgirard, V., Albajar, F. and Johner, J. 2000 Design of next step tokamak: consistent analysis of plasma performance flux composition and poloidal field system. In: *18th IAEA Fusion Energy Conf., Sorrento,* `http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/33/029/33029055.pdf`.

Ariola, M. and Pironti, A. 2008 *Magnetic Control of Tokamak Plasmas*. London: Springer.

Artaud, J. F. et al. 2010 The CRONOS suite of codes for integrated tokamak modelling. *Nucl. Fusion* **50**(4), 043 001.

Bielak, J. and MacCamy, R. C. 1991 Symmetric finite element and boundary integral coupling methods for fluid-solid interaction. *Q. Appl. Math.* **49**(1), 107–119.

Blum, J. 1989 *Numerical Simulation and Optimal Control in Plasma Physics: With Applications to Tokamaks*. Paris: Wiley/Gauthier-Villars.

Blum, J., Boulbe, C. and Faugeras, B. 2012 Reconstruction of the equilibrium of the plasma in a tokamak and identification of the current density profile in real time. *J. Comput. Phys.* **231**(3), 960–980.

Blum, J. and Heumann, H. 2014 Optimal control for quasi-static evolution of plasma equilibrium in tokamaks. *Technical Report*, INRIA, Sophia Antipolis, HAL Id: hal-00988045, version 1.

Blum, J., Le Foll, J. and Thooris, B. 1981 The self-consistent equilibrium and diffusion code SCED. *Comput. Phys. Commun.* **24**, 235–254.

Bucalossi, et al. 2011 Feasibility study of an actively cooled tungsten divertor in tore supra for ITER technology testing. *Fusion Eng. Des.* **86**(6–8), 684–688.

Chen, G. and Zhou, J. 1992 *Boundary Element Methods (Computational Mathematics and Applications)* . London: Academic Press, Ltd.

Ciarlet, P. G. 1978 *The Finite Element Method for Elliptic Problems*. Amsterdam: North-Holland Publishing Co.

Cooper, W. A., Hirshman, S. P., Merkel, P., Graves, J. P., Kisslinger, J., Wobig, H. F. G., Narushima, Y., Okamura, S. and Watanabe, K. Y. 2009 Three-dimensional anisotropic pressure free boundary equilibria. *Comput. Phys. Commun.* **180**(9), 1524–1533.

Costabel, M. 1987 Principles of boundary element methods. In: *Finite Elements in Physics (Lausanne, 1986)*, Amsterdam: North-Holland, pp. 243–274.

Costabel, M. and Stephan, E. P. 1990 Coupling of finite and boundary element methods for an elastoplastic interface problem. *SIAM J. Numer. Anal.* **27**(5), 1212–1226.

Coster, D. P., Basiuk, V., Pereverzev, G., Kalupin, D., Zagórksi, R., Stankiewicz, R., Huynh, P. and Imbeaux, F. 2010 The european transport solver. *IEEE Trans. Plasma Sci.* **38**(9), 2085–2092.

Davis, T. A. 2011 Suitesparse: a suite of sparse matrix software. `http://faculty.cse.tamu.edu/davis/suitesparse.html`.

Degtyarev, L. M. and Drozdov, V. V. 1985 An inverse variable technique in the MHD-equilibrium problem. *Comput. Phys. Rep.* **2**(7), 341–387.

Degtyarev, L. M. and Drozdov, V. V. 1991 Adaptive mesh computation of magnetic hydrodynamic equilibrium. *Int. J. Mod. Phys.* C **02**(01), 30–38.

Delfour, M. C. and Zolésio, J.-P. 2011 *Shapes and Geometries*, 2nd edn., *Advances in Design and Control*, Vol. 22. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

DeLucia, J., Jardin, S. C. and Todd, A. M. M. 1980 An iterative metric method for solving the inverse tokamak equilibrium problem. *J. Comput. Phys.* **37**(2), 183–204.

Falchetto, G. L.et al. 2014 The European Integrated Tokamak Modelling (ITM) effort: achievements and first physics results. *Nucl. Fusion* **54**(4), 043 018.

Feistauer, M. and Sobotikova, V. 1990 Finite element approximation of nonlinear elliptic problems with discontinuous coefficients. *ESAIM: Math. Modelling Numer. Anal. - Modlisation Mathématique et Analyse Numrique* **24**(4), 457–500.

Feneberg, W. and Lackner, K. 1973 Multipole tokamak equilibria. *Nucl. Fusion* **13**(4), 549.

Fitzgerald, M., Appel, L. C. and Hole, M. J. 2013 EFIT tokamak equilibria with toroidal flow and anisotropic pressure using the two-temperature guiding-centre plasma. *Nucl. Fusion* **53**(11), 113 040.

Freidberg, J. P. 1987 *Ideal Magnetohydrodynamics*. US: Plenum.

Gatica, G. N. and Hsiao, G. C. 1995 The uncoupling of boundary integral and finite element methods for nonlinear boundary value problems. *J. Math. Anal. Appl.* **189**(2), 442–461.

Glowinski, R. and Marrocco, A. 1974 Analyse numérique du champ magnétique d'un alternateur par éléments finis et sur-relaxation ponctuelle non linéaire. *Comput. Methods Appl. Mech. Eng.* **3**(1), 55–85.

Goedbloed, J. P., Keppens, R. and Poedts, S. 2010 *Advanced Magnetohydrodynamics: with Applications to Laboratory and Astrophysical Plasmas*. Cambridge: Cambridge University Press.

Goedbloed, J. P. and Lifschitz, A. 1997 Stationary symmetric magnetohydrodynamic flows. *Phys. Plasmas (1994–present)* **4**(10), 3544–3564.

Goedbloed, J. P. and Poedts, S. 2004 *Principles of Magnetohydrodynamics: with Applications to Laboratory and Astrophysical Plasmas*. Cambridge: Cambridge University Press.

Grad, H. 1967 Toroidal containment of a plasma. *Phys. Fluids (1958–1988)* **10**(1), 137–154.

Grad, H. and Rubin, H. 1958 Hydromagnetic equilibria and force-free fields. In: *Proc. 2nd UN Conf. on the Peaceful Uses of Atomic Energy*, United Nations Publications, Geneva, Vol. 31, p. 190.

Grandgirard, V. 1999 Modélisation de l'équilibre d'un plasma de tokamak. *PhD thesis*, l'Université de Franche-Comté.

Gruber, R., Iacono, R. and Troyon, F. 1987 Computation of MHD equilibria by a quasi-inverse finite hybrid element approach. *J. Comput. Phys.* **73**(1), 168–182.

Guazzotto, L., Betti, R., Manickam, J. and Kaye, S. 2004 Numerical study of tokamak equilibria with arbitrary flow. *Phys. Plasmas (1994–present)* **11**(2), 604–614.

Helton, F. J. and Wang, T. S. 1978 MHD equilibrium in non-circular tokamaks with field-shaping coil systems. *Nucl. Fusion* **18**(11), 1523.

Hertout, P., Boulbe, C., Nardon, E., Blum, J., Bremond, S., Bucalossi, J., Faugeras, B., Grandgirard, V. and Moreau, P. 2011 The cedres++ equilibrium code and its application to ITER, JT-60SA and Tore Supra. *Fusion Eng. Des.* **86**, 1045–1048.

Hinton, F. L. and Hazeltine, R. D. 1976 Theory of plasma transport in toroidal confinement systems. *Rev. Mod. Phys.* **48**, 239–308.

Hiptmair, R. 2003 Coupling of finite elements and boundary elements in electromagnetic scattering. *SIAM J. Numer. Anal.* **41**(3), 919–944.

Hirshman, S. P. and Betancourt, O. 1991 Preconditioned descent algorithm for rapid calculations of magnetohydrodynamic equilibria. *J. Comput. Phys.* **96**(1), 99–109.

Hirshman, S. P. and Jardin, S. C. 1979 Two-dimensional transport of tokamak plasmas. *Phys. Fluids* **22**(4), 731–742.

Hofmann, F. and Tonetti, G. 1988 Tokamak equilibrium reconstruction using Faraday rotation measurements. *Nucl. Fusion* **28**(10), 1871.

Huysmans, G. T. A. and Czarny, O. 2007 MHD stability in X-point geometry: simulation of ELMs. *Nucl. Fusion* **47**(7), 659.

Huysmans, G. T. A., Goedbloed, J. P. and Kerner, W. 1991 Isoparametric bicubic Hermite elements for solution of the Grad–Shafranov equation. In: *Proc. CP90 Conf. on Comp. Phys.*, Vol. 2, pp. 371–376.

ITM 2013 Integrated tokamak modelling. `http://portal.efda-itm.eu/`, integrated Tokamak Modelling.

Jardin, S. C. 2010 *Computational Methods in Plasma Physics*. Boca Raton, Florida: CRC Press/Taylor and Francis.

Jardin, S. C., Pomphrey, N. and DeLucia, J. 1986 Dynamic modeling of transport and positional control of tokamaks. *J. Comput. Phys.* **66**, 481–507.

Johnson, J. L. et al. 1979 Numerical determination of axisymmetric toroidal magnetohydrodynamic equilibria. *J. Comput. Phys.* **32**(2), 212–234.

Lackner, K. 1976 Computation of ideal MHD equilibria. *Comput. Phys. Commun.* **12**(1), 33–44.

Lao, L. L., Ferron, J. R., Geoebner, R. J., Howl, W., St. John, H. E., Strait, E. J. and Taylor, T. S. 1990 Equilibrium analysis of current profiles in Tokamaks. *Nucl. Fusion* **30**(6), 1035.

Lao, L. L., Hirshman, S. P. and Wieland, R. M. 1981 Variational moment solutions to the Grad–Shafranov equation. *Phys. Fluids* **24**(8), 1431–1440.

Lao, L. L., John, H. St., Stambaugh, R. D., Kellman, A. G. and Pfeiffer, W. 1985 Reconstruction of current profile parameters and plasma shapes in tokamaks. *Nucl. Fusion* **25**(11), 1611.

Li, J., Melenk, J. M., Wohlmuth, B. and Zou, J. 2010 Optimal a priori estimates for higher order finite elements for elliptic interface problems. *Appl. Numer. Math.* **60**(1–2), 19–37.

Ling, K. M. and Jardin, S. C. 1985 The Princeton spectral equilibrium code: PSEC. *J. Comput. Phys.* **58**(3), 300–335.

Lüst, R. and Schlüter, A. 1957 Axialsymmetrische magnetohydrodynamische Gleichgewichtskonfigurationen. Z. Naturforsch. **A12**, 850–854.

Luxon, J. L. and Brown, B. B. 1982 Magnetic analysis of non-circular cross-section tokamaks. *Nucl. Fusion* **22**(6), 813.

Maschke, E. K. and Perrin, H. J. 1984 An analytic solution of the stationary {MHD} equations for a rotating toroidal plasma. *Phys. Lett.* A **102**(3), 106–108.

Mc Carthy, P. J., Martin, P. and Schneider, W. 1999 The CLISTE interpretive equilibrium code. *Technical Report IPP Report 5/85*. Max-Planck-Institut fur Plasmaphysik.

Murat, F. and Simon, J. 1976 Sur le contrôle par un domaine géométrique. *Technical Report 76015*. Laboratoire d'Analyse Numérique, Université de Paris 6.

Nédélec, J.-C. 2001 *Acoustic and Electromagnetic Equations (Applied Mathematical Sciences, 144)*. New York: Springer-Verlag.

Nocedal, J. and Wright, S. J. 2006 *Numerical Optimization*, 2nd edn. *Springer Series in Operations Research and Financial Engineering*. New York: Springer.

Parail, V. et al. 2013 Self-consistent simulation of plasma scenarios for iter using a combination of 1.5d transport codes and free-boundary equilibrium codes. *Nucl. Fusion* **53**(11), 113 002.

Park, W., Belova, E. V., Fu, G. Y., Tang, X. Z., Strauss, H. R. and Sugiyama, L. E. 1999 Plasma simulation studies using multilevel physics models. *Phys. Plasmas (1994-present)* **6**(5), 1796–1803.

Pechstein, C. and Jüttler, B. 2006 Monotonicity-preserving interproximation of B-H-curves. *J. Comput. Appl. Math.* **196**(1), 45–57.

Pustovitov, V. D. 2010 Anisotropic pressure effects on plasma equilibrium in toroidal systems. *Plasma Phys. Control. Fusion* **52**(6), 065 001.

Renard, Y. and Pommier, J. 2014 GetFEM++, an open-source finite element library. `http://download.gna.org/getfem/html/homepage/index.html`.

Schwab, C. 2004 *p- and hp- Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics*. Oxford: Clarendon Press.

Shafranov, V. D. 1958 On magnetohydrodynamical equilibrium configurations. *Sov. J. Exp. Theor. Phys.* **6**, 545.

Shewchuk, J. R. 1996 Triangle: Engineering a 2D quality mesh generator and delaunay triangulator. In: *Applied Computational Geometry: Towards Geometric Engineering*, Vol. 1148 (ed. M. C. Lin and D. Manocha), *Lecture Notes in Computer Science*, Springer-Verlag, from the First ACM Workshop on Applied Computational Geometry, Springer, Berlin, pp. 203–222.

Stephan, E. P. 1992 Coupling of finite elements and boundary elements for some nonlinear interface problems. *Comput. Methods Appl. Mech. Eng.* **101**(1–3), 61–72.

Turkington, B., Lifschitz, A., Eydeland, A. and Spruck, J. 1993 Multiconstrained variational problems in magnetohydrodynamics: equilibrium and slow evolution. *J. Comput. Phys.* **106**(2), 269–285.

Wesson, J. and Campbell, D. J. 2004 *Tokamaks*. *The International Series of Monographs in Physics*, 3rd edn. Vol. 118, Oxford: Clarendon Press.

Zhao, K., Vouvakis, M. N. and Lee, J.-F. 2006 Solving electromagnetic problems using a novel symmetric fem-bem approach. *IEEE Trans. Magn.* **42**(4), 583–586.

Zwingmann, W., Eriksson, L.-G. and Stubberfield, P. 2001 Equilibrium analysis of tokamak discharges with anisotropic pressure. *Plasma Phys. Control. Fusion* **43**(11), 1441.

Article F : [5] : G. L. Falchetto, D. Coster, R. Coelho, B.D. Scott, L. Figini, D. Kalupin, E. Nardon, L.L. Alves, J.F. Artaud, V. Basiuk, J. Bizarro, C. Boulbe, A. Dinklage, D. Farina, B. Faugeras, J. Ferreira, A. Figueiredo, P. Huynh, F. Imbeaux, I. Ivanova-Stanik, T. Jonsson, H.-J. Klingshirn, C. Konz, A. Kus, N.B. Marushchenko, E. Nardon, S. Nowak, G. Pereverzev, M. Owsiak, E. Poli, Y. Peysson, R. Reimer, J. Signoret, O. Sauter, R. Stankiewicz, P. Strand, I. Voitsekhovitch, E. Westerhof, T. Zok, W. Zwingmann, ITM-TF contributors, the ASDEX Upgrade Team et JET-EFDA Contributors. The European Integrated Tokamak Modelling (ITM) effort : achievements and first physics results. *Nucl. Fusion* 54.4 (2014), p. 043018. doi : 10.1088/0029-5515/54/4/043018

The European Integrated Tokamak Modelling (ITM) effort: achievements and first physics results

2014 Nucl. Fusion 54 043018

(http://iopscience.iop.org/0029-5515/54/4/043018)

IOP Publishing | International Atomic Energy Agency Nuclear Fusion

Nucl. Fusion **54** (2014) 043018 (19pp) doi:10.1088/0029-5515/54/4/043018

The European Integrated Tokamak Modelling (ITM) effort: achievements and first physics results

G.L. Falchetto[1], D. Coster[2], R. Coelho[3], B.D. Scott[2], L. Figini[4], D. Kalupin[5], E. Nardon[1], S. Nowak[4], L.L. Alves[3], J.F. Artaud[1], V. Basiuk[1], João P.S. Bizarro[3], C. Boulbe[6], A. Dinklage[7], D. Farina[4], B. Faugeras[6], J. Ferreira[3], A. Figueiredo[3], Ph. Huynh[1], F. Imbeaux[1], I. Ivanova-Stanik[8], T. Jonsson[9], H.-J. Klingshirn[2], C. Konz[2], A. Kus[7], N.B. Marushchenko[7], G. Pereverzev[2,c], M. Owsiak[10], E. Poli[2], Y. Peysson[1], R. Reimer[7], J. Signoret[1], O. Sauter[11], R. Stankiewicz[8], P. Strand[12], I. Voitsekhovitch[13], E. Westerhof[14], T. Zok[10], W. Zwingmann[15], ITM-TF Contributors[a], the ASDEX Upgrade Team and JET-EFDA Contributors[b]

[1] CEA, IRFM, F-13108 Saint-Paul-lez-Durance, France
[2] Max-Planck-Institut für Plasmaphysik, EURATOM-IPP Association, Garching, Germany
[3] Associação EURATOM/IST, Instituto de Plasmas e Fusão Nuclear, Instituto Superior Técnico, Universidade Técnica de Lisboa 1049-001 Lisboa, Portugal
[4] Istituto di Fisica del Plasma CNR, Euratom-ENEA-CNR Association, 20125 Milano, Italy
[5] EFDA-CSU Garching, Boltzmannstr. 2, D-85748, Garching, Germany
[6] Univ. Nice Sophia Antipolis, Lab. JA Dieudonne, UMR 7351, F-06108 Nice 02, France
[7] Max Planck Institut für Plasmaphysik, EURATOM Association, Greifswald, Germany
[8] Institute of Plasma Physics and Laser Microfusion, EURATOM Association, 00-908 Warsaw, Poland
[9] Association EURATOM-VR, Fusion Plasma Physics, EES, KTH, SE-10037 Stockholm, Sweden
[10] Poznan Supercomputing and Networking Center, IChB PAS, Noskowskiego 12/14, Poznan, Poland
[11] Ecole Polytechnique Federale de Lausanne (EPFL), Centre de Recherches en Physique des Plasmas (CRPP), Association Euratom-Confederation Suisse, Lausanne, Switzerland
[12] Department of Earth and Space Sciences, Chalmers University of Technology, Euratom-VR Association, SE-352 96 Göteborg, Sweden
[13] EURATOM/CCFE Fusion Association, Culham Science Centre, Abingdon OX14 3DB, UK
[14] FOM Institute DIFFER, Association EURATOM-FOM, Nieuwegein, The Netherlands
[15] European Commission, Directorate-General for Research and Innovation, B-1049 Brussels, Belgium

E-mail: gloria.falchetto@cea.fr

Received 8 March 2013, revised 27 January 2014
Accepted for publication 14 February 2014
Published 20 March 2014

Abstract

A selection of achievements and first physics results are presented of the European Integrated Tokamak Modelling Task Force (EFDA ITM-TF) simulation framework, which aims to provide a standardized platform and an integrated modelling suite of validated numerical codes for the simulation and prediction of a complete plasma discharge of an arbitrary tokamak. The framework developed by the ITM-TF, based on a generic data structure including both simulated and experimental data, allows for the development of sophisticated integrated simulations (workflows) for physics application.

[a] See the appendix.
[b] See the appendix of Romanelli F. *et al* 2012 *Proc. 24th IAEA Fusion Energy Conf. (San Diego, CA, 2012)* (www-naweb.iaea.org/napc/physics/FEC/FEC2012/html/proceedings.pdf).
[c] Deceased.

The equilibrium reconstruction and linear magnetohydrodynamic (MHD) stability simulation chain was applied, in particular, to the analysis of the edge MHD stability of ASDEX Upgrade type-I ELMy H-mode discharges and ITER hybrid scenario, demonstrating the stabilizing effect of an increased Shafranov shift on edge modes. Interpretive simulations of a JET hybrid discharge were performed with two electromagnetic turbulence codes within ITM infrastructure showing the signature of trapped-electron assisted ITG turbulence. A successful benchmark among five EC beam/ray-tracing codes was performed in the ITM framework for an ITER inductive scenario for different launching conditions from the equatorial and upper launcher, showing good agreement of the computed absorbed power and driven current. Selected achievements and scientific workflow applications targeting key modelling topics and physics problems are also presented, showing the current status of the ITM-TF modelling suite.

Keywords: integrated modelling, simulation, code verification, turbulence, transport

(Some figures may appear in colour only in the online journal)

1. Introduction

The European Integrated Tokamak Modelling Task Force (ITM-TF) [1, 2] aims at providing a standardized platform and an integrated modelling suite of validated numerical codes for the simulation and prediction of a complete plasma discharge in arbitrary tokamaks. In order to address such a challenge, the ITM-TF approach builds on a modelling infrastructure, focusing on the development of a data and communication ontology, i.e. standardizing the data exchange between different codes, through a generic data structure incorporating both simulated and experimental data. The elements of this data structure are identified as 'Consistent Physical Objects', or CPO [3]. Physics modules of various complexities can be easily adapted to the data structure, which is code and language agnostic.

Thanks to the standardization of I/O through CPOs, physics modules can be seamlessly coupled into different integrated simulations (workflows); also, modules describing the same physics (e.g. equilibrium, transport modules, heating) can be easily interchanged within the same workflow, so to allow the physicist to choose and easily integrate the more appropriate model to tackle a specific physics problem. Moreover, in the ITM-TF framework all machine related data are extracted into standardized machine descriptions (MD) so that physics modules, like equilibrium reconstruction tools, also become independent of the specific tokamak experiment.

The ITM-TF uses the open-source Kepler[16] scientific workflow manager and orchestrator tool, which allows for a user-friendly graphical construction of the integrated simulation. Physics modules enter as *actors* of a Kepler workflow; all the data transfer among actors within a workflow occurs via CPOs. In Kepler, semantic types can be defined which allow one to distinguish different CPOs and therefore verify whether a CPO output of an actor is correctly connected to the corresponding CPO input of the subsequent actor. Furthermore, the Kepler framework allows for interactive steering of simulations, through its capability to pause the simulation and alter some parameters; also, users can easily include actors for visualizing the present state of a simulation.

The ITM-TF uses the Kepler framework for simple run orchestration (workflows without convergence loops, named hereafter *loosely coupled workflows*) as well as more complicated workflows, involving mutual interactions among different codes (within loops, named hereafter *tightly coupled workflows*). It has to be noted, though, that the generic data structure (CPOs) is totally independent of the used workflow orchestrator tool (Kepler or other), all the advantages of the generic data structure remain if the physics modules are called in a classic Fortran workflow, with CPOs as arguments.

The framework developed by the ITM-TF has allowed for the development of sophisticated workflows for physics applications. Among those, the European Transport Simulator (ETS) [4] workflow, a leading ITM tool for both interpretive and predictive transport simulations and scenario modelling, incorporating a sophisticated module for synergy effects between heating schemes, several equilibrium modules, pellets, impurities, neutrals, sawteeth and neoclassical tearing mode (NTM) modules, as well as a variety of neoclassical and turbulence-transport modules of different complexity. In this paper, selected achievements targeting key modelling topics and physics problems are outlined, showing the present status of the ITM-TF modelling suite. Moreover, it is worth mentioning that the modules which can be coupled into ITM workflows can be either centrally distributed (residing on the common ITM Gateway cluster) or may be supplied by the user (in whichever programming language, including interpreted languages like Python and Matlab, provided they are CPO compliant independent modules).

First, we present applications of simple (loosely coupled) workflows. Physics results on the MHD equilibrium and linear stability of the plasma edge of ASDEX Upgrade and ITER hybrid scenario [5] as well as interpretive studies of a JET discharge using gyrofluid and gyrokinetic turbulence models are reported in section 2. We conclude section 2.2, with an illustration of tightly coupled turbulence-transport workflows developed in the ITM framework.

The physics modules integrated into the different ITM workflows are being cross-verified within the ITM framework, as well as against existing integrated modelling codes to guarantee both their interchangeability and their validation. Results from a thorough benchmarking of electron cyclotron heating and current drive codes [6] on an ITER H-mode scenario for different launching conditions both from the equatorial launcher (EL) and upper launcher (UL) are shown in section 3.1. Section 3.2 reports the ETS successful

[16] http://kepler-project.org.

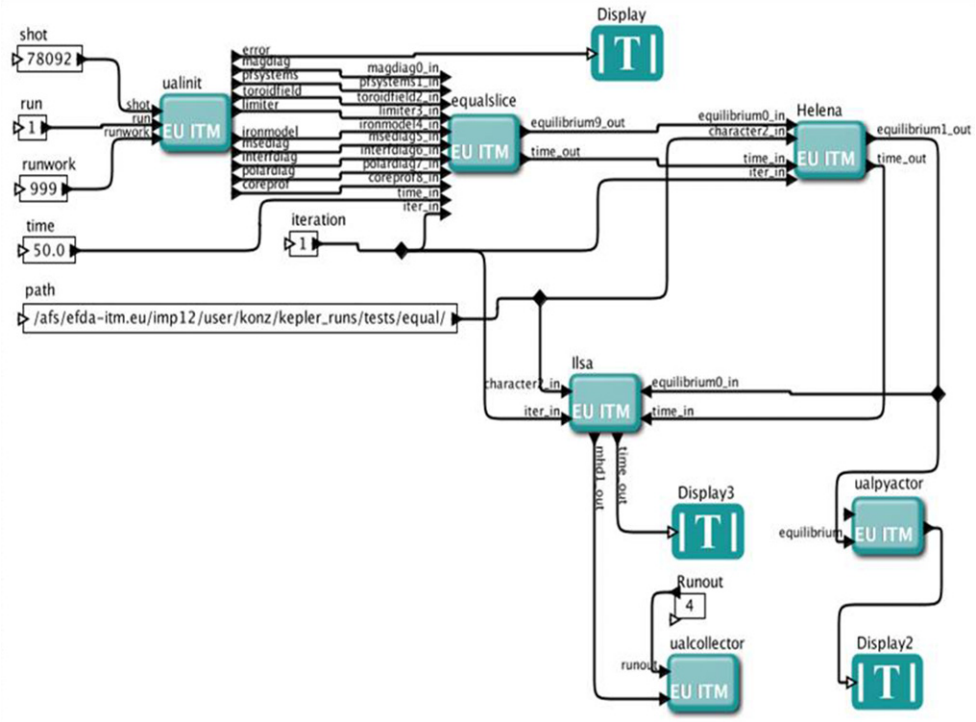

Figure 1. ITM-TF Kepler workflow for MHD linear stability coupling: an initialization module (ualinit) reading experimental data, EQUAL, HELENA and ILSA modules. A python script actor (ualpyactor) provides the visualization of the reconstructed equilibrium. Replacing equal slice with the j-alpha module allows one to perform a parameter study by modifying pressure and plasma current.

benchmarking against leading tokamak plasma core transport codes on a JET hybrid discharge [7]. In order to illustrate the flexibility and wide range of use cases for scientific workflows, section 4 focuses on other relevant examples of tightly coupled workflows developed by the ITM-TF. Firstly, an application of a direct coupling of the ETS core transport solver to a two-dimensional (2D) edge transport code, demonstrated for the particular case of steady state and multiple impurities [8], is shown. The second example addresses the effect of NTMs on plasma transport and confinement, incorporated in ETS workflows via a dedicated NTM module that calculates the island frequency, width and associated reshaping in transport coefficients. Finally, a successful proof-of-principle application of an ETS workflow including the coupling with a free-boundary equilibrium (FBE) code, to the simulation of a vertical displacement event (VDE), is presented together with details on the coupling algorithm.

Finally, in section 5 recent results are shown of the ongoing effort in ITM-TF to incorporate synthetic diagnostics [9] into the modelling framework (fusion products, three-dimensional (3D) reflectometry, motional Stark effect (MSE), neutron and neutral particle analyser (NPA)), focusing on synthetic MSE spectra and comparison to the experimental data.

2. Physics results

A selection of some of the first physics results produced using the ITM-TF framework is presented in the following subsections.

2.1. Equilibrium reconstruction and linear MHD stability

The first demonstration of the use of ITM-TF integrated simulation workflows for physics studies on experimental data addressed equilibrium reconstruction, refinement and linear MHD stability calculations [5]. The corresponding Kepler workflow is illustrated in figure 1, *actors* for FBE reconstruction (e.g. EQUAL [10, 11]), high-resolution fixed-boundary Grad–Shafranov solver (e.g. HELENA [12] or CHEASE [13]), and linear MHD stability (e.g. ILSA [14] or MARS-F [15]) are seamlessly integrated in the workflow environment. The machine independent equilibrium reconstruction code EQUAL developed within the ITM-TF has been extensively validated (at a first stage with magnetic data only) on JET discharges [16].

An analysis of the edge MHD stability of ASDEX Upgrade type-I ELMy H-mode discharges was carried out, using the stability chain coupling CLISTE, HELENA and ILSA [5]. CLISTE is a FBE reconstruction code using input from poloidal field (PF) coil currents, magnetic and possibly kinetic plasma profile diagnostic measurements. The reconstructed coarse equilibrium is then passed to the high-resolution reconstruction code HELENA and this refined equilibrium is used by the linear MHD stability code suite ILSA (in the particular case addressed here the ideal MISHKA code module of ILSA was used [13]).

Replacing the equilibrium actor with a JALPHA actor, which reads a previously calculated fixed-boundary equilibrium from the database, modifies the pressure profile and/or the flux surface averaged current density and computes the new high-resolution equilibrium, a j–α workflow is created. Stability diagrams can then be automatically

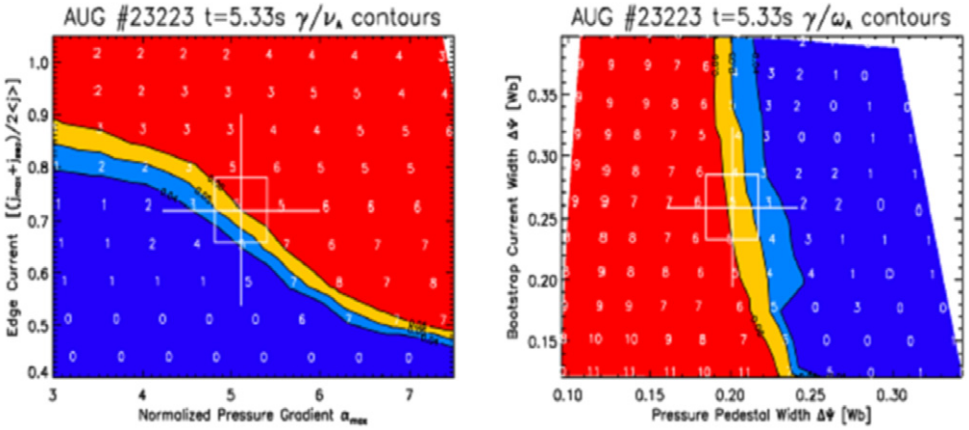

Figure 2. Pedestal height (left) and width (right) study for ASDEX Upgrade obtained with the J-alpha stability workflow ([8] with kind permission of the European Physical Journal (EPJ)). The plot shows the contours of the linear ideal MHD growth rates γ (normalized to the Alfvén frequency v_A) of the fastest growing edge modes (toroidal mode numbers are indicated by the white integers) in the plane defined by the maximum normalized edge pressure gradient α_{max} and the normalized edge current density. Contours indicate the level of the diamagnetic drift frequency separating the stable (blue) from the unstable (red) region. The crosshair indicates the experimental equilibrium including error bars.

computed using Kepler, by wrapping the linear j–α workflow in a double loop over the pressure and current scaling parameters. Computation times being substantial for such scans, the ITM-TF developed, in cooperation with the FP7 project EUFORIA [17], Kepler workflows for automatic job submission to Grid and Cloud infrastructures.

For pedestal height studies, the pressure and current density profiles in the edge can be scaled by a constant factor, while the core profiles are adapted to keep the plasma energy W_{MHD} and the total plasma current I_p unchanged. For pedestal width studies, the widths of the pressure and current density pedestals can be scaled independently, again adjusting the core profiles such that W_{MHD} and I_p remain the same. In this case, the pressure at the pedestal top and the amplitude of the bootstrap current remain constant, only the gradients change through variation of the width. Therefore, the total current flowing in the edge is smaller if the width is reduced.

Figure 2 shows the stability diagrams for the variation of the pedestal height and width for ASDEX Upgrade shot #23223 at $t = 5.33$ s. The profiles were taken just before the crash of type-I ELMs. As expected, the experimental equilibrium is marginally unstable with a toroidal mode number ($n = 5$) indicating a strong peeling component. Reducing the pedestal width, and thereby increasing the gradients, clearly drives the equilibrium unstable. It may also be noted (figure 2 right) that the drive from the current density gradient (small bootstrap current width) dominates the drive from larger edge current (large bootstrap current width).

Core and pedestal scans of the normalized plasma beta β_N (applying, respectively, a scaling factor only on the core pressure profile or on the full profile) were also performed using the linear MHD stability chain for the ASDEX Upgrade type-I ELMy shot #20116 at $t = 3.59$ s as well as for an ITER hybrid 'scenario 2' (kinetic profiles of the used ITER scenario 2, for the reference $\beta_N = 1.8$, are shown in figure 3).

The most unstable mode growth rates for the two scans are shown in figure 4. It is evident from the computed growth rates in dashed lines that the increased Shafranov shift helps stabilizing edge modes (external kink modes of intermediate n). When scaling the entire pressure profile (solid lines), the destabilizing effect of the larger edge pressure gradient strongly dominates over the stabilizing effect by the Shafranov shift, inducing the destabilization of a (high n) ballooning mode. The ITER case shows a slight destabilization of a (low n) pure peeling mode for large Shafranov shifts ($\beta_N > 2.75$).

2.2. Turbulence simulations

A simple workflow allows conventional methods of comparing a turbulence code's transport results to experimental measurements and transport analysis. Run in a double-blind fashion, the result is almost always discrepant. Physical insight into the problem usually depends on diagnosing these discrepancies. A hybrid JET shot (#77922) was used as a very interesting test-bed for radially local turbulence/transport computations, which happen to fail due to the set of parameters in the core-confinement region (between 0.4 and 0.7 in normalized radius).

Discussions of the observed discrepancies among different turbulence/transport code simulations have highlighted several issues of provenance, namely what is used for the equilibrium flux surface structure, and what is used to define the dimensionless parameters of the runs (in this case, gradients). The profiles of the case under study turned out to be close enough to stability thresholds that small differences in magnetic shear or in the choice of radial coordinates (e.g. ρ_{tor} versus the midplane-cut minor radius) are enough to make the difference between stability and weak turbulence.

The prescribed case was profile data from JET shot 77922 at time 47.7 s. The input data were provided by TRANSP [18] in interpretive mode from the actual experimental data which determine the profiles. Profiles of the electron density, electron and ion temperatures (ions hotter), and the toroidal current and pitch parameter q are shown in figure 5.

The case is read from the database into coreprof and equilibrium CPOs, and then fed to the rest of the workflow, represented in figure 6. Since the equilibrium_CPO did not

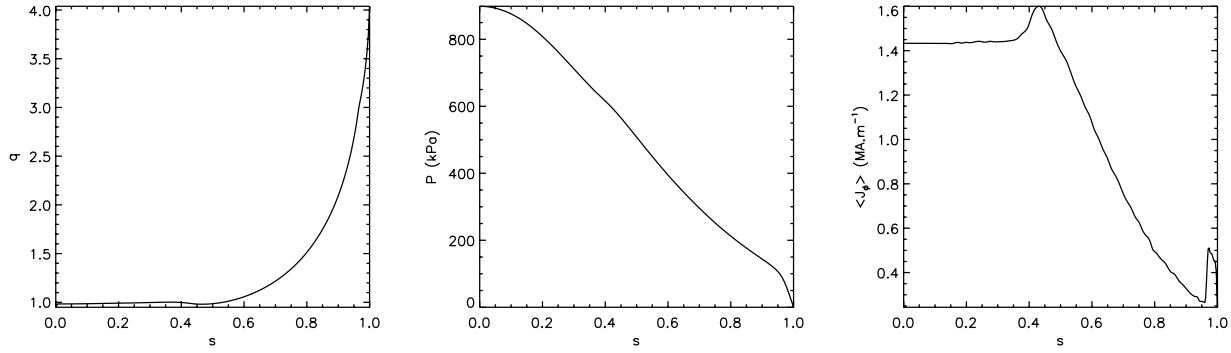

Figure 3. Safety factor, pressure and toroidal current profile of the used ITER scenario 2, for the reference $\beta_N = 1.8$. The radial coordinate is $s = \mathrm{sqrt}(\psi/\psi_{\mathrm{boundary}})$.

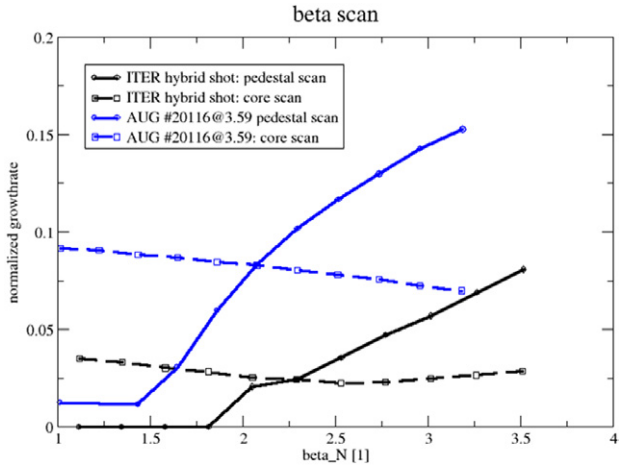

Figure 4. Core and pedestal scans of the normalized plasma beta for ASDEX Upgrade type-I ELMy shot #20116 (blue) and an ITER hybrid scenario (black) [9]. The dashed lines show modification of the plasma β_N via modification of the core pressure profile while keeping the pedestal pressure unchanged. The solid lines, on the other hand, show modification of the plasma β_N via scaling of the entire pressure profile.

contain sufficient information as needed by the successive flux-tube turbulence code modules (namely the pressure profile and the straight-field-line coordinate metric were missing), the workflow consisted of three actors: EQUPDATE which constructs equilibrium profile inputs for pressure and toroidal current from coreprof_CPO, and passes the equilibrium boundary surface, in this case the experimental separatrix, then the fixed-boundary Grad–Shafranov solver GKMHD which also fills the coord_sys element in the equilibrium_CPO, and then the turbulence code GEM, a flux tube gyrofluid model [19]. GEM actor is executed in batch on HPC-FF, running in parallel one flux tube at each of 0.4, 0.5, 0.6 and 0.7 normalized midplane-cut minor radii; it fills the coretransp_CPO and also provides the standard post-process diagnostics for turbulence.

The same workflow was used replacing GEM with delta-FEFI, a delta-f gyrokinetic turbulence code (parent model to GEM otherwise similar in structure [20]), for direct comparisons between the two models.

The use of the GKMHD module was needed because a theoretical s–α model was found to be a very poor approximate to these experimental cases which are in the shaped geometry of a diverted tokamak.

GKMHD sets up a regular triangular grid logically the same as placing flux surfaces onto nested hexagons. Each iteration consists of solving $-\Delta \times \psi = \mu_0 \langle J_{\mathrm{tor}} R\rangle + \mu_0(\mathrm{d}p/\mathrm{d}\psi)(R^2 - \langle R^2\rangle)$ where $\langle\rangle$ denotes flux surface average, p and $\langle J_{\mathrm{tor}} R\rangle/R_0$ are the input profiles, and then moving the grid points towards or away from the axis such that the prescribed normalized ψ of the surface agrees with the new values of $\psi(R,Z)$. Otherwise it is a conventional Grad–Shafranov solver taking pressure and current on input. Afterwards, the resulting equilibrium_CPO is filled with coordinate metric information needed by flux tube models. The midplane-cut minor radius is defined as $(r_{\mathrm{outboard}} - r_{\mathrm{inboard}})/2$ from the equilibrium_CPO; the normalized version is denoted as r_a below.

Both GEM and delta-FEFI take the straight-field-line coordinate metric on input and construct a field-aligned, shifted-metric coordinate system based on Hamada coordinates [21, 22]. The fluctuations are initialized as a single Maxwellian density structure localized at nonlinear amplitude with Gaussian profiles to 10 ion sound gyroradii (ρ_s) in the drift plane and to qR_0 along the field lines. The finite electron pressure launches shear-Alfvén waves and then a drift wave field at nonlinear amplitude, and the system proceeds to fully developed turbulence unless it is nonlinearly stable [21, 23].

Gyrofluid runs are held in saturation or decay for 4000 gyro-Bohm times ($\tau_{\mathrm{GB}} = L_\perp/c_s$, where $L_\perp$ is the steepest gradient scale length and c_s the ion sound speed). Gyrokinetic runs only went to $1000 L_\perp/c_s$ due to the far greater computational expense. Each flux tube is an independent run, with its own normalized units including normalized time, τ_{GB} being different for each case. The time step is $0.002\tau_{\mathrm{GB}}$, allowing for extreme transients which are found in the early stages of some core-parameter cases. The domain size is $20\pi\rho_s$ in the radial direction, $80\pi\rho_s$ in the drift-angle direction, and one connection length $2\pi qR$ in the parallel direction. The grid is $128\times128\times32$ in these directions, respectively. The numerical scheme is given in [19], mostly following [23]. Delta-FEFI uses the same scheme as GEM with the additional ingredient being the phase-space parallel bracket [20].

We concentrate on the case $r_a = 0.6$ since both codes found stability or on-threshold behaviour at 0.7. The normalized parameters (defined as in [24]) at $r_a = 0.6$ are beta-hat $= 0.38$, mu-hat $= 0.022$, $C = 3 \times 10^{-4}$, $T_i/T_e = 1.25$, $R/L_{T_i} = 6.30$, $L_{T_i}/L_{T_e} = 0.68$, $L_{T_i}/L_n = 0.38$ and $qR_0/L_{T_i} = 9.0$.

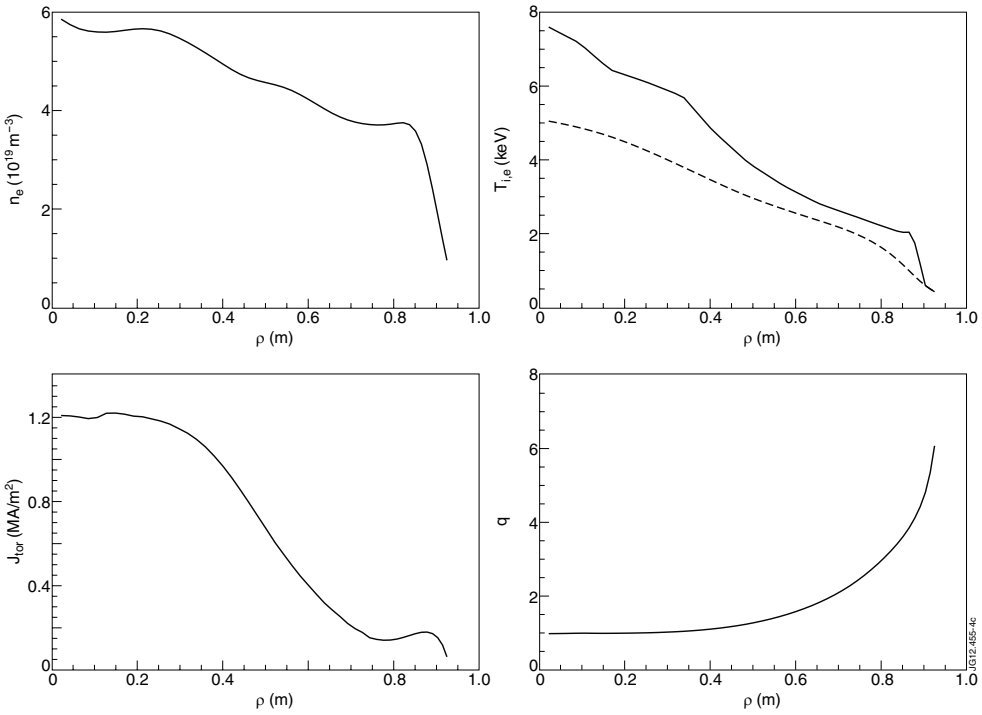

Figure 5. Profiles of the electron density, electron (dashed) and ion temperatures (full line), toroidal current and pitch parameter q, for the JET shot 77922 at 47.7 s given by the TRANSP interpretation through coreprof and equilibrium CPOs in the database case as discussed in the text.

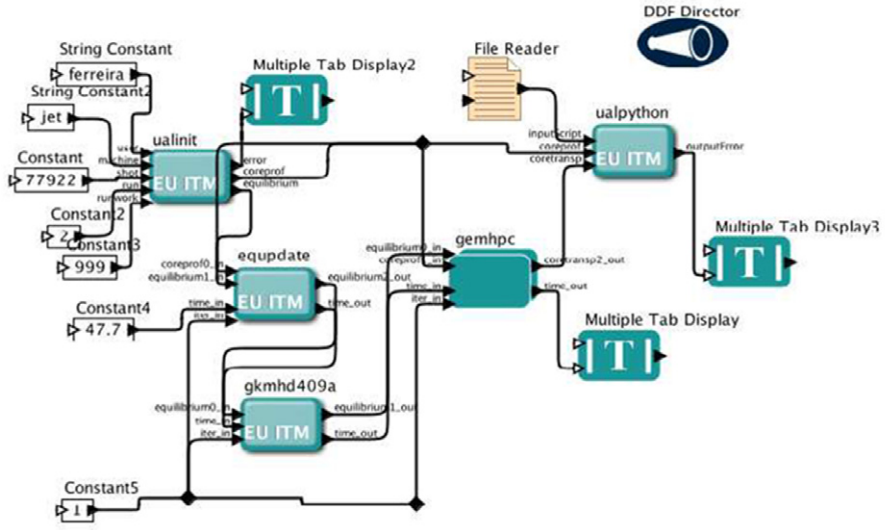

Figure 6. Turbulence workflow: JET shot data are read from the ITM database, the actor EQUPDATE provides equilibrium profile data and the equilibrium boundary, i.e. the separatrix, to actor GKMHD which adds the metrics information; both equilibrium_CPO and coreprof_CPO are input to the turbulence actor GEMHPC which runs in batch GEM gyrofluid flux-tube code in parallel on the HPC-FF.

GEM found a weak-to-stable ITG case at $r_a = 0.6$ (dominant $E \times B$/ion-gradient energetics), with ion heat flux $Q_i < 0.1$ in gyro-Bohm units of $p_e c_s (\rho_s/L_\perp)^2$, whereas it showed stability at 0.4, a very weak ITG case at 0.5 and approximately null growth at 0.7. The delta-FEFI results are quite different: at $r_a = 0.4$ and 0.5, the code crashed apparently due to difficulty with the kinetic ballooning mode, KBM (delta-FEFI has never managed a saturated nonlinear-KBM case) whereas, interestingly, GEM had not found it; the $r_a = 0.7$ case was definitively stable and the $r_a = 0.6$ case produced what can be identified as a strongly trapped-electron enhanced ITG turbulence case, the evidence of which is worth showing.

Figure 7 shows the velocity-space distribution of the contributions of delta-f to the turbulent $E \times B$ fluxes: all of the activity in the electrons and almost all in the ions is in the trapped domain (smaller $v_\parallel$ for finite μB). This is the clearest-possible identifier for a role of trapped electrons despite the ion-dominant energetics and is the basis for the named *trapped-electron assisted ITG turbulence*. It has to be mentioned that this result differs from that provided by GEM, as its gyrofluid model does not include trapped electrons [24].

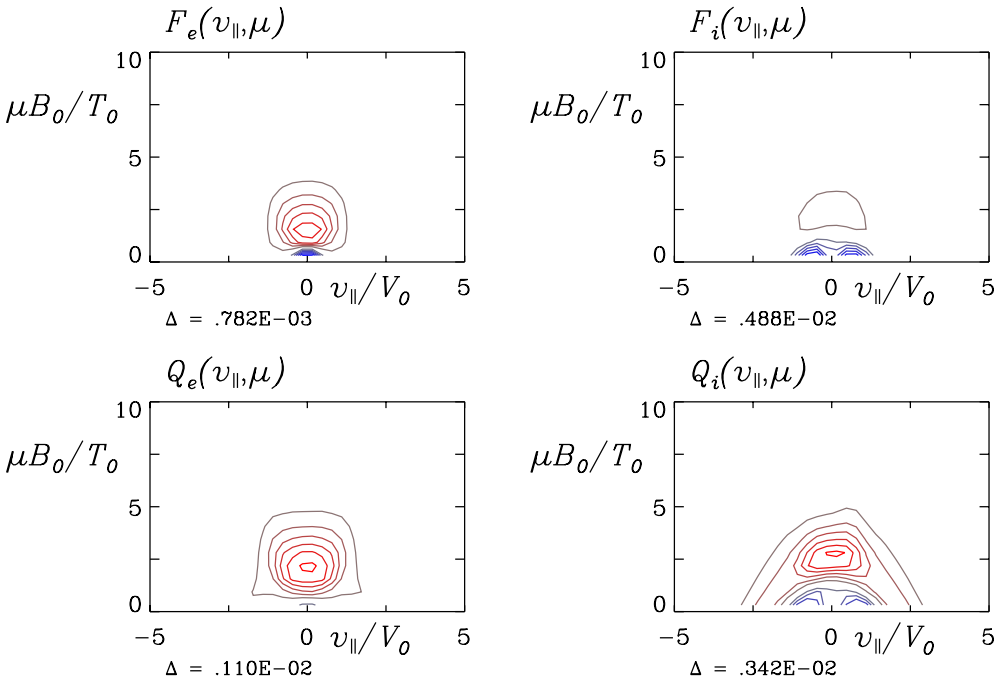

Figure 7. Velocity-space distribution of the $E \times B$ particle (F) and heat fluxes (Q) in the electrons and ions, in a snapshot at the end of the delta-FEFI run. The trapped zone is roughly the 60° cone centred upon the vertical axis where $v_\parallel = 0$. Almost all of the activity in the ions, and essentially all of it in the electrons, is in the trapped zone. These trapped-electron features together with the dominant ITG energetics (not shown) yield the description 'trapped-electron enhanced ITG turbulence'.

In addition to the above simple workflow used for a transparent cross-benchmarking of turbulence modules producing the above physics results, tightly coupled transport workflows were developed, wherein the turbulence module provides heat and particle diffusion coefficients to a transport solver, similar to the strategy used by other models [25–27]. A demonstration of ITM progress towards turbulence-transport workflows with the ETS core transport solver is given in [28]. A parallel effort within a different scientific workflow framework is presented in [29].

Herein, we describe our progress in the coupling with an equilibration model which solves a statistical steady-state equilibrium rather than a time-dependent transport problem. This essentially replaces the time step loop with a convergence loop for the time-independent problem. Nevertheless, the term 'time step' is convenient to the generic transport workflow structure (figure 8).

The main Kepler workflow including the time loop is shown in figure 8, in which the 'workflow core' is a composite actor representing one time/convergence step of the transport workflow.

The sequence of operations of the workflow core, shown in figure 9, is detailed in the following. Here, BPROFS is the transport equilibration model, used in place of the ETS module implemented in [28, 29].

The UALINIT actor is executed only once, at the beginning, to read the input data from the database; it provides the initial coreprof_CPO and equilibrium_CPO at the initial time (for time steps after the first one, the 'coreprof' and 'equilibrium' boxes shown at the top replace the UALINIT actor, representing the previous step's output). At each time step the EQUPDATE actor sets up a new equilibrium_CPO using the pressure and current from the coreprof_CPO and

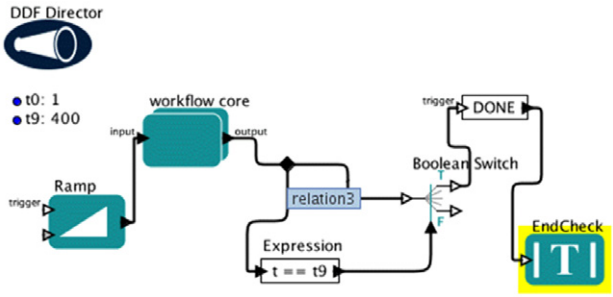

Figure 8. Topmost level Kepler workflow. The actor 'ramp' corresponds to the control of the time loop (it generates integers from 1 to the maximum value of the time step); 'workflow core' is a composite actor representing one time step of the transport workflow.

the last closed flux surface (LCFS) boundary from the equilibrium_CPO. This is then fed into the CHEASE actor to calculate a new, updated equilibrium_CPO. The coreprof_CPO and equilibrium_CPO are then used as inputs for the remaining actors: GEM, which provides a coretransp_CPO, NEOART, a neoclassical transport module providing a neoclassic_CPO, and BSOURCE, a simple analytical source model which provides a coresource_CPO. All of these are fed into the BPROFS actor, which updates the coreprof_CPO according to a simplified profile-equilibration model using running exponential averages [30] of the transport to relax the profiles into a state of transport equilibrium (the aim is not a transport simulation but a procedure to find a steady state). Review information and detailed comparison of relaxation methods can be found in [31].

The only parallel actor is GEMHPC, the first call to which launches a batch job on the HPC-FF, consisting of eight flux

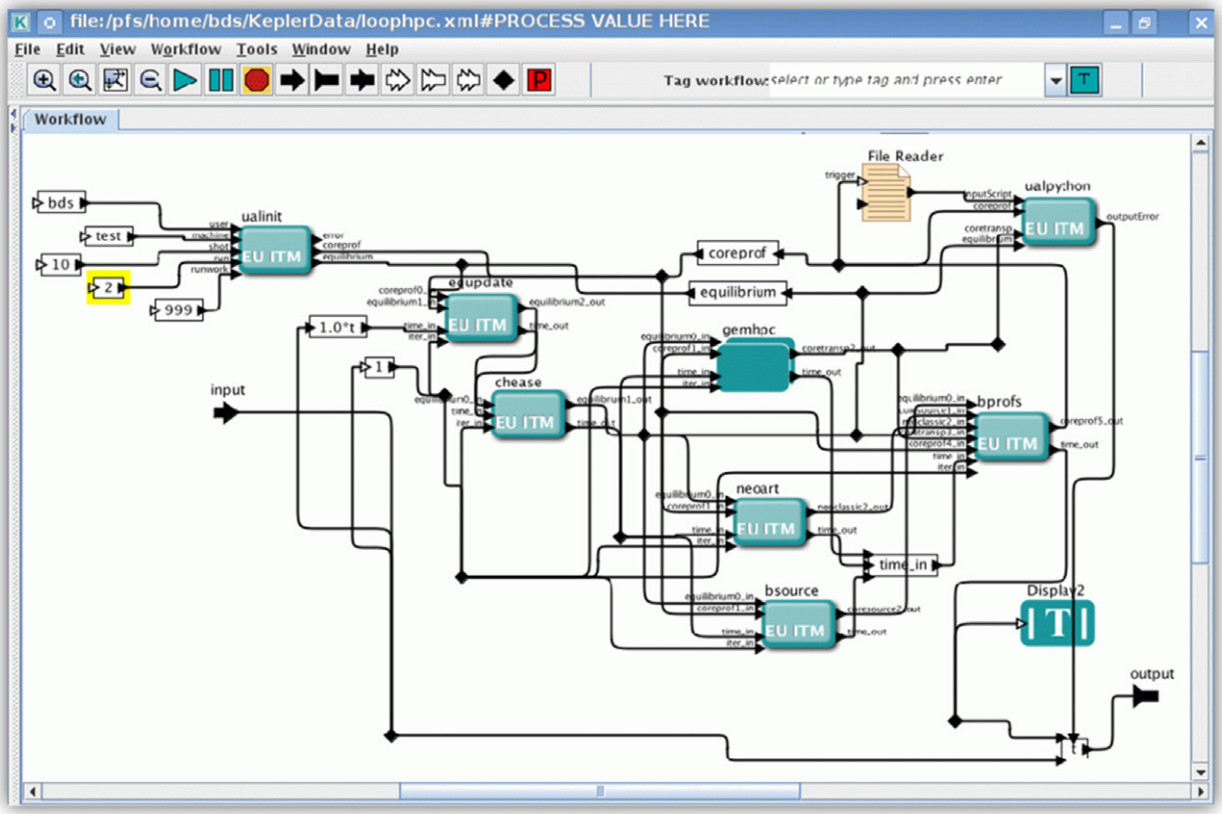

Figure 9. Workflow core, representing the sequence of operations performed during one time step of the transport workflow of figure 8. Detailed description is given in the text.

tubes on the profile, arranged as in the single-run case described above, on 1024 cores. The HPC-FF job runs one segment of $10\tau_{GB}$, returns a coretransp_CPO, and waits. Subsequent loop steps 'fire' the actor again, and it sends a message to the job containing the coreprof_CPO and equilibrium_CPO and instructions to evolve GEM's state for another $10\tau_{GB}$ and return running exponential averages of the turbulent flux profiles to the coretransp_CPO, after which the job waits again and the GEMHPC actor sends the coretransp_CPO on to the BPROFS actor.

The batch job is kept running until it either crashes or accepts a stop signal from the workflow indicating completion—that is, the batch job needs only be submitted and wait in the queue once.

This workflow was applied to a JET-sized case with a model LCFS boundary for ITER (i.e. the R and Z are halved). A power source of 10 MW for each species was specified by the BSOURCE actor. The loop ran for 120 time steps (close to a relaxation time for many cases). The main transport workflow ran on the ITM Gateway cluster, while GEM's batch job ran on the HPC-FF on 1024 cores.

Temperature profile modifications induced by the turbulence coefficients occurred only in the edge, producing a fast profile steepening (the core being marginally stable and turbulence delayed), that eventually crashed the equilibrium reconstruction. Figure 10 shows the outcome of the python visualization actor (see figure 9 top-right), which allows the monitoring of the simulation at each transport time step during the workflow.

This work is in progress, as the workflow scheme is only mature and robust for s–α model cases actually running only GEM and BPROFS by themselves.

3. Verification and validation

The ITM-TF framework is a valuable environment for a rigorous cross-verification of codes describing the same physics processes with different models, since by interchanging those as modules within the same workflow the possible external sources of discrepancy are minimized. Considering the fundamentals of an integrated transport simulator, it is essential to address the benchmarking of the equilibrium and core profile evolution solvers as well as the transport and turbulence or heating and current drive modules. Equilibrium codes went through benchmarking both within the ETS workflow [25] and independently, whereas cross-verification of turbulence and MHD codes is ongoing on specified test cases within dedicated workflows. In this section, the benchmarking of standalone electron cyclotron heating and current drive codes on an ITER scenario and the ETS validation against existing integrated modelling transport codes on a JET hybrid discharge are presented. It has to be mentioned that the ETS was previously extensively verified [6, 32]. The very good agreement achieved for the simulated quantities and applied modules lays the foundations for the use of ETS for both predictive and interpretative runs on present devices and ITER, in a variety of scenarios.

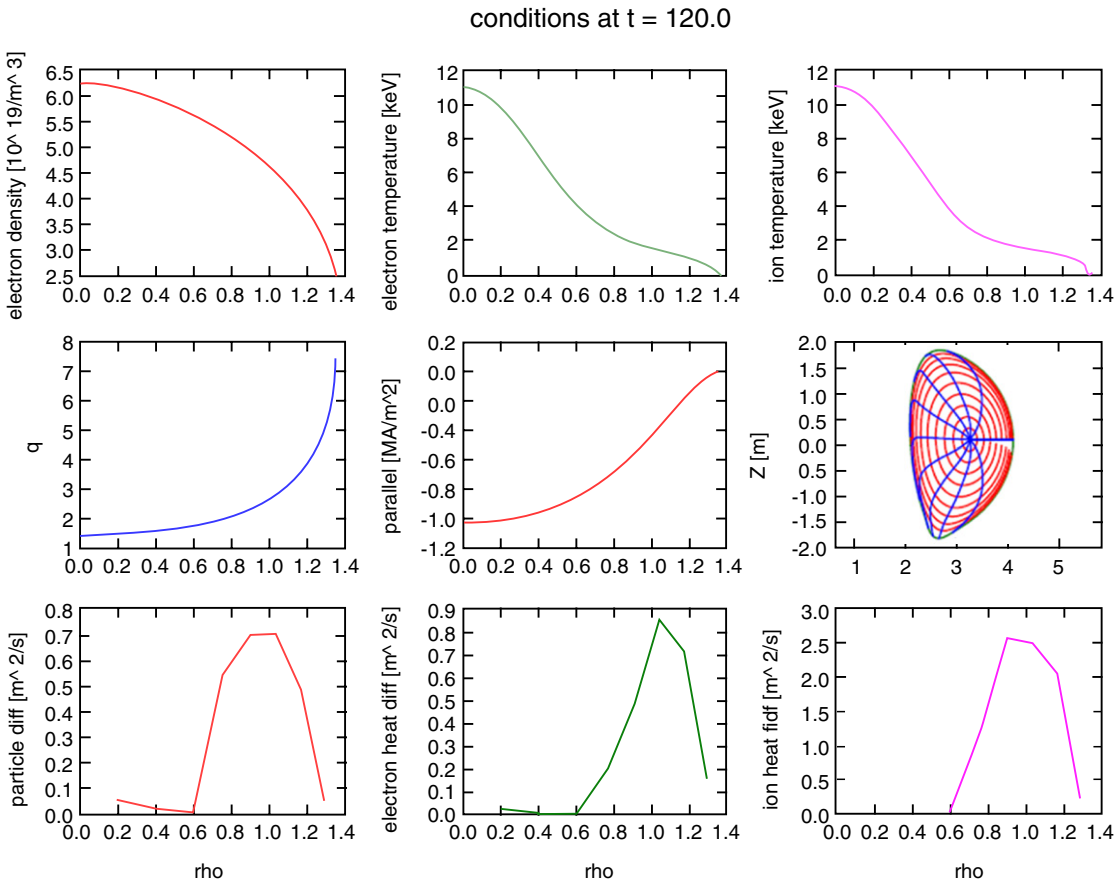

Figure 10. Instantaneous profiles and magnetic equilibrium after 120 time steps of the coupled turbulence-transport workflow (outcome of the ualpython visualization actor, see figure 9). Top and middle row, from left to right: electron density (not evolved), ion and electron temperature, safety factor profile, parallel current (coreprof_CPO); magnetic equilibrium (equilibrium CPO). Bottom row: particle, electron heat and ion heat diffusivities (coretransp_CPO).

Table 1. Launching conditions used in the benchmark. The poloidal and toroidal launching angles are defined as $\alpha = \tan^{-1}(k_{0,z}/k_{0,R})$ and $\beta = \sin^{-1}(k_{0,\phi}/k_0)$, where $(k_{0,R}, k_{0,\phi}, k_{0,z})$ are the cylindrical wave vector components of the launched wave. The beam has a Gaussian profile, with waist w_0 at a distance d from the launching point. The model considered here corresponds to simple astigmatism, when the spot ellipse and the curvature ellipse are rotated by the same angle.

Case	R_m (m)	z_m (m)	α (°)	β (°)	w_0 (m)	d (m)
EL25	9.27	0.62	0	25	0.030	0.00
EL40	9.27	0.62	0	40	0.030	0.00
UL	6.90	4.18	48	18	0.021	1.62

3.1. Benchmarking of electron cyclotron heating and current drive codes on an ITER scenario

A benchmark among five European EC beam/ray-tracing codes (C3PO [33], GRAY [34], TORAY-FOM [35], TORBEAM [36], TRAVIS [37]) has been successfully performed [6] within the ITM framework for a standard inductive H-mode ITER scenario ('Scenario 2') at the end of burn phase, for three different launching conditions both from the EL and UL, see table 1.

The three cases have been selected to cover different geometries and physics of interest in ITER: injection of divergent beams from the EL, either absorbed in the core at quite large T_e (EL25), or characterized by quite long trajectories and large $N_\parallel$ (EL40), and of a highly focused beam that drives the EC current in a narrow channel (UL). The frequency of the launched beam is 170 GHz and the input power is 1 MW. Figure 11 represents the used plasma equilibrium and beam trajectories.

Among these codes, GRAY, TORBEAM and TORAY-FOM had also participated in the benchmark exercise in [38] that was run on the same ITER scenario using only a divergent Gaussian beam launched from the UL. Since then, the codes have been modified and updated to include different physics modules as, e.g., the current drive model.

The fact that the codes run in the same ITM workflow simplifies the verification and, at the same time, guarantees a more detailed check of the various steps of the benchmark. Note that any module can be switched into an ETS time-dependent simulation since they share the same interface through CPOs.

The steps taken in the benchmarking study consisted in (i) an extensive check of matching between ITM's and all codes' coordinate and sign conventions as well as physical quantities definitions, to ensure that the input and output data were correctly interpreted and written by the codes; (ii) a comparison among the computed wave trajectories, with particular consideration of the vacuum–plasma transition;

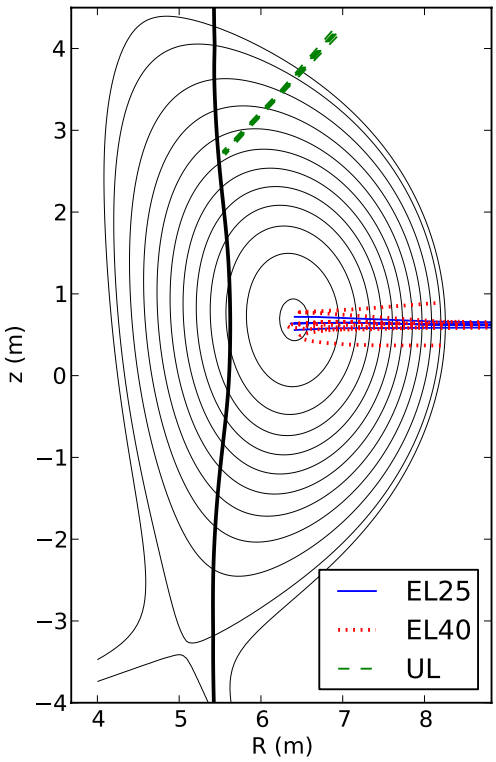

Figure 11. Magnetic equilibrium of the used ITER scenario 2. The cold resonance at 170 GHz is shown (thick vertical line), with the beam-tracing computed by the GRAY code for the three considered launching (from [6] with kind permission of the European Physical Journal, EPJ).

(iii) a comparison of the power absorption and current drive results.

Good agreement was found, with differences in total current $|\delta I_{CD}/I_{CD}| < 15\%$, and with peak values of power density dP/dV and driven current density typically matching within 10%, and the position of the profiles matching within $\delta\rho \sim 0.02$ in normalized radius units (figure 12, partially taken from [6]).

Small discrepancies can be ascribed to the different models used for wave propagation and absorption and current drive. The EL25 case considers a beam trajectory passing very close to the magnetic axis, where small differences in the interpolation of the equilibrium data or in the beam trajectory may result in relevant changes in the flux averaged power density dP/dV value, as can be seen in the TORAY-FOM result at $\rho < 0.15$. In the EL40 case, Doppler broadening dominates the effect of finite beam size in the determination of profile width, and all the codes here agree very well. Some differences can also be appreciated among the results for the UL case, mainly because this case is more demanding in terms of spatial resolution required to reconstruct the actual shape of the absorption profile, since the focused beam considered here produces a much narrower (full width at $1/e$ $\Delta\rho \sim 0.015$) absorption profile than those obtained with the EL. In this respect, the difference of C3PO/LUKE here results from a coarser grid considered in the calculation.

In the UL case, despite the focused beam, the profiles are reasonably well reconstructed also by ray-tracing codes, giving results comparable to those obtained by the codes which account for diffraction effects. The large edge density gradient, and long path from boundary to absorption region, amplifies the impact of edge refraction on beam propagation. Nonetheless, the influence of the observed discrepancies on computed power and current density profiles is still moderate. Only in the case of a strongly focused beam, as in the UL case, may the uncertainty approach the profile width. A deeper analysis of the discrepancies among the different codes and underlying models used for wave propagation, absorption and current drive is ongoing and will be presented in a following paper.

3.2. ETS validation

A rigorous benchmarking of the ETS against ASTRA [31] and CRONOS [40] integrated modelling transport codes was performed by using the parameters of JET hybrid discharge #77922 with current overshoot, $B_{tor} = 2.3$ T, $I_{pl} = 1.7$ MA, high triangularity (0.38), 18 MW of NBI, $n_l = 4.8 \times 10^{19}$ m^{-3}, $\beta_N = 2.8$. The equilibrium was reconstructed in CRONOS and ASTRA with the solvers available within those, respectively HELENA [12] and the three moment equilibrium module EMEQ [35]; the latter module was also implemented in the ETS. It is worth mentioning that the flexibility of the ETS advantageously allows for an easy integration of additional equilibrium codes other than those it already supports. Evidently, ideally a rigorous benchmark should have been required for all codes to use the same equilibrium reconstruction.

Self-consistent evolution of electron and ion temperatures, current diffusion and equilibrium was simulated. Spitzer resistivity was used for the current transport, and the heat transport coefficients were provided by a Bohm–gyro-Bohm transport model. Neoclassical heat transport was not included. The simulations were performed with a fixed electron density profile measured at 7.7 s of shot #77922. Gaussian heating and current drive profiles (centred at $\rho = 0$, half-width $\Delta\rho = 0.3$), with a total heating power $P_{tot} = 18$ MW, distributed 70/30 between ions and electrons, were used in all codes. A beam-driven current $I_{ni} = 0.12$ MA was imposed in all simulations while the bootstrap current was neglected. With these assumptions, the simulations were performed for 40 s reaching a steady-state solution.

Satisfactory agreement was obtained on the temperatures and q-profile simulated by the three codes as well as on the computed thermal diffusivities (figure 13) [6]. The slight differences in profiles can be attributed to the different equilibrium solvers used within the compared integrated modelling transport codes.

4. Tightly coupled workflows developed by ITM-TF

4.1. Core–edge coupling

Coupling codes, besides the complexity of dealing with separate codes eventually presenting mixed-language programming, which is indeed overcome by the ITM-TF approach, introduces a number of issues to be dealt with: disparity in time-scales, different physics assumptions and scheduling the interaction between the coupled codes.

The core–edge coupled system does introduce a disparity in time-scales, with a characteristic time-scale for the core

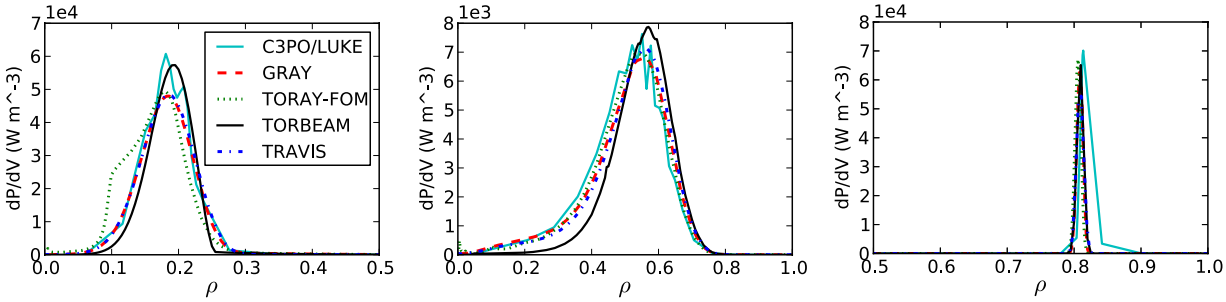

Figure 12. Power density profiles computed for the launching conditions of table 1: EL25 (left), EL40 (centre) ([6] with kind permission of the European Physical Journal (EPJ)) and UL (right).

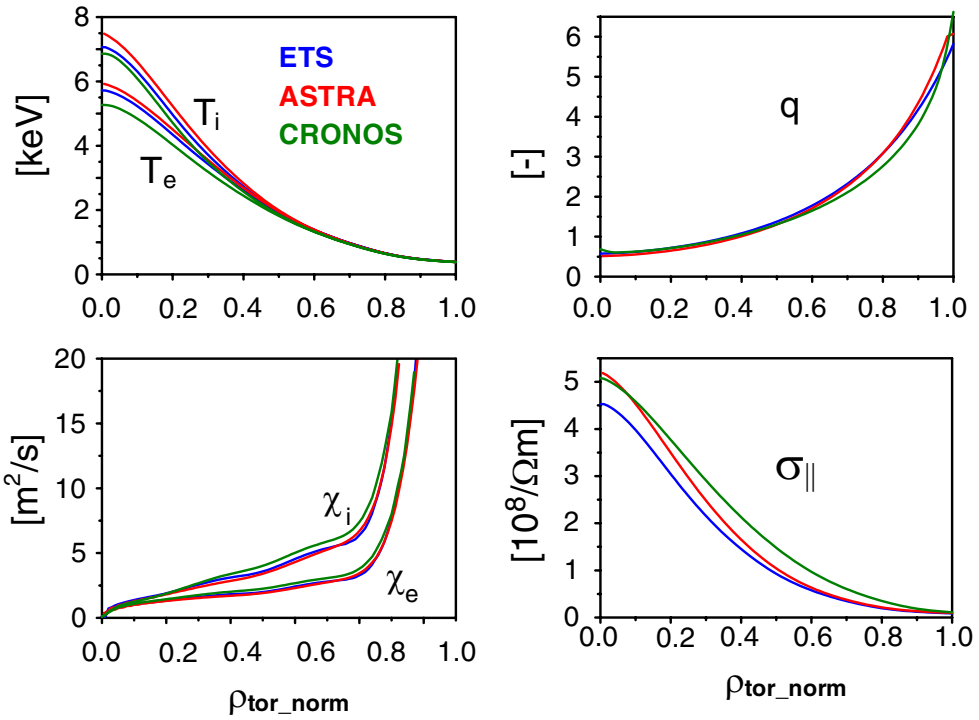

Figure 13. Benchmarking between ETS (blue), ASTRA (red) and CRONOS (green) integrated modelling transport codes for the conditions of JET hybrid discharge #77922. Profiles of ion and electron temperature, safety factor, ion and electron heat diffusivity and parallel resistivity are plotted in the steady state, after 40 s of evolution.

being an energy confinement time or longer (seconds), whereas the scrape-off layer (SOL) typically has a time-scale of milliseconds with some phenomena being even faster. Another disparity is the computational complexity: transport solvers for the core are typically one-dimensional (1D) (radial) codes solving a set of reaction–convection–diffusion equations evolving the density, toroidal momentum and energies for the species considered; edge transport solvers are typically a 2D (radial and poloidal with toroidal symmetry assumed) or 3D code solving for the density, parallel momentum and energies for the species considered and are thus considerably more expensive computationally. Moreover, impurities in the core are often split off from the main ion species and only the density equations are solved for the various impurity charge states. The coupling effort is significantly simplified in the case where one is interested in finding a consistent steady-state solution between the core and edge codes, which is the problem addressed here.

Three approaches for core–edge coupling can be used, as described in [41]: mediated, where the edge codes are used to provide boundary conditions (BCs) for the core codes on the basis of fitting coefficients to the results of a number of edge runs; direct where the edge and core codes are directly coupled; and avoided where the edge code is extended all the way to the centre of the plasma. There have been several previous independent core–edge coupling projects: [42] describes the coupling of the core code Corsica to the edge code UEDGE; [44] describes the coupling of JETTO, EDGE2D and SANCO. A very similar approach describing the coupling of SOLPS and ASTRA is described in [45, 46]. An alternative approach is that described in [41], where the coupling issue was avoided by extending the SOLPS grid to the centre of the machine. In [47], scaling laws were derived on the basis of SOLPS simulations and then used for core simulations.

Here we present the direct coupling of an edge and a core transport code via a Fortran workflow using the ITM-TF infrastructure (i.e. CPOs) for the particular case of steady state and multiple impurities [7]. The edge 2D transport code (SOLPS) [43] was coupled with the 1D core main plasma transport code ETS [4] including a core impurity

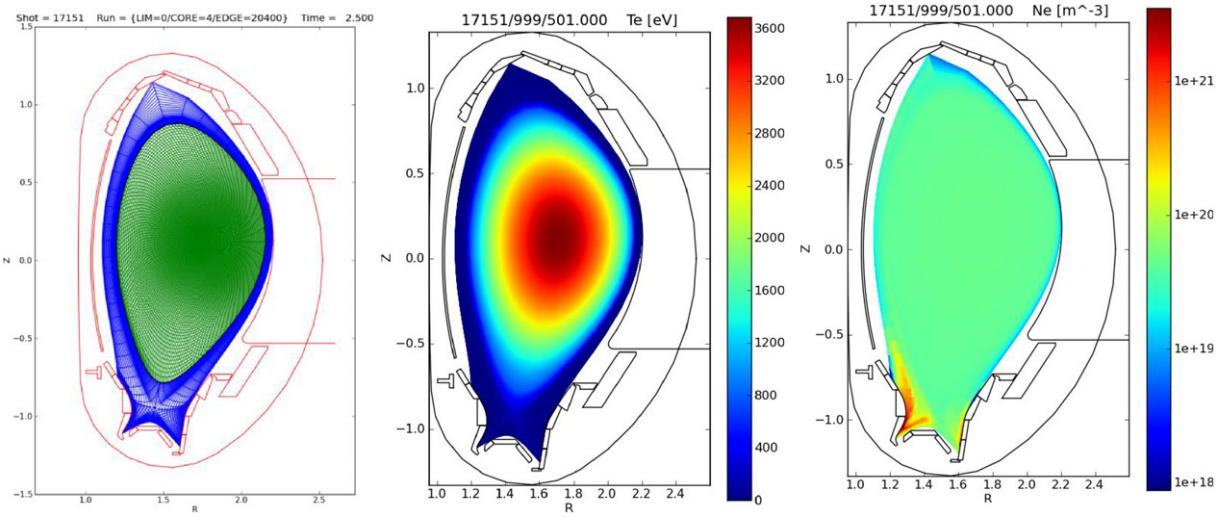

Figure 14. Left, the combined core and edge grids for ASDEX Upgrade shot 17151. The boundary surface separating the calculation domains between the core and edge codes is clearly visible. T_e (centre) and n_e (right) for the final state of the D + He + C + Ar + Ne case [7]. All plot data are derived from CPOs.

transport code, developed within the framework of the ITM-TF. In this work a Fortran version of the ETS workflow was used, including the equilibrium code HELENA [12] and simple models for particle and energy sources as well as transport coefficients. ASDEX Upgrade shot #17151 equilibrium at 2.5 s was imported into equilibrium and limiter CPOs. These CPOs enter the HELENA code providing equilibrium to the core transport code and were used to create the SOLPS grid (figure 14, left).

The location of the transition surface between the core and edge code was chosen at 95% of the normalized poloidal flux for the case shown below, as this is the usual stopping point for 1D core transport and equilibrium codes, making comparisons with existing results easier. Standalone SOLPS runs, pertaining to the demonstration case chosen here, showed the poloidal asymmetries to have averaged out at this depth. More generally, at least for H-mode cases, any poloidal asymmetries introduced by the edge physics do not penetrate past the pedestal, as demonstrated by a comparison study between SOLPS and ONETWO codes [48].

The two codes, ETS and SOLPS, were then called alternately and individually run until converged, with information about the BCs transferred from one to the other, until convergence of the workflow is obtained.

For the most complicated test case, SOLPS treated all of the charge states of D, He, C, Ar and Ne (including the neutrals), a total of 42. The ETS treated D^+ and He^{2+} as main ions, and the core impurity code treated the individual charge states of C, Ar and Ne. The core codes did not, in this case, treat the neutrals. Electron and ion energy fluxes as well as D^+ and He^{2+} particle fluxes are passed from the core to the edge code. Values of density and ion temperature on the boundary are passed from SOLPS to the ETS and densities of C, Ar and Ne charge states to the core impurity code. SOLPS used a zero-flux BC for neutrals, all of the charge states of C, Ar, Ne and for He^{1+}. The fluxes are implemented in the edge code via a feedback loop, which varies a constant density on the boundary so that the desired flux is obtained; this avoids

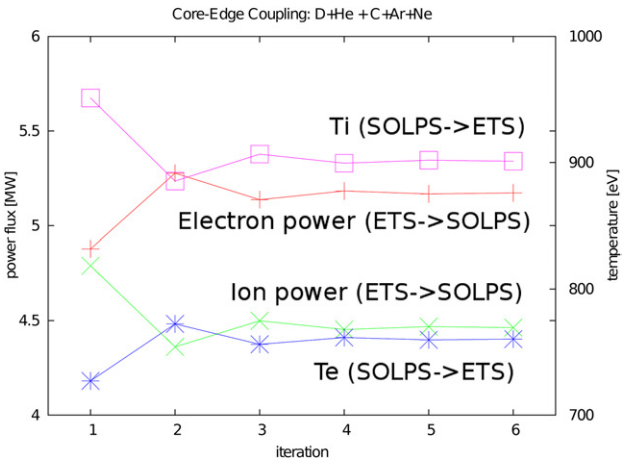

Figure 15. Core–edge workflow convergence of the boundary powers and temperatures with respect to iteration number.

the problem of having a poloidal variation of the density on the boundary which would then need to be averaged in some way before being passed back to the core code (if the flux BC is implemented directly using a prescription of equal flux per unit area, a strong poloidal variation for some of the impurity densities can be found).

Convergence was obtained with five iterations, as is shown in figure 15. To illustrate the convergence figure 16 shows the time evolution of the electron and ion temperature profiles at the outer midplane. The results for the steady-state electron temperature and density are shown in figure 14 (centre, right); densities for C, Ne and Ar charge states in figure 17.

Figure 18 shows a visualization of the core plasma temperature (simulated in 1D by the ETS core transport solver), the edge plasma temperature (simulated in 2D by SOLPS) together with the 3D wall.

Recently SOLPS has been modified so that it can accommodate time-dependent BCs. The coupling has thus been automated as follows: the ETS Fortran workflow calls SOLPS just after the convergence loop, SOLPS receives as input the

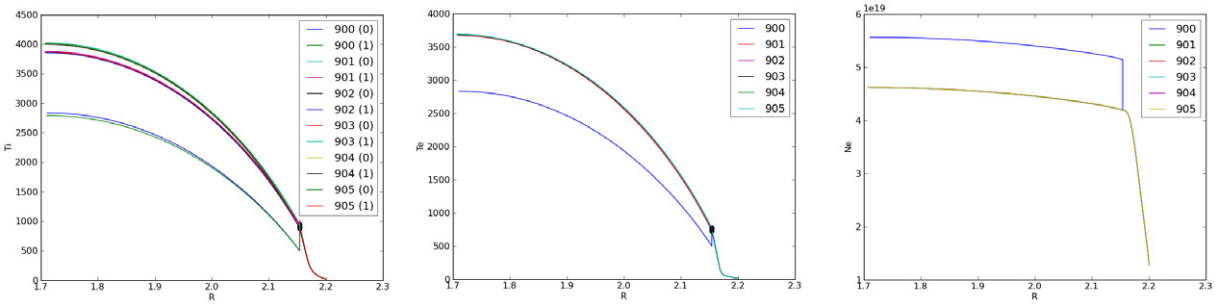

Figure 16. Time evolution of the ion and electron temperature (left, centre, respectively) and electron density (right) profiles at the outer midplane.

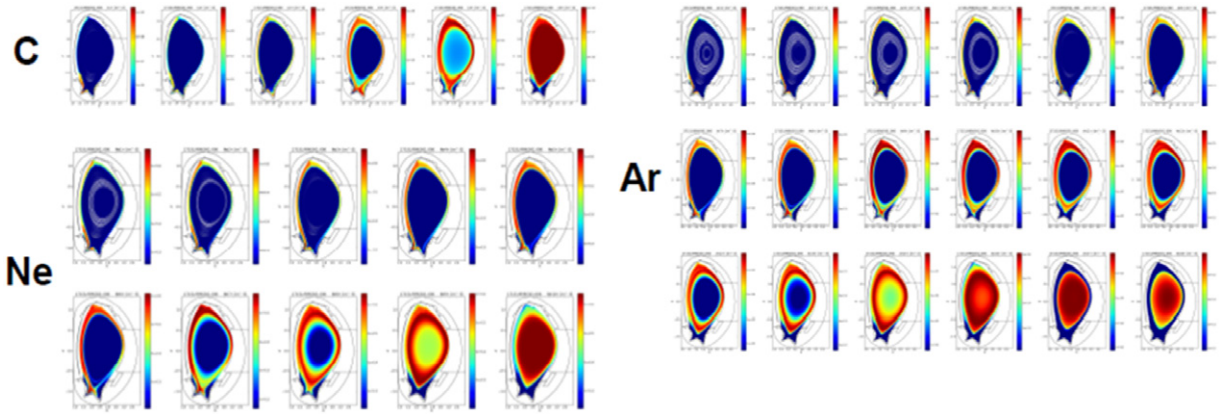

Figure 17. Density plots in the steady state for all the charges states of C Ne (left, top bottom) and Ar (right) ([7] with kind permission of the European Physical Journal (EPJ)).

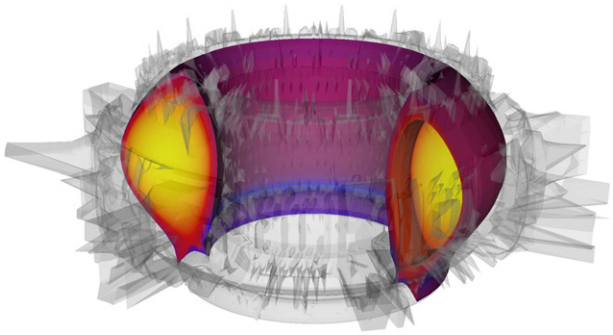

Figure 18. Visualization of the core–edge coupled simulation results: T_e calculated in the core with the ETS, in the edge with SOLPS, within the 3D defeatured first wall of ASDEX Upgrade obtained using a ray-tracing rasterization and smoothing ([49] with kind permission of the European Physical Journal (EPJ)). All data are stored in CPOs and plot with VisIT.

necessary BCs from the core CPOs, runs for one or more time-steps and calculates new core CPOs with new BCs based on the edge results, then the ETS continues with a new time step. The advantage of the new automated coupling scheme goes beyond just speeding up the calculation. The initial approach relied on a manual coupling which required the user not to make mistakes in the coupling procedure, and was also limited to steady-state scenarios. For impurities, there was also a limitation in that only cases with net zero flux could be implemented, and this was then done by charge state rather than the more physically correct summation over charge states. The new approach does not require manual intervention during the run, is not limited to steady-state simulations, and removes the issue related to zero-flux BCs for impurities at the coupling interface.

4.2. Transport simulations including NTMs

A module which simulates the time behaviour of NTMs [50] can also be integrated in the ETS workflows. Here we present a demonstration of the ETS workflow including the NTM module reproducing the effect of NTMs on transport evolution.

NTMs are resistive instabilities breaking the flux surfaces into magnetic islands at the rational surfaces $q = m/n$ (i.e. located at radius r_s). The modes are destabilized by a loss of bootstrap current proportional to the plasma pressure. The simulated modes grow starting from a specified onset time, up to the saturated state. Their growth affects the local electron and ion temperature and density by changing the perpendicular transport coefficients around the mode location. The transport is modified by the NTM module, which adds a Gaussian perturbation of given amplitude and width to the unperturbed transport coefficients [51]. The width is calculated self-consistently by solving the modified Rutherford equation at each simulation time step, with parameters as in [52], evaluated in toroidal geometry, except for assuming $\Delta' = -m/r_s$ (effective Δ' in the case of a large island, Δ' being the usual tearing parameter due to the perturbation of the equilibrium current). This approach enables the reproduction of density and temperature profiles very close to the experimental ones.

Figure 19 presents the temporal evolution of the electron temperature and total perpendicular heat diffusivity profiles,

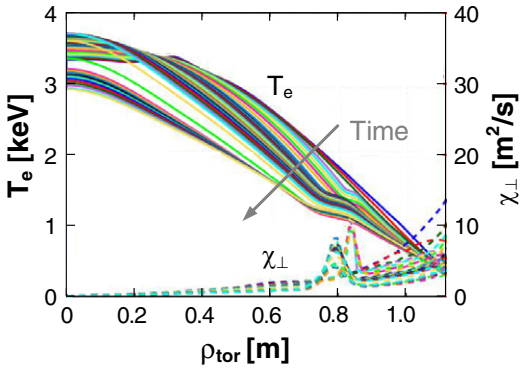

Figure 19. Modification of the heat transport coefficient by NTMs, assumed to be located at ρ_{tor} ~0.8, and its effect on the electron temperature profile.

during an ETS-NTM simulation performed for typical JET H-mode plasma conditions. The effect on these profiles of an $m/n = 2/1$ magnetic island, assumed to be located at $\rho_{tor} \sim 0.8$, is shown. NCLASS [53] is used within the ETS [33] to calculate the resistivity and bootstrap current. The equilibrium is evolving and parameters in the modified Rutherford equation are recalculated according to [52]. The q-profile is thus slightly evolving, which leads to a radial displacement of the $q = m/n$ surface and therefore a change of location of the increased transport due to the island (peak on χ). The increase in radial transport due to the presence of the magnetic island leads to a flattening of the temperature profile around the 2/1 surface. The mode is predicted to grow up to a saturated island of 8 cm width on a resistive time scale of about 150 ms; this leads to a 16% drop in the stored energy. Validation against experimental data will be the next step and requires first a validation of the transport model.

4.3. FBE coupled to transport

A key feature for a tokamak simulator is the inclusion of the PF system, i.e. the PF coils and their power supplies as well as passive conducting structures. This allows including important operational limits and real-time magnetic control issues in the design of scenarios [54]. The ETS now has such a capability thanks to the inclusion of a FBE solver, at present, either CEDRES++ [55] or FREEBIE [56, 57]. The circuit equations for the PF coils and passive structures are embedded in the FBE code. A switch in the ETS workflow allows one to select one of these solvers in place of a fixed-boundary one. Coupling a FBE code to a 1D transport solver is not trivial [55, 58], therefore the coupling algorithm is detailed in the following.

In order to ensure consistency between the equilibrium and the profiles, the FBE–transport coupling scheme relies on a convergence loop performed at each time step, which is represented in figure 20.

A time step starts with one transport step $t \rightarrow t+\mathrm{d}t$. In addition to the coreprof_CPO, which contains the profiles at time t, the transport solver needs as an input an equilibrium_CPO, for two reasons: the transport equations involve metric coefficients which depend on the equilibrium (e.g. such as the flux surface average of $1/R$) and the flux diffusion equation (FDE) needs a BC at the edge, which has to be provided by the equilibrium. This BC is a central point

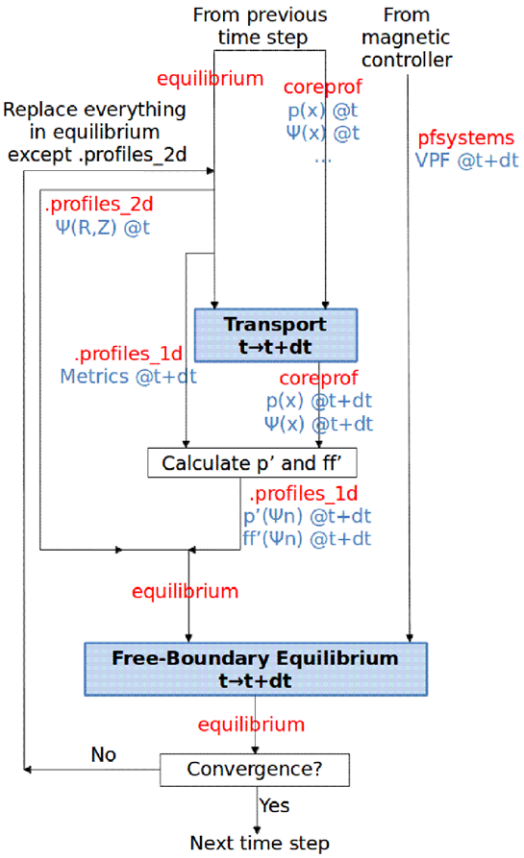

Figure 20. The convergence loop performed at each time step in the coupled FBE–ETS workflow. The labels next to the arrows comprise, in red, the names of the transferred CPOs or CPO fields and in blue, the names of the main variables of interest which are contained within them.

of the FBE–transport coupling. It has to guarantee in particular the consistency of the poloidal flux Ψ, which is evolved both by the FBE solver (Ψ_{eq}) and the transport solver (Ψ_{tr}). The natural choice for the BC of the FDE, $\Psi_{tr,x=1} = \Psi_{eq,x\mathrm{b}}$, where $\Psi_{eq,x\mathrm{b}}$ is the poloidal flux at the plasma boundary provided by the equilibrium solver at the previous iteration of the convergence loop (or at the previous time step for the first iteration) and x is the normalized square root of the toroidal flux, tends to generate unphysical current sheets at the edge driven entirely by numerical noise on $\Psi_{eq,x\mathrm{b}}$. Therefore a BC of the type

$$\left.\frac{\mathrm{d}\psi}{\mathrm{d}x}\right|_{x=1} = \frac{-2\pi\mu_0 I_p^*}{C_2(x=1)}$$

is used, where C_2 is a metric coefficient (see [54]). In order to guarantee consistency the following expression is used:

$$I_p^* = I_{p,eq}\left[1 + \tanh\left(\frac{\psi_{tr,x\mathrm{b}} - \psi_{eq,x\mathrm{b}}}{\psi_{eq,x0} - \psi_{eq,x\mathrm{b}}}\right)\right], \qquad (1)$$

where subscripts $x0$ and $x\mathrm{b}$ indicate the magnetic axis and the plasma boundary, respectively.

This expression is based on the following considerations. A correction term ΔI_p is added $I_p^* = I_{p,eq} + \Delta I_p$ ($I_{p,eq}$ indicating the plasma current found by the FBE solver) aiming at ensuring the consistency between Ψ_{eq} and Ψ_{tr}, which otherwise diverge one from the other. As a measure of the distance between those

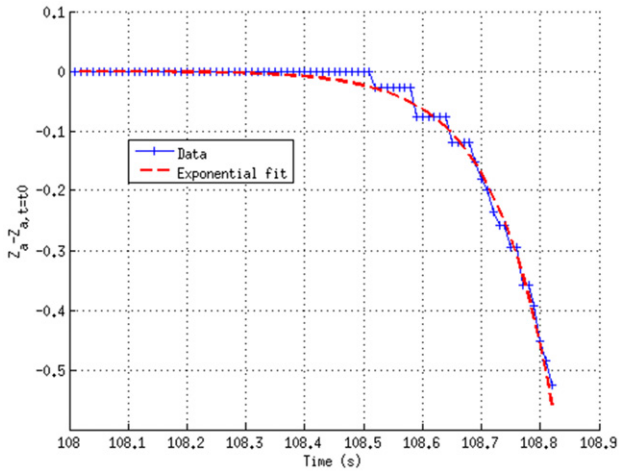

Figure 21. Evolution of the magnetic axis vertical position with respect to its initial value.

$\Delta\Psi_{xb} \equiv \Psi_{tr,xb} - \Psi_{eq,xb}$ is used; therefore, $\Delta I_p = \Delta\Psi_{xb}/L_p$, where L_p has the dimension of an inductance and is chosen as $L_p = (\Psi_{eq,x0} - \Psi_{eq,xb})/I_{p,eq}$. Finally, a tanh function is used in order to provide a saturation of the correction term in case the initial difference $\Delta\Psi_{xb}$ would be large. Evidently, at the end of the convergence process, $\Delta\Psi_{xb}$ is small so that the tanh does not make a difference.

After the transport time step, in order to prepare the input data for the FBE time step, an intermediate step is necessary to calculate the $p'(\Psi_n)$ and $ff'(\Psi_n)$ profiles from the new p and Ψ profiles and the metric coefficients. Ψ_n is the normalized Ψ equal to 0 on the magnetic axis and 1 at the plasma boundary, $f = RB_\varphi$ is the diamagnetic function and the $'$ denotes the derivative with respect to Ψ. First, p' is calculated as $p' = (\mathrm{d}p/\mathrm{d}x)/(\mathrm{d}\psi/\mathrm{d}x)$, ff' is then obtained from the averaged Grad–Shafranov equation [54]. An FBE step $t \rightarrow t+\mathrm{d}t$ can then be made, using as inputs the p' and ff' profiles as well as the voltages in the PF coils (at present those are prescribed but will eventually be provided by a magnetic controller). The FBE time step calculates a new equilibrium at $t+\mathrm{d}t$, including new metrics and a new I_p. These are then injected back into the transport solver, and the whole process is repeated.

The convergence criterion ensures that the difference in the averaged current density $j_{av} \equiv \langle j_\varphi/R\rangle/\langle 1/R\rangle$ (where brackets denote a flux surface average) and $\Psi_{eq,xb}$ between two iterations is smaller than a given tolerance.

It has to be noted that this algorithm works for both limited and diverted plasmas. As a demonstration of the coupled FBE–ETS workflow we present here a simulation of a VDE in ITER. The initial plasma has $I_p = 11.8$ MA, an elongation $\kappa = 1.49$, and is limited on the high-field side (HFS). PF voltages are set to 0. Figure 21 shows the behaviour of the magnetic axis vertical position Z_a. As expected, it has an exponential behaviour. The time constant is $\tau_{VDE} = 102$ ms, which is typical of ITER [54]. In figure 22, the toroidal current density $j_\varphi(R,Z)$ is shown at two times in the simulation: 108.50 and 108.82 s (the simulation starts arbitrarily at $t = 108$ s, as shown in figure 21). It can be seen that as the plasma moves down, negative currents are induced in the passive structures (as one may expect), in particular the triangular support (small oblique plate on the LFS) and, to a smaller extent, the divertor inboard rail (small vertical plate on the HFS) and the lower part of the vacuum vessel shells (mostly the inner one). Interestingly, a positive current sheet grows at the edge of the plasma towards the end of the simulation. This has to be analysed in detail but it is likely a consequence of the growth of the negative currents in the passive structures. We note that the global current diffusion (L/R) time of the plasma here is of several thousands of seconds whereas the local time at the plasma edge is of the order of a few seconds, on the spatial scale of the observed current sheet, consistently with the simulation. As for the previous section case, the plasma resistivity is calculated with NCLASS within the ETS. There is no bootstrap current spike at the edge however, because this plasma has L-mode-like profiles.

Subsequent to this first demonstration of the FBE–ETS workflow, a cross-benchmarking of the FBE codes (CEDRES++ and FREEBIE) within the above detailed workflow has been started; possibly, optimizations of the algorithm and benchmarks with existing similar efforts are foreseen as well.

The main following step is the implementation of a feedback controller to allow for scenario simulations. Preliminary work has already been performed in this direction with the inclusion of a controller actor produced from a TCV Simulink controller in a CEDRES++ workflow (without coupling to the ETS, hence using prescribed p' and ff' profiles), which allowed reproducing a 'yo–yo' TCV discharge, i.e. the plasma is moved up and down the vessel by the magnetic controller.

5. Synthetic diagnostics integration

The ongoing efforts on synthetic diagnostic integration in the ITM-TF platform focus on reflectometry, neutron and NPA diagnostics and spectral MSE.

A full-wave 3D code (ERC3d) valid for both O and X-mode polarizations has been developed, ported and tested on the ITM platform and work is under way to enhance the kernel to cope with high levels of turbulence and high injection angles (Doppler reflectometry operation). A generic framework for neutron synthetic diagnostics has been integrated which is composed of three different modules: calculation of the effective solid angle of the detector from small plasma volumes (LINE21 code); a Directional RElativistic Spectrum Simulator (DRESS) to derive the energy spectra and source rates of particles created in fusion reactions emitted in a specified direction and a diagnostic response function. Integration of JET neutron camera setup is ongoing. The integration of NPA diagnostics in the ITM platform was also carried out using modules of the ASCOT code package [59] and calculating the fraction of the tokamak chamber and born neutrals (with a given pitch velocity) that are within the sight of the NPA collimator. A spectral MSE forward model [60] that calculates the emissivity for each MSE channel and the resultant radiance Balmer-alpha MSE spectra as well as the charge exchange of the plasma with the beam has been integrated. Full, half and third beam energy components are considered and a collisional–radiative beam–plasma model is used to determine the coupled densities of charged states along

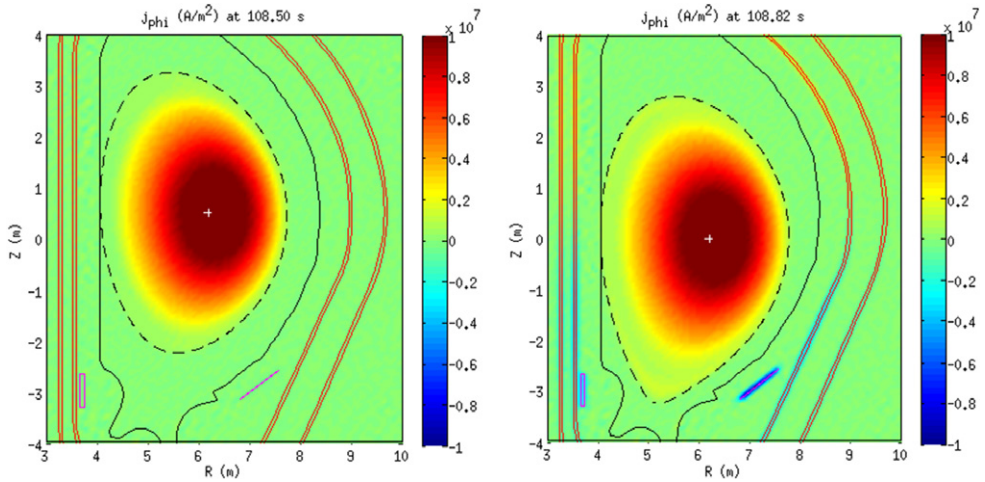

Figure 22. Snapshots of the toroidal current density j_φ at $t = 108.50$ s (left) and 108.82 s (right). The white cross indicates the magnetic axis, the black dashed line the separatrix, the black full line the first wall, and the red and magenta lines the vacuum vessel shells and passive structures, respectively (divertor inboard rail and triangular support).

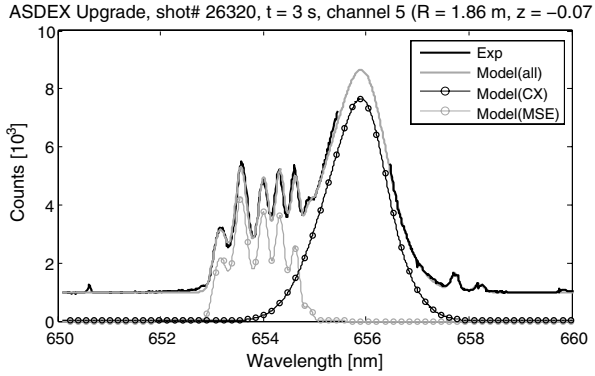

Figure 23. MSE emissivity wavelength spectra for ASDEX Upgrade shot #26320. The contribution from half and third beam energy components, beam divergence and unshifted Dα emission are shown. An offset of ~1000 counts is added to the MSE+CX synthetic counts to account for the characteristic background level of the measured signal by the CCD.

the diagnostic neutral beam path. Preliminary results on the MSE synthetic diagnostic validation on ASDEX Upgrade data (shot #26320) are presented in figure 23, showing the simulated and experimental emissivities.

6. Conclusions

The EU ITM-TF standardized, modular and flexible integrated modelling framework allows building complex workflows for physics application and is a valuable environment to benchmark codes describing similar physics processes with different model sophistication, by interchanging those as modules within the same workflow. Moreover, both the orchestration engine Kepler, and the ITM-TF developments performed in collaboration with the FP7 EUFORIA project [16] and the HLST[17], allow to run workflows or only part of those (the main workflow residing on the central ITM Gateway cluster) on GRID or HPC-FF, thus rendering possible highly computationally demanding calculations.

The first application of the ITM-TF simulation chain coupling equilibrium reconstruction, refinement and linear MHD stability modules addressed edge stability of ASDEX Upgrade ELMy H-Mode and ITER hybrid scenario. Turbulence code interpretative runs starting from given experimental profiles of a JET hybrid discharge, challenging case near to stability threshold conditions, were performed with two different electromagnetic codes, a gyrofluid and a gyrokinetic one, within an ITM workflow. Only one radial position ($r_a = 0.6$ in normalized radius) was found unstable in the gyrokinetic run, highlighting trapped-electron assisted ITG turbulence characteristics. A benchmark among EC beam/ray-tracing codes for a standard inductive H-mode ITER scenario for three different launching conditions, showed good agreement of the five EU codes even in the more demanding test cases, like central ECCD at high temperature, and beam focused close to the resonance region. Benchmarking of the European Transport Simulator (ETS) against ASTRA and CRONOS transport codes, on a JET discharge, showed very good agreement among the simulated quantities, laying the foundations for its usage for both predictive and interpretative runs on present devices and ITER.

Some selected examples of ITM scientific workflow applications have also been outlined. Automated direct coupling of a core and edge transport code was demonstrated for the particular case of steady state and multiple impurities. The effect of NTMs on heat transport coefficients and temperature profiles was reproduced via a dedicated NTM module incorporated into the ETS. Coupling of the ETS to a free-boundary equilibrium solver was tested on a vertical displacement event (VDE) for an ITER scenario. Finally, ongoing efforts on the integration and testing of synthetic diagnostics in the ITM-TF platform have been reported, namely, the validation of spectral MSE forward model on ASDEX Upgrade data.

[17] www.efda-hlst.eu/.

Acknowledgments

This work, supported by the European Communities under the contract of Association between EURATOM and CEA, CCFE, ENEA, FOM, IPP, IPPLM, IST, Swiss Confederation, VR, was carried out within the framework of the Task Force on Integrated Tokamak Modelling of the European Fusion Development Agreement. Part of the research leading to these results received funding from the European Community's Research Infrastructures initiative of the 7th Framework Programme FP7 (2007–2013) under grant agreement No 211804 (EUFORIA). The views and opinions expressed herein do not necessarily reflect those of the European Commission. Dr X. Bonnin and O. Hoenen are gratefully acknowledged for their collaboration in the revision of the paper.

Appendix. List of ITM-TF Contributors

M. Airila[1], S. Äkäslompolo[2], L. Allegretti[3], L. Alves[4], T. Aniel[3], L. Appel[5], J-F. Artaud[3], O. Asunta[2], C.-V. Atanasiu[6], F. Aumayr[7], O. Barana[3], M. Baruzzo[8], V. Basiuk[3], R. Bilato[9], J. Bizarro[4], E. Blanco[21], J. Blum[10], T. Bolzonella[8], X. Bonnin[11], D. Borodin[12], A. Bottino[9], C. Boulbe[10], S. Brémond[3], S. Briguglio[13], Y. Buravand[3], A. Cardinali[13], C. Cianfarani[13], J. Citrin[14], R. Coelho[4], D. Coster[9], G. Csepany[17], D. Dimopoulos[18], A. Dinklage[19], C. DiTroia[13], C. Dritselis[18], D. Dumitru[6], R. Dumont[3], E. Fable[9], G.L. Falchetto[3], D. Farina[20], B. Faugeras[10], J. Ferreira[4], L. Figini[20], A. Figueiredo[4], G. Fogaccia[13], V. Fusco[13], K. Gal[41], J. Garcia[3], P. Garcia-Muller[21], L. Garzotti[5], J. Geiger[9], E. Giovannozzi[13], G. Giruzzi[3], V. Goloborodko[16], R. Goswami[3], B. Guillerminet[3], S. Hacquin[3], A. Hannan[22], J. Heikkinen[1], T. Hellsten[22], S. Heuraux[23], J. Hillairet[3], E. Hirvijoki[2], J. Hobirk[9], D. Hogeweij[14], L. Höök[22], P. Huynh[3], K. Igenbergs[24], F. Imbeaux[3], H. Isliker[15], I. Ivanova-Stanik[25], S. Janhunen[2], F. Jenko[9], T. Johnson[22], S. Kakarantzas[18], N. Kanaris[26], S. Kassinos[26], A. Keim[7], V. Kiptily[5], A. Kirschner[12], T. Kiviniemi[2], H.-J. Klingshirn[9], F. Köchl[7], Y. Kominis[35], T. Korpilo[2], L. Kos[27], T. Koskela[2], S. Kulovec[27], C. Lechte[28], E. Lerche[29], X. Litaudon[3], F. Liu[3], Y. Liu[5], J. Lönnroth[5], T. Lunt[9], G. Manduchi[8], N. Marushchenko[19], S. Mastrostefano[30], R. Mayo[21], J. Miettunen[2], S. Moradi[31], D. Moreau[3], P. Moreau[3], A. Morillas[21], D. Muir[5], Q. Mukhtar[22], F. Nabais[4], E. Nardon[3], F. Nave[4], A.H. Nielsen[32], R. Nouailletas[3], S. Nowak[20], M. O'Mullane[5], M. Ottaviani[3], M. Owsiak[33], V. Pais[6], B. Palak[33], A. Papadopoulos[35], G. Papp[17], N. Pelekasis[18], Y. Peysson[3], T. Pisokas[15], M. Plociennik[33], G. Pokol[17], E. Poli[9], G. Poulipoulis[36], I. Pusztai[31], H. Radhakrishnan[26], N. Ravenel[3], H. Reimerdes[37], D. Reiser[12], M. Romanelli[5], J. Rydén[31], A. Salmi[1], T. Samaras[35], O. Sauter[36], P. Scheier[37], K. Schmid[9], M. Schneider[3], K. Schöpf[16], B.D. Scott[9], J. Signoret[3], F. Silva[4], S. Sipilä[2], P. Siren[1], A. Sirinelli[5], A. Snicker[2], R. Stankiewicz[25], J. Storrs[5], P. Strand[38], E. Sundén[39], T. Tala[1], S. Tholerus[22], G. Throumoulopoulos[36], K. Tokesi[40], C. Tsironis[35], D. Tskhakaya[16], O. Tudisco[13], J. Urban[42], D.V. Eester[29], L. Villard[36], F. Villone[30], B. Viola[13], S. Viorica[6], G. Vlad[13], I. Voitsekhovitch[5], E. Westerhof[14], R. Wieggers[14], M. Wischmeier[9], D. Yadykin[38], P. Zestanakis[35], T. Zok[33]

[1] VTT Technical Research Centre of Finland, Association Euratom-TEKES, P O Box 1000, 02037 VTT, Finland
[2] Department of Applied Physics, Aalto University, Association Euratom-TEKES, PO Box 13500, FI-00076 AALTO, Finland
[3] CEA, IRFM, F-13108 Saint-Paul-lez-Durance, France
[4] Associação EURATOM/IST, Instituto de Plasmas e Fusão Nuclear, Instituto Superior Tècnico, Universidade Tècnica de Lisboa 1049-001 Lisboa, Portugal
[5] EURATOM/CCFE Fusion Association, Culham Science Centre, Abingdon OX14 3DB UK
[6] National Institute for Laser, Plasma and Radiation Physics, Association MEdC-EURATOM, Bucharest, Romania
[7] Association EURATOM-ÖAW/ATI, Atominstitut, TU Wien, 1020 Vienna, Austria
[8] Associazione EURATOM-ENEA sulla Fusione, Consorzio RFX, 29127 Padova, Italy
[9] Max-Planck-Institut für Plasmaphysik, EURATOM-IPP Association, Garching, Germany
[10] Univ. Nice Sophia Antipolis, Lab. JA Dieudonne, UMR 7351, F-06108 Nice 02, France
[11] CNRS-LSPM, Université Paris XIII, F-93430 Villetaneuse, France
[12] Institut für Energie und Klimaforschung Plasmaphysik, Forschungszentrum Jülich, Association Euratom-FZJ, Germany
[13] Associazione Euratom-ENEA sulla Fusione, C.R. ENEA-Frascati, Via E. Fermi 45, 00037 Frascati, Roma, Italy
[14] FOM Institute DIFFER, Association EURATOM-FOM, Nieuwegein, The Netherlands
[15] Section of Astrophysics, Astronomy and Mechanics, Department of Physics, University of Thessaloniki, Association Euratom-Hellenic Republic, Thessaloniki, Greece
[16] Association EURATOM-ÖAW, Institute for Theoretical Physics, University of Innsbruck, A-6020, Innsbruck, Austria
[17] Department of Nuclear Techniques, Budapest University of Technology and Economics, Association EURATOM, H-1111 Budapest, Hungary
[18] Department of Mechanical Engineering, University of Thessaly, Pedion Areos, Volos 38334, Greece-Association Euratom-Hellenic Republic
[19] Max Planck Institut für Plasmaphysik, EURATOM Association, Greifswald, Germany
[20] Istituto di Fisica del Plasma CNR, Euratom-ENEA-CNR Association for Fusion, Milan, Italy
[21] Asociación EURATOM CIEMAT, Madrid 28034, Spain
[22] Association EURATOM-VR, Fusion Plasma Physics, EES, KTH, SE-10037 Stockholm, Sweden
[23] Universitè de Lorraine, IJL, UMR 7198, BP 70233, Vandoeuvre, F-54506 Cedex, France
[24] Vienna Univ. Technol., Inst. Appl. Phys., A-1034 Vienna, Austria
[25] Institute of Plasma Physics and Laser Microfusion, EURATOM Association,00-908 Warsaw, Poland
[26] Computational Sciences Laboratory–UCY-CompSci, Department of Mechanical and Manufacturing Engineering, University of Cyprus, Nicosia, Cyprus
[27] University of Ljubljana, Faculty of Mech. Eng., Askerceva 6, SI-1000 Ljubljana, Slovenia

[28] Institute for Plasma Research, University of Stuttgart, 70569 Stuttgart, Germany
[29] Ecole Royale Militaire-Koninklijke Militaire School, Laboratoire de Physique des Plasmas, B-1000 Brussels, Belgium
[30] Associazione EURATOM/ENEA/CREATE, DIEI, Università degli Studi di Cassino e del Lazio Meridionale, Via Di Biasio 43, 03043 Cassino (FR), Italy
[31] Nuclear Engineering, Department of Applied Physics, Chalmers University of Technology, Euratom-VR Association, SE-35296 Göteborg, Sweden
[32] Association EURATOM-DTU, 3400 Roskilde, Denmark
[33] Poznan Supercomputing and Networking Center (PSNC), IChB PAS, Noskowskiego 12/14, Poznan, Poland
[34] School of Electrical and Computer Engineering, National Technical University of Athens, Association EURATOM-Hellenic Republic, Zografou, Athens 15773, Greece
[35] University of Ioannina, Association Euratom–Hellenic Republic, Section of Theoretical Physics, Ioannina, Greece
[36] Ecole Polytechnique Federale de Lausanne (EPFL), Centre de Recherches en Physique des Plasmas (CRPP), Association Euratom-Confederation Suisse, Lausanne, Switzerland
[37] Institut für Ionenphysik und Angewandte Physik, Universität Innsbruck
[38] Department of Earth and Space Sciences, Chalmers University of Technology, Euratom-VR Association, SE-352 96 Göteborg, Sweden
[39] Uppsala University, VR-Euratom Association, Box 516, 751 20 Uppsala, Sweden
[40] Institute of Nuclear Research of the Hungarian Academy of Sciences (ATOMKI), 3401 Debrecen, Hungary
[41] Institute for Particle and Nuclear Physics, Wigner Research Centre for Physics, Hungarian Academy of Sciences, EURATOM Association HAS, POB 49, H-1525, Budapest, Hungary
[42] Association EURATOM/IPP.CR, IPP AS CR, Prague, Czech Republic

References

[1] Becoulet A. *et al* 2007 *Comput. Phys. Commun.* **177** 55–9
[2] Strand P.I. *et al* 2010 *Fusion Eng. Des.* **85** 383–7
[3] Imbeaux F. *et al* 2010 *Comput. Phys. Commun.* **181** 987–98
[4] Coster D. *et al* 2010 *IEEE Trans. Plasma Sci.* **38** 2085
[5] Konz C. *et al* 2011 First physics applications of the Integrated Tokamak Modelling (ITM-TF) tools to the MHD stability analysis of experimental data and ITER scenarios *38th EPS Conf. on Plasma Physics (Strasbourg, France, 2011)* vol 35G (ECA) O2.103 ISBN 2-914771-68-1, http://ocs.ciemat.es/EPS2011PAP/pdf/O2.103.pdf
[6] Figini L. *et al* 2012 Benchmarking of electron cyclotron heating and current drive codes on ITER scenarios within the European Integrated Tokamak Modelling framework *EPJ Web of Conferences 32, 01011*. EC-17-*17th Joint Workshop on Electron Cyclotron Emission and Electron Cyclotron Resonance Heating (Deurne, 7–10 May 2012)* http://dx.doi.org/10.1051/epjconf/20123201011
[7] Kalupin D. *et al* 2011 Verification and validation of the European Transport Solver *38th EPS Conf. on Plasma Physics (Strasbourg, France, 2011)* vol 35G (ECA) P4.111 ISBN 2-914771-68-1, http://ocs.ciemat.es/EPS2011PAP/pdf/P4.111.pdf
[8] Coster D.P. *et al* 2012 Core–edge coupling: developments within the EFDA task force on Integrated Tokamak Modelling *39th EPS Conf. on Plasma Physics & 16th Int. Congress on Plasma Physics (Stockholm, Sweden, 2012)* vol 36F (ECA) P1.073 ISBN 2-914771-79-7, http://ocs.ciemat.es/epsicpp2012pap/pdf/P1.073.pdf
[9] Coelho R. *et al* 2013 *Fusion Sci. Technol.* **63** 1–8
[10] Zwingmann W. 2003 *Nucl. Fusion* **43** 842
[11] Zwingmann W. *et al* 2008 *PLASMA (2007) (Greifswald, Germany) AIP Conf. Proc.* **993** 11
[12] Huysmans G., Goedbloed J. and Kerner W. 1991 *CP90 Conf. Proc. on Computational Physics (Amsterdam, 10–13 Sept. 1990)* (Singapore: World Scientific) p 371
[13] Luetjens H., Bondeson A. and Sauter O. 1996 *Comput. Phys. Commun.* **97** 219
[14] Huysmans G.T.A. *et al* 2001 *Phys. Plasmas* **8** 4292
[15] Liu Y.Q., Bondeson A., Fransson C.M., Lennartson B. and Breitholtz C. 2000 *Phys. Plasmas* **7** 3681
[16] Zwingmann W. *et al* 2010 Validation procedure of the tokamak equilibrium reconstruction code equal with a scientific workflow system *37th EPS Conf. on Plasma Physics (Dublin,Ireland, 2010)* vol 34A (ECA) P4-180 ISBN 2-914771-62-2 P4-180 http://ocs.ciemat.es/EPS2010PAP/pdf/P4.180.pdf
[17] Coster D.P., Strand P. and Contributors to the EUFORIA Project 2010 EUFORIA: Exploring E-science for fusion (*Advances in Parallel Computing* vol 19) (Amsterdam: IOS Press) 520–9
[18] Goldston R.J. *et al* 1981 *J. Comput. Phys.* **43** 61
[19] Scott B. 2005 *Phys. Plasmas* **12** 102307
[20] Scott B. 2010 *Phys. Plasmas* **17** 102306
[21] Scott B. 2000 *Phys. Plasmas* **7** 1845–56
[22] Scott B. 2001 *Phys. Plasmas* **8** 447–58
[23] Naulin V. 2003 *Phys. Plasmas* **10** 4016–28
[24] Scott B. 2006 *Plasma Phys. Control. Fusion* **48** B277–93
[25] Shestakov A.I. *et al* 2003 *J. Comput. Phys.* **185** 399
[26] Candy J. *et al* 2009 *Phys. Plasmas* **16** 060704
[27] Barnes M. *et al* 2010 *Phys. Plasmas* **17** 056109
[28] Kalupin D. *et al* 2013 Numerical analysis of JET discharges with the European Transport Simulator *Nucl. Fusion* **53** 123007
[29] Hoenen O. *et al* 2013 Designing and running turbulence transport simulations using a distributed multiscale computing approach *40th EPS Conf. on Plasma Physics (Helsinki, Finland, 2013)* P4.155 http://ocs.ciemat.es/EPS2013PAP/pdf/P4.155.pdf
[30] Roberts S.W. 1959 *Technometrics* **1** 239–50
[31] Pereverzev G.V. and Corrigan G. 2008 *Comput. Phys. Commun.* **179** 579
[32] Basiuk V. *et al* 2010 European Transport Solver: first results, validation and benchmark *37th EPS Conf. on Plasma Physics (Dublin, Ireland, 2010)* vol 34A (ECA) P 1.1009 ISBN 2-914771-62-2; http://ocs.ciemat.es/EPS2010PAP/pdf/P1.1009.pdf
[33] Peysson Y. *et al* 2012 *Plasma Phys. Control. Fusion* **54** 045003
[34] Farina D. 2007 *Fusion Sci. Technol.* **52** 154
[35] Westerhof E. 1989 Implementation of TORAY at JET *Rijnhuizen Report* RR-89–183
[36] Poli E. *et al* 2001 *Comput. Phys. Commun.* **136** 90
[37] Marushchenko N.B. *et al* 2007 *J. Plasma Fusion Res.* **2** S1000
[38] Prater R. *et al* 2008 *Nucl. Fusion* **48** 035006
[39] Pereverzev G. and Yushmanov P.N. 2002 ASTRA Automated system for transport analysis in a tokamak *IPP*5/98 Max-Planck Institut für Plasmaphysik
[40] Artaud J.-F. *et al* 2010 *Nucl. Fusion* **50** 043001
[41] Coster D.P. 2009 *J. Nucl. Mater.* **390–391** 826
[42] Tarditi A. *et al* 1996 *Contrib. Plasma Phys.* **36** 132–5
[43] Schneider R. *et al* 2006 *Contrib. Plasma Phys.* **46** 3
[44] Fichtmüller M. *et al* 1998 *Czech. J. Phys.* **48** 25
[45] Senichenkov I.Y. *et al* 2012 Progress in ASTRA-B2SOLPS coupling for integrated tokamak modeling *39th EPS Conf. on Plasma Physics (Stockholm, Sweden, 2012)* vol 36F

(ECA) P2.042, ISBN 2-914771-79-7 http://ocs.ciemat.es/epsicpp2012pap/pdf/P2.042.pdf
[46] Senichenkov I.Y. *et al* 2011 Integrated modeling of H mode plasma in ASDEX Upgrade and Globus-M *38th EPS Conf. on Plasma Physics* (*Strasbourg, France, 2011)* vol 35G (ECA) P5.115 ISBN 2-914771-68-1, http://ocs.ciemat.es/EPS2011PAP/pdf/P5.115.pdf
[47] Kukushkin A.S. *et al* 2003 *Nucl. Fusion* **43** 716
[48] Owen L.W. *et al* 2010 *Nucl. Fusion* **50** 064017
[49] Äkäslompolo S. *et al* 2012 Preparing tokamak 3D wall and magnetic data for particle tracing simulations *39th EPS Conf. on Plasma Physics (Stockholm, Sweden, 2012)* vol 36F (ECA) P5.058 ISBN 2-914771-79-7 http://ocs.ciemat.es/epsicpp2012pap/pdf/P5.058.pdf
[50] Sauter O. *et al* 2002 *Plasma Phys. Control. Fusion* **44** 1999
[51] Turri G. *et al* 2008 *Proc. 22nd Int. Conf.* on *Fusion Energy (Geneva, Switzerland,2008)* (Vienna: IAEA) CD-ROM file EX/P3-06 and www-naweb.iaea.org/napc/physics/FEC/FEC2008/html/index.htm
[52] Sauter O. *et al* 1997*Phys. Plasmas* **4** 1654
[53] Houlberg W.A. *et al* 1997 *Phys. Plasmas* **4** 3230
[54] Gribov Y. *et al* 2007 Progress in the ITER Physics Basis: chapter 8. Plasma operation and control *Nucl. Fusion* **47** S385
[55] Hertout P. *et al* 2011 *26th Proc. Symp. on Fusion Technology (SOFT26, Porto, 2010) Fusion Eng. Des.* **86** 1045–8
[56] Urban J. *et al* 2012 Free-boundary equilibrium transport simulations of ITER scenarios under control *39th EPS Conf. on Plasma Physics (Stockholm, Sweden, 2012)* vol 36F (ECA) P1.019 ISBN 2-914771-79-7 http://ocs.ciemat.es/epsicpp2012pap/pdf/P1.019.pdf
[57] Artaud J.F. and Kim S.H. 2012 A new free-boundary equilibrium evolution code, FREEBIE *39th EPS Conf. on Plasma Physics (Stockholm, Sweden, 2012)* vol 36F (ECA) P4.023 ISBN 2-914771-79-7 http://ocs.ciemat.es/epsicpp2012pap/pdf/P4.023.pdf
[58] Blum J. and Le Foll J. 1984 *Comput. Phys. Rep.* **1** 465–94
[59] Fable E. *et al* 2013 *Nucl. Fusion* **53** 033002.
[60] Heikkinen J.A. *et al* 1993 *Comput. Phys. Commun.* **76** 215
[61] Dinklage A. *et al* 2011 *Fusion Sci. Technol.* **59** 406

Article G : [15] B. Faugeras et O. Maury. An advection-diffusion-reaction size-structured fish population dynamics model combined with a statistical parameter estimation procedure : Application to the Indian Ocean skipjack tuna fishery. *Math. Biosciences and Engineering* 2.4 (2005), p. 719–741

MATHEMATICAL BIOSCIENCES
AND ENGINEERING
Volume **2**, Number **4**, October **2005**

http://www.mbejournal.org/

pp. **1–23**

AN ADVECTION-DIFFUSION-REACTION SIZE-STRUCTURED FISH POPULATION DYNAMICS MODEL COMBINED WITH A STATISTICAL PARAMETER ESTIMATION PROCEDURE: APPLICATION TO THE INDIAN OCEAN SKIPJACK TUNA FISHERY

Blaise Faugeras

CNRS I3S,
Les Algorithmes, 2000 route des lucioles, BP 121, 06903, Sophia Antipolis Cedex, France

Olivier Maury

Institut de Recherche pour le Développement,
Centre de Recherche Halieutique, avenue Jean Monnet, BP 171, 34200 Sète, France

(Communicated by ??)

Abstract. We develop an advection-diffusion size-structured fish population dynamics model and apply it to simulate the skipjack tuna population in the Indian Ocean. The model is fully spatialized, and movements are parameterized with oceanographical and biological data; thus it naturally reacts to environment changes. We first formulate an initial-boundary value problem and prove existence of a unique positive solution. We then discuss the numerical scheme chosen for the integration of the simulation model. In a second step we address the parameter estimation problem for such a model. With the help of automatic differentiation, we derive the adjoint code which is used to compute the exact gradient of a Bayesian cost function measuring the distance between the outputs of the model and catch and length frequency data. A sensitivity analysis shows that not all parameters can be estimated from the data. Finally twin experiments in which pertubated parameters are recovered from simulated data are successfully conducted.

1. **Introduction.** Fish population dynamics models together with parameter estimation techniques are essential to provide assessment of the fish abundance and fishery exploitation level. Their use forms the basis of scientific advice for fisheries managements. This is particularly true for tuna fisheries, which are among the most valuable in the world and subject to increasing fishing pressure and to the effects of climate changes.

Discrete age-structured models with crude representations of space are most of the time used for fisheries stock assessments [1, 2]. The classical data used in fishery science to calibrate models are fishing effort, catch and length frequency data.

Length frequency data are not straightforward to use. Fish of the same age can exhibit very different sizes depending of their history [3, 4]. Therefore, to compare

2000 *Mathematics Subject Classification.* 92D25, 92D40, 86A05, 35K15, 35K20, 35K57, 65M06, 86A22, 65K10, 93B30.

Key words and phrases. Population dynamics model, size structure, well-posed initial-boundary value problem, statistical parameter estimation, tuna fisheries, stock-assessment.

the outputs of age-structured models with length frequency data, a Gaussian size distribution is generally added to each age class. However, because of non-uniform mortality over sizes, bias on growth and mortality estimates may result from this procedure [5].

Another point concerning tuna fisheries is that they are highly heterogeneous in space and time. This has a significant effect on their functioning. Important migrations of fish occur at various scales, so that fish movements have to be explicitly represented using spatialized models [6].

These are some of the main problems of current stock assessment models. It is necessary to carry on the modelling effort by proposing and testing more complex models dealing more accurately with size distributions and spatial heterogeneity. This paper follows this direction and its purpose is twofold.

First we describe in section 2 a model of population dynamics in which both size and space are taken as structure variables to account for growth, movements of fish, environmental variability and variable distribution of fishing effort. The model consists of an advection-diffusion-reaction equation. Spatial advection-diffusion models have a long history in ecology [7, 8, 9], but their use in fishery science has grown recently, particularly for tuna population modeling purposes [10, 11, 6]. To model tropical tuna population in the Indian Ocean realistically, our model needs to reflect the heterogeneous distribution and movements of the population linked to the environment and fishing effort heterogeneity. Thus in the model fish movements depend on oceanographical and biological data through a habitat suitability index. Recruitment, that is to say the input of young fishes in the model, is modeled as a source term involving a nonlocal nonlinearity. The structure of the model enables a direct and simultaneous comparison with the two main types of data available for tuna fisheries: catches and size frequencies.

We assess the mathematical well-posedness of the model in section 3. We formulate an initial-boundary value problem, introduce a variational formulation and show existence of a unique weak solution. As often with nonlinear problems the proof uses a fixed-point argument. We also show the positivity of the solution.

Our second goal is to develop and test a data assimilation procedure to estimate the parameters of the model in a realistic skipjack tuna fishing simulation. Indeed one of the main objectives of tuna population modeling is to provide robust evaluations of stocks which are hardly possible nowadays for tuna fisheries in the Indian Ocean because of the lack of robust estimations of many biological parameters, such as natural and fishing mortality rates or recruitment parameters. Section 4 deals with the numerical implementation of the simulation model. Then in Section 5 we describe the data assimilation method developed for parameter estimation as well as the Bayesian likelihood approach used to formulate a cost function measuring the distance between the outputs of the model and the data. The paper ends with some numerical experiments conducted in section 6 to validate the algorithm in the case of the Indian Ocean skipjack fishery.

2. **The model.**

The dynamics of the population of fish is described through a density function $p(x, y, s, t)$, where position $(x, y) \in \Omega$ the bounded domain representing the ocean, size or length $s \in (S_0, S_1)$ and time $t \in (0, T)$. The number of fish of size between s_1 and s_2 at time t and position (x, y) is given by the integral

$$\int_{s_1}^{s_2} p(x, y, s, t) ds.$$

The population density follows an advection-diffusion process in space. Let

$$D(x,y,s,t) = diag(D_u(x,y,s,t), D_v(x,y,s,t))$$

be the space diffusion matrix and

$$V(x,y,s,t) = (u(x,y,s,t), v(x,y,s,t))^T$$

be the velocity field. The population density also follows an advection-diffusion process in the size variable (see section 2.1 for more details). Let $d(s)$ denote the dispersion coefficient in size and $\gamma(s)$ be the growth rate. Finally let $m(s)$ and $F(x,y,s,t)$ denote the natural and fishing mortality rates, and $R(x,y,s,t,p)$ the recruitment source term (see sections 2.4 and 2.3). The density function p follows the balance law,

$$\begin{aligned} \partial_t p &= \operatorname{div}(D\nabla p) - \operatorname{div}(Vp) \\ &\quad +\partial_s(d\partial_s p) - \partial_s(\gamma p) \\ &\quad -(m+F)p + R(p), \quad \text{in } \Omega \times (S_0, S_1) \times (0,T), \end{aligned} \tag{1}$$

where ∇ and div are the usual differential operators on Ω.

This equation has to be completed with initial conditions

$$p(x,y,s,0) = p^0(x,y,s), \quad \forall (x,y,s) \in \Omega \times (S_0, S_1) \tag{2}$$

and boundary conditions

$$\partial_s p(x,y,S_0,t) = \partial_s p(x,y,S_1,t) = 0, \quad \forall (x,y,t) \in \Omega \times (0,T), \tag{3}$$

and

$$\nabla p(x,y,s,t) \cdot n(x,y) = 0, \quad \text{in } \partial\Omega, \ \forall (s,t) \in (S_0,S_1) \times (0,T), \tag{4}$$

where $n(x,y)$ is the unit normal vector pointing outside Ω. Homogeneous Neumann boundary conditions at $s = S_0$ and $s = S_1$ express the fact that the size of individuals can not reach values lower than S_0 or larger than S_1.

The parameterizations of the processes involved in the time evolution of the population are described in detail in the following subsections.

2.1. **Movements: Advection-diffusion in space.** Diffusion and velocity in space have a physical and a biological component. The biological components depend on a habitat suitability index function, $hsi(x,y,t)$, and its first space derivatives. The index hsi depends on temperature, $T(x,y,t)$, and forage, $Food(x,y,t)$ which are input data for the model. The biotic affinities for these environmental factors are defined as

$$f_T(x,y,t) = 1/(1 + exp(-\alpha_T(T(x,y,t) - T_0))), \tag{5}$$

and

$$f_{Food}(x,y,t) = Food(x,y,t)/(K_{Food} + Food(x,y,t)). \tag{6}$$

All parameters are given in Table 1, and Fig. 2 shows plots of f_T and f_{Food}. In its most general form the index is defined as

$$hsi(x,y,t) = (f_T(x,y,t))^{p_T} (f_{Food}(x,y,t))^{p_{Food}}. \tag{7}$$

The velocity field is computed as

$$V(x,y,s,t) = V_{phy}(x,y,t) + V_{hsi}(x,y,s,t), \tag{8}$$

TABLE 1. Habitat suitability index parameters

name	value	unit
α_T	0.35	$(degree\ C)^{-1}$
T_0	20	$degree\ C$
K_{Food}	1000	J
p_T	1	
p_{Food}	1	

where $V_{phy} = (u_{phy}, v_{phy})^T$ represents the physical velocity (computed by a hydrodynamical model) and $V_{hsi} = (u_{hsi}, v_{hsi})^T$ represents biological velocity defined by

$$\begin{aligned} u_{hsi}(x,y,s,t) &= u_{hsi0}(1-hsi(x,y,t))(\frac{\partial_x hsi(x,y,t)}{k_{hsi}+|\partial_x hsi(x,y,t)|})(\frac{s}{S_1}),\\ v_{hsi}(x,y,s,t) &= u_{hsi0}(1-hsi(x,y,t))(\frac{\partial_y hsi(x,y,t)}{k_{hsi}+|\partial_y hsi(x,y,t)|})(\frac{s}{S_1}). \end{aligned} \tag{9}$$

V_{hsi} is proportional to length (large fish can swim faster than small ones) and to $(1-hsi)\nabla hsi$ (the model transports the population towards the most suitable places for fish living according to the temporal habitat index evolution). The diffusion coefficients are defined as follows:

$$\begin{aligned} D_u(x,y,s,t) &= D_{min} + (D_{max}-D_{min})\\ &\quad \times(1-hsi(x,y,t))\\ &\quad \times(1-\frac{|\partial_x hsi(x,y,t)|}{k_{hsi}+|\partial_x hsi(x,y,t)|})(\frac{s}{S_1})^2,\\ D_v(x,y,s,t) &= D_{min} + (D_{max}-D_{min})\\ &\quad \times(1-hsi(x,y,t))\\ &\quad \times(1-\frac{|\partial_y hsi(x,y,t)|}{k_{hsi}+|\partial_y hsi(x,y,t)|})(\frac{s}{S_1})^2. \end{aligned} \tag{10}$$

The interpretation of such a parameterization is similar to the one given for V_{hsi} and all parameters are given in Table 2.

TABLE 2. Movements parameters

name	value	unit
D_{min}	10^4	$m^2.s^{-1}$
D_{max}	10^5	$m^2.s^{-1}$
u_{hsi0}	10	$m.s^{-1}$
k_{hsi}	2.5×10^{-7}	m^{-1}

2.2. **Growth and dispersion in size.** As time goes on and fish grow older, their size or length increases with a growth rate $\gamma(s)$ (see Eq. (11) and Table 3). A diffusion term in the size variable with a dispersion rate $d(s)$ (see Eq. (12) and Table 3) is included to account for individuals having the same age but different sizes. Indeed, in a fish population individuals of the same age can often differ markedly in size [4]. This variability in growth can result from many mechanisms, including genetic or behavorial traits that confer different performances to individuals, and factors such as environmental heterogeneity and variability [3]. In fishery science,

this variability is usually taken into account in age-structured models using a length-at-age relation perturbed by a Gaussian noise (see, for example, [12]). The model discussed here is size-structured and uses a diffusion term in the size variable with dispersion rate $d(s)$ to account for individuals having the same age but different sizes [13, 5]. The advection-diffusion term in size can be seen as the limit of a random walk model in which each individual grows with an average velocity but has at each time step a small binomial probability to grow faster or slower than this average (see [8] for more details). We consider that fish growth follows a Von Bertalanfy curve:

$$\gamma(s) = \gamma_1 - \gamma_2(\frac{s - S_0}{S_1 - S_0}), \tag{11}$$

$$d(s) = d_1 + d_2\frac{\gamma(s)}{\gamma_1}. \tag{12}$$

TABLE 3. Parameters for growth and dispersion in size

name	value	unit
γ_1	3.858×10^{-9}	$m.s^{-1}$
γ_2	3.858×10^{-9}	$m.s^{-1}$
d_1	3.215×10^{-12}	$m^2.s^{-1}$
d_2	3.215×10^{-12}	$m^2.s^{-1}$

2.3. **Recruitment.** Recruitment is computed as a function of the stock spawning biomass,

$$B(x, y, t, p) = \int_{s_{mat}}^{S_1} fr(s)w(s)p(x, y, s, t)ds, \tag{13}$$

where s_{mat} is the minimum size at maturity, the fecundity rate $fr(s)$ is given by

$$fr(s) = \frac{b_f}{(1 + exp(-a_f(s - s_{mat}))}, \tag{14}$$

and the weight $w(s)$ of a fish of size s by

$$w(s) = a_w s^{b_w}. \tag{15}$$

We use a Beverton and Holt [3] stock-recruitment relation and obtain,

$$R(x, y, s, t, p) = \mathbb{1}_{(S_0,s_r)}(s)\frac{b_0 B(x, y, t, p)}{k_B + B(x, y, t, p)}, \tag{16}$$

where $\mathbb{1}_{(S_0,s_r)}$ is the usual characteristics function and s_r is the maximum size of recruitment.

2.4. **Natural and fishing mortality.** The mortality rate is split into size-dependent natural mortality $m(s)$ [14] and a fishing mortality rate $F(x, y, s, t)$. The fishing mortality rate is defined as the sum of the N_f fishing mortality induced by each fleet,

$$F(x, y, s, t) = \sum_{f=1}^{N_f} F_f(x, y, s, t). \tag{17}$$

The mortality rate induced by each fleet is described by the following equation:

$$F_f(x, y, s, t) = q_f(s)E_f(x, y, t), \tag{18}$$

TABLE 4. Recruitment parameters

name	value	unit
b_f	0.5	
a_f	1	m^{-1}
s_{mat}	0.5	m
a_w	4.82×10^{-6}	$kg.m^{-b_w}$
b_w	3.36	
b_0	0.35×10^{-9}	$m^{-1}.s^{-1}$
k_B	0.5×10^{-4}	kg
s_r	0.5	m

where $q_f(s)$ is the size-dependent catchability coefficient for fleet f (that is the probability for a fish of size s to be caught by a unit of fishing effort of fleet f), and $E_f(x, y, t)$ is the observed fishing effort.

3. **Mathematical well-posedness.** In this section we prove existence and uniqueness of a positive weak solution to the model.

3.1. **Functional spaces.** Let us introduce some functional spaces. The study is conducted on the open set $\mathcal{Q} = \Omega \times (S_0, S_1)$. Let $T < \infty$ be a fixed time. The population density, p, is considered as an element of the functional space $H = L^2(\mathcal{Q})$, whose Hilbert space machinery is convenient to use. H is equipped with the scalar product

$$(p, q)_H = \int_\Omega \int_{S_0}^{S_1} pqdsdxdy,$$

and we denote by $||.||_H$ the induced norm. We also consider the separable Hilbert space defined by $H^1 = H^1(\mathcal{Q})$ and equipped with the scalar product

$$(p, q)_{H^1} = \int_\Omega \int_{S_0}^{S_1} (pq + \nabla p.\nabla q + \partial_s p \partial_s q)dsdxdy.$$

We denote by $||.||_{H^1}$ the induced norm on H^1.

We will also have to consider the Banch space $L^\infty = L^\infty(\mathcal{Q} \times (0, T))$ equipped with the norm

$$||p||_\infty = inf\{M, |p(x, y, s, t)| \leq M \ a.e. \ in \ \mathcal{Q} \times (0, T)\}.$$

$L^2(0, T, H)$ is the space of functions L^2 in time with values in H, equipped with the norm,

$$||p||_{L^2(0,T,H)} = [\int_0^T ||p(t)||_H^2 dt]^{1/2},$$

and $L^\infty(0, T, H)$ is the space of functions L^∞ in time with values in H, equipped with the norm,

$$||p||_{L^\infty(0,T,H)} = inf\{M, ||p(t)||_H \leq M \ a.e \ in \ (0, T)\}.$$

Similarly $C([0, T], H)$ is the space of continuous functions on $[0, T]$ with values in H. Further, $C([0, T], H)$, $L^\infty(0, T, H)$ and $L^2(0, T, H)$ are Banach spaces.
Classically H' denotes the dual of H and $(H^1)'$ the dual of H^1. When H is identified with its dual, we have the scheme

$$H^1 \subset H = H' \subset (H^1)',$$

where each space is dense in the following and the imbeddings are continuous. Let us denote by $W(H^1)$ the Hilbert space

$$W(H^1) = \{p \in L^2(0,T,H^1), \frac{dp}{dt} \in L^2(0,T,(H^1)')\}.$$

LEMMA 3.1. *Every $p \in W(H^1)$ is a.e equal to a continuous function from $[0,T]$ to H. Moreover we have the following continuous imbedding,*

$$W(H^1) \subset C([0,T],H).$$

Proof. See, for example, Dautray and Lions [15] □

3.2. **Assumptions on the data and preliminary transformation of the system.** The mortality rates are assumed to satisfy

- $m(s),\ F(x,y,s,t) \geq 0$ a.e in $\mathcal{Q} \times (0,T)$, $m,\ F \in L^\infty$.

If we assume that the input temperature and forage fields, $T(x,y,t)$ and $Food(x,y,t)$, are positive and regular enough, it appears clearly from section 2 that

- $D_u(x,y,s,t),\ D_v(x,y,s,t) \geq D_{min} > 0$, a.e in $\mathcal{Q} \times (0,T)$, $D_u,\ D_v \in L^\infty$;
- $u(x,y,s,t),\ v(x,y,s,t)$ are differentiable with respect to x and y, respectively and $u,\ v,\ \partial_x u,\ \partial_y v \in L^\infty$.

It is also clear from section 2 that

- $d(s) \geq d_1 > 0$, a.e in (S_0,S_1), $d \in L^\infty$;
- $\gamma(s)$ is differentiable with respect to s, and $\gamma, \partial_s\gamma \in L^\infty$;
- $fr(s),\ w(s) \geq 0$ a.e in (S_0,S_1), $fr,\ w \in L^\infty$.

We also assume that the initial distribution $p^0(x,y,s)$ satisfies

- $p^0(x,y,s) \geq 0$ a.e in $\mathcal{Q}$, $p^0 \in H$.

To prove our existence-uniqueness result, it is convenient to perform a change of unknown function: p satisfies (1)-(4) if and only if $\hat{p} = e^{-\lambda t}p$ is a solution to the same equations where $-(m+F)p$ is replaced with $-(m+F+\lambda)p$ in Eq. (1) and the recruitment $R(x,y,s,t,p)$ is replaced by

$$\hat{R}(x,y,s,t,\hat{p}) = \mathbb{1}_{[S_0,s_r]}(s)\frac{b_0 e^{-\lambda t}\hat{B}(x,y,t,\hat{p})}{k_B e^{-\lambda t} + \hat{B}(x,y,t,\hat{p})}, \tag{19}$$

$$\hat{B}(x,y,t,\hat{p}) = \int_{s_{mat}}^{S_1} fr(s)w(s)\hat{p}(x,y,s,t)ds. \tag{20}$$

In the remaining part of the mathematical analysis, this change of unknown is implicitly done and we omit the $\hat{p}$ notation. The constant λ will be fixed to a convenient value below. Moreover, the possible nullification of the term $k_B e^{-\lambda t} + \hat{B}(x,y,t,\hat{p})$ invites us to define

$$R(x,y,s,t,p) = \mathbb{1}_{[S_0,s_r]}(s)\frac{b_0 e^{-\lambda t}B(x,y,t,p)}{k_B e^{-\lambda t} + |B(x,y,t,p)|}, \tag{21}$$

This formulation is being used in the following. We will show that if the initial distribution, p^0 is nonnegative then $p \geq 0$ a.e. in $\mathcal{Q} \times (0,T)$; thus the two formulations are equivalent.

3.3. **Variational formulation.**

3.3.1. *The bilinear form $a(t,p,q)$.* Formally multiplying Eq. (1) by a function q and integrating by parts on $\mathcal{Q}$ leads to the definition of the following bilinear form. For $p,q \in H^1$ let us define

$$\begin{aligned} a(t,p,q) &= \int_{\mathcal{Q}} D\nabla p\nabla q dxdyds + \int_{\mathcal{Q}} V.\nabla pqdxdyds \\ &+ \int_{\mathcal{Q}} d(\partial_s p)(\partial_s q)dxdyds + \int_{\mathcal{Q}} \gamma(\partial_s p)qdxdyds \\ &+ \int_{\mathcal{Q}} (m + F + \lambda + \mathrm{div}(V) + \partial_s\gamma)pqdxdyds. \end{aligned} \tag{22}$$

LEMMA 3.2. *For a.e. $t \in (0,T)$, $a(t,p,q)$ is continuous on $H^1 \times H^1$, and for λ large enough, $a(t,p,q)$ is coercive on H^1. There exist two constants $C_1 > 0$ and $C_2 > 0$, depending on $||D||_\infty$, $||d||_\infty$, $||u||_\infty$, $||v||_\infty$, $||\gamma||_\infty$, $||\partial_x u||_\infty$, $||\partial_y v||_\infty$, $||\partial_s\gamma||_\infty$, $||F||_\infty$, $||m||_\infty$, D_{min}, d_1 and λ, such that*

$$|a(t,p,q)| \leq C_1||p||_{H^1}||q||_{H^1}, \ \forall p,q \in H^1, \tag{23}$$

$$a(t,p,p) \geq C_2||p||^2_{H^1}, \ \forall p \in H^1. \tag{24}$$

Proof. The proof is classical and we omit it. □

3.3.2. *The nonlinear operator $\mathcal{R}$.* In this section we show that the recruitment term $R(x,y,s,t,,p)$ (cf. Eqs. (13)-(16)) allows us to define a Lipschitz continuous operator $\mathcal{R}$ on $L^2(0,T,H)$.

LEMMA 3.3. *Let*

$$\Lambda = \frac{b_0(S_1 - S_0)||fr||_\infty||w||_\infty}{k_B};$$

then the application

$$p(x,y,s,t) \mapsto R(x,y,s,t,p)$$

defines a bounded nonlinear operator, $\mathcal{R}$, Lipschitz continuous from $L^2(0,T,H)$ to $L^2(0,T,H)$ with Lipschitz constant Λ.

Proof. Let us first notice that the application $(t,B) \mapsto h(t,B) = \dfrac{b_0 e^{-\lambda t}B}{k_B e^{-\lambda t} + |B|}$ from $[0,T] \times \mathbb{R}$ to $\mathbb{R}$ satisfies

$$|h(t,B)| \leq \frac{b_0}{k_B}|B|. \tag{25}$$

Furthermore $h(t,B)$ is Lipschitz continuous in B uniformly in $t \in [0,T]$,

$$|h(t,B^1) - h(t,B^2)| \leq \frac{b_0}{k_B}|B^1 - B^2|, \quad \forall B^1,B^2 \in \mathbb{R}, \ \forall t \in [0,T]. \tag{26}$$

Since $B(x,y,t,p)^2 = (\int_{s_{mat}}^{S_1} fr(s)w(s)p(x,y,s,t)ds)^2$, we obtain using Cauchy-Schwarz

$$B(x,y,t,p)^2 \leq (S_1 - S_0)||fr||^2_\infty||w||^2_\infty||p(x,y,.,t)||^2_{L^2(S_0,S_1)}. \tag{27}$$

Hence from (25) and (27) we deduce that $\forall t \in [0,T]$,

$$\begin{aligned} ||\mathcal{R}p(t)||^2_H &= \int_\Omega \int_{S_0}^{S_1} [\mathbb{1}_{[S_0,s_r]}(s)h(t,B(x,y,t,p))]^2 dxdyds, \\ &\leq \Lambda^2||p(t)||^2_H, \end{aligned} \tag{28}$$

and that $\mathcal{R}$ is well-posed on $L^2(0,T,H)$.

In the same way, if to p^1 (resp. p^2) we associate B^1 (resp. B^2), we deduce from (26) that $\forall t \in [0,T]$,

$$\begin{aligned} ||\mathcal{R}p^1(t) - \mathcal{R}p^2(t)||_H^2 &= \int_\Omega \int_{S_0}^{S_1} [\mathbb{1}_{[S_0,s_r]}(s)(h(t,B^1(x,y,t,p^1)) \\ &\quad -h(t,B^2(x,y,t,p^2)))]^2 dxdyds, \\ &\leq \Lambda^2 ||p^1(t) - p^2(t)||_H^2, \end{aligned} \tag{29}$$

and thus $\mathcal{R}$ is Lipschitz continuous on $L^2(0,T,H)$. □

3.3.3. *Weak solutions.* We can now give the definition of a weak solution to system (1)-(4).

Definition 3.1. *We say that $p \in W(H^1)$, is a weak solution of system (1)-(4) if*

$$\forall q \in H^1, \quad (\frac{dp}{dt}, q)_H + a(t,p,q) = (\mathcal{R}p, q)_H, \tag{30}$$

in the $\mathcal{D}'(]0,T[)$ sens,
and $p(0) = p^0$.

Then we can prove the following result.

Theorem 3.1. *System (1)-(4) admits a unique non-negative weak solution.*
Proof.

Existence and uniqueness. The proof consists mainly in defining a nonlinear operator Θ by freezing the nonlinear term $\mathcal{R}p$ and applying Banach fixed-point theorem to Θ. The fixed point is the desired solution.

Step 1. Let $\hat{p}$ be fixed in $W(H^1)$ and in Eq. (30) let us replace $(\mathcal{R}p,q)_H$ by $(\mathcal{R}\hat{p},q)_H$. The problem becomes linear in p and admits a unique solution (e.g, [15]). This solution defines an operator Θ on $W(H^1)$, $\Theta\hat{p} = p$.

Step 2. Let us show that for T sufficiently small Θ satisfies the following properties:

1. Θ leaves invariant the ball,

$$B_r = \{p \in W(H^1),\ ||p||_{L^\infty(0,T,H)} \leq r,\ r \geq \frac{||p^0||_H}{\sqrt{(1-(\Lambda^2 T/2C_2))}}\}$$

that is, $\Theta B_r \subset B_r$.
Taking $q = p$ as test function in (30), integrating on $[0,t]$, using the coerciveness of a and Cauchy-Schwarz inequality, we obtain

$$\int_0^t \frac{1}{2}\frac{d}{dt}||p(\sigma)||_H^2 + C_2||p(\sigma)||_{H^1}^2 d\sigma \leq \int_0^t ||\mathcal{R}\hat{p}(\sigma)||_H ||p(\sigma)||_H d\sigma.$$

For all $\alpha > 0$, Young inequality leads to

$$||p(t)||_H^2 + 2C_2 \int_0^t ||p(\sigma)||_{H^1}^2 d\sigma \leq \int_0^t \frac{1}{\alpha}||\mathcal{R}\hat{p}(\sigma)||_H^2 d\sigma + \int_0^t \alpha||p(\sigma)||^2 d\sigma + ||p^0||_H^2,$$

and choosing $\alpha = 2C_2$ gives

$$||p(t)||_H^2 \leq \frac{1}{2C_2}\int_0^t ||\mathcal{R}\hat{p}(\sigma)||_H^2 d\sigma + ||p^0||_H^2.$$

Then using Eq. (28) we obtain

$$||p(t)||^2_{L^\infty(0,T,H)} \leq \frac{\Lambda^2 T}{2C_2}||\hat{p}||^2_{L^\infty(0,T,H)} + ||p^0||^2_H.$$

If $||\hat{p}||_{L^\infty(0,T,H)} \leq r$ then $||p||_{L^\infty(0,T,H)} \leq r$ for $\frac{\Lambda^2 T}{2C_2}r^2 + ||p^0||^2_H \leq r^2$ that is to say $r^2(1 - \frac{\Lambda^2 T}{2C_2}) \geq ||p^0||^2_H$. This implies $\frac{\Lambda^2 T}{2C_2} < 1$ which is valid for small T, and $r \geq \dfrac{||p^0||_H}{\sqrt{(1 - (\Lambda^2 T/2C_2))}}$.

2. Θ is a strict contraction on B_r, there exists $0 < k < 1$ such that $\forall p^1,\ p^2 \in B_r$, $||\Theta p^1 - \Theta p^2||_{L^\infty(0,T,H)} \leq k||p^1 - p^2||_{L^\infty(0,T,H)}$.
Let $p^1 = \Theta\hat{p}^1$ and $p^2 = \Theta\hat{p}^2$. Substracting the two associated Eq. (30), taking $p^1 - p^2$ as test function and again using the coerciveness of a, Cauchy-Schwarz and Young inequality leads to

$$\frac{d}{dt}||p^1(t) - p^2(t)||^2_H \leq \frac{1}{2C_2}||\mathcal{R}\hat{p}^1(t) - \mathcal{R}\hat{p}^2(t)||^2_H.$$

Since $p^1(0) = p^2(0) = p^0$, we deduce

$$||p^1(t) - p^2(t)||^2_H \leq \frac{1}{2C_2}\int_0^t ||\mathcal{R}\hat{p}^1(\sigma) - \mathcal{R}\hat{p}^2(\sigma)||^2_H d\sigma,$$

and

$$||p^1 - p^2||^2_{L^\infty(0,T,H)} \leq \frac{\Lambda^2 T}{2C_2}||\hat{p}^1 - \hat{p}^2||^2_{L^\infty(0,T,H)}.$$

Then for $\frac{\Lambda^2 T}{2C_2} < 1$, Θ is a strict contraction.

Step 3. For T small enough, by Banach-fixed point theorem Θ admits a unique fixed point which is the desired solution on $(0, T)$. Since T does not depend on p^0, the same procedure can be applied on $(T, 2T)$, ... until a solution is found on the desired time interval.

Positivity. Let $p_1 \geq 0$ be given in $W(H^1)$, and let us define the sequence $(p_n)_{n\geq 1}$ by $\Theta p_n = p_{n+1}$. Let us prove that p_2 is non-negative:
Taking $p_2^- = max(0, -p_2)$ as test function in (30) leads to

$$(\frac{d}{dt}p_2, p_2^-)_H + a(t, p_2, p_2^-) = (\mathcal{R}p_1, p_2^-)_H,$$

and therefore to

$$\frac{1}{2}\frac{d}{dt}||p_2^-||^2_H \leq \frac{1}{2}\frac{d}{dt}||p_2^-||^2_H + a(t, p_2^-, p_2^-) = -(\mathcal{R}p_1, p_2^-)_H,$$

Since $p_1 \geq 0$ then $\mathcal{R}p_1 \geq 0$ and $-(\mathcal{R}p_1, p_2^-)_H \leq 0$. It results that

$$\frac{d}{dt}||p_2^-||^2_H \leq 0,$$

that is,

$$||p_2^-(t)||^2_H \leq ||p_2^-(0)||^2_H = ||p^{0-}||^2_H = 0,$$

and $p_2 \geq 0$. An induction then shows that $p_n \geq 0,\ \forall n \geq 1$, and since the sequence converges to the solution p, this latter is non-negative. □

4. **Numerical treatment of the model.** In the approximation procedure of the model a centered finite difference discretization is used. Equation (1) is solved on a grid with a spatial resolution of 2 degrees, i.e. $\Delta y = 120$ nautical miles in the latitudinal direction and $\Delta x = \Delta y \cos(\theta)$ in the longitudinal direction (θ is the latitude angle), a discrete length step, Δs of 4 cm, and a discrete time step Δt of one day is used. The discretization points are denoted by (x_i, y_j, s_l, t_n) with $i \in [1 : I]$, $j \in [1 : J]$ (assuming here for simplicity that the domain Ω is rectangular), $l \in [1 : L]$ and $n \in [1 : N]$. In what follows, $p^n_{i,j,l}$ denotes the numerical approximation of $p(x_i, y_j, s_l, t_n)$.

Several difficulties arise in the computation of the solution to Eq. (1). First the numerical scheme has to be very stable because of possible strong variations in space and time of the advection and diffusion coefficients, u, v, D_u, D_v. Moreover the numerical solution of Eq. (1) is to be used in a numerical function minimization procedure to obtain estimates of model parameters. Therefore the solution algorithm must be fast because the model and its adjoint may have to be solved hundreds of time. Moreover, the function minimization algorithm may test parameter values that do not necessarily guarantee numerical stability.

The selected scheme combines a splitting method [16, 17] and the use of the MUSCL scheme for advection terms (monotonic upstream centered scheme for conservation laws [18]). At each time step, given an approximation p^n of $p(x, y, s, t_n)$, the computation of p^{n+1} from p^n is achieved through four steps. The advection-diffusion equation in the x variable is integrated first, on $[t_n, t_{n+1}]$:

$$\begin{aligned} &\partial_t p(x, y, s, t) = \partial_x(D_u(x, y, s, t)\partial_x p) - \partial_x(u(x, y, s, t)p), \\ &p(x, y, s, t^n) = p^n. \end{aligned} \tag{31}$$

It results in a first approximation $p^{n+1,1}$. Then the advection-diffusion equation in the y variable is integrated on $[t_n, t_{n+1}]$ starting from $p^{n+1,1}$:

$$\begin{aligned} &\partial_t p(x, y, s, t) = \partial_y(D_v(x, y, s, t)\partial_y p) - \partial_y(v(x, y, s, t)p), \\ &p(x, y, s, t^n) = p^{n+1,1}. \end{aligned} \tag{32}$$

It results in a second approximation $p^{n+1,2}$. Then the advection-diffusion term in the s variable is integrated on $[t_n, t_{n+1}]$ starting from $p^{n+1,2}$:

$$\begin{aligned} &\partial_t p(x, y, s, t) = \partial_s(d(s)\partial_s p) - \partial_s(\gamma(s)p), \\ &p(x, y, s, t^n) = p^{n+1,2}. \end{aligned} \tag{33}$$

It results in a third approximation $p^{n+1,3}$. Finally mortality and recruitment are integrated on $[t_n, t_{n+1}]$ starting from $p^{n+1,3}$:

$$\begin{aligned} &\partial_t p(x, y, s, t) = -(m(s) + F(x, y, s, t))p + R(x, y, s, t, p), \\ &p(x, y, s, t^n) = p^{n+1,3}. \end{aligned} \tag{34}$$

It results in the final value p^{n+1}.

In each of the first three steps, diffusion is treated implicitly in time, and the MUSCL scheme is used for the advection term. For example, the discretization used to solve Eq. (31) can be written as follows:

$$
\begin{aligned}
&-\tfrac{\Delta t}{2\Delta x^2}(D^{n+1}_{u;i-1,j,l}+D^{n+1}_{u;i,j,l})p^{n+1}_{i-1,j,l}\\
&+(1+\tfrac{\Delta t}{2\Delta x^2}(D^{n+1}_{u;i-1,j,l}+2D^{n+1}_{u;i,j,l}+D^{n+1}_{u;i+1,j,l})p^{n+1}_{i,j,l}\\
&-\tfrac{\Delta t}{2\Delta x^2}(D^{n+1}_{u;i,j,l}+D^{n+1}_{u;i+1,j,l})p^{n+1}_{i+1,j,l}\\
&=\tfrac{\Delta t}{\Delta x}(f^n_{i+1/2,j,l}-f^n_{i-1/2,j,l}),
\end{aligned}
\tag{35}
$$

where

$$
f^n_{i+1/2,j,l}=\begin{cases}
u^n_{i+1/2,j,l}(p^n_{i,j,l}+1/2\Delta^n_i(1-u^n_{i+1/2,j,l}\frac{\Delta t}{\Delta x})), \text{ if } u^n_{i+1/2,j,l}\geq 0,\\
u^n_{i+1/2,j,l}(p^n_{i+1,j,l}-1/2\Delta^n_{i+1}(1+u^n_{i+1/2,j,l}\frac{\Delta t}{\Delta x})), \text{ otherwise.}
\end{cases}
\tag{36}
$$

In this last equation, if $(p^n_{i-1,j,l}\leq p^n_{i,j,l}\leq p^n_{i+1,j,l})$
or if $(p^n_{i+1,j,l}\leq p^n_{i,j,l}\leq p^n_{i-1,j,l})$ then

$$
\begin{aligned}
&\Delta_{min}=\min(\tfrac{|p^n_{i+1,j,l}-p^n_{i-1,j,l}|}{2},2|p^n_{i+1,j,l}-p^n_{i,j,l}|,2|p^n_{i,j,l}-p^n_{i-1,j,l}|),\\
&\Delta^n_i=sign(p^n_{i+1,j,l}-p^n_{i-1,j,l})\Delta_{min},
\end{aligned}
\tag{37}
$$

and otherwise,

$$
\Delta^n_i=0. \tag{38}
$$

5. **Parameter estimation.** In this section we describe the data assimilation algorithm developed to estimate the parameters of the model.

5.1. **The outputs of the model corresponding to the data.** Total catches in weight as well as length frequencies of the catches are computed and compared to observations to estimate the parameters of the model. In each cell (i,j) of the grid, where during month m (30 days), the fishing effort is nonzero, catches of fleet f are computed as follows:

$$
C_{i,j,m,f}=\sum_{l=1}^{L}\sum_{n=30(m-1)+1}^{30m}q_{f,l}E^n_{f,i,j,l}p^n_{i,j,l}w_l\Delta s\Delta x\Delta y\Delta t, \tag{39}
$$

and length frequencies as

$$
Q_{i,j,l,m,f}=\frac{\displaystyle\sum_{n=30(m-1)+1}^{30m}q_{f,l}E^n_{f,i,j,l}p^n_{i,j,l}\Delta s\Delta x\Delta y\Delta t}{\displaystyle\sum_{l=1}^{L}\sum_{n=30(m-1)+1}^{30m}q_{f,l}E^n_{f,i,j,l}p^n_{i,j,l}\Delta s\Delta x\Delta y\Delta t}. \tag{40}
$$

5.2. **The cost function.** The parameters of the model are denoted in what follows by $K\in\mathbb{R}^{N_p}$ where N_p is the number of parameters. K is being estimated in a Bayesian context by computing the mode of the posterior density function of the parameters knowing the data. We use the maximum of posterior distribution method [19], which involves minimizing the sum of the negative log-likelihood of the data plus the log of prior density functions.

We assume that the observation errors for catch data follow a log-normal distribution. Therefore the contribution of total catches to the negative log-likelihood

is

$$J_C(K) = \frac{1}{2\sigma_C^2} \sum_{i,j} \sum_m \sum_f (\log(C_{i,j,m,f}) - \log(C^{obs}_{i,j,m,f}))^2. \tag{41}$$

The observation errors for length frequency data are assumed to be normal and the contribution of frequency data to the negative log-likelihood reads

$$J_Q(K) = \frac{1}{2\sigma_Q^2} \sum_{i,j} \sum_l \sum_m \sum_f (Q_{i,j,l,m,f} - Q^{obs}_{i,j,l,m,f})^2. \tag{42}$$

The negative log of prior density functions for the parameters is

$$J_P(K) = \sum_n \frac{1}{2\sigma_n^2} (K_n - K_n^0)^2, \tag{43}$$

where K_n^0 are the reference a priori parameters given in Tables 1-4. The cost function to be minimized is the sum of those three terms:

$$J(K) = J_C(K) + J_Q(K) + J_P(K). \tag{44}$$

The parameters have different units and orders of magnitude. To avoid any numerical difficulties that might arise from this during the minimization, we adimensionalize the parameter vector K, dividing each parameter K_i by its first guess a priori value K_i^0. Let $D = diag(K_i^0)$ then the adimensionalized control vector is $k = D^{-1}K$. Such an adimensionalization procedure can be regarded as a preconditioning for the minimization. The final cost function is

$$j(k) = j(D^{-1}K) = J(K), \tag{45}$$

and the a priori reference adimentionalized parameter vector is $k^0 = 1$.

5.3. **Optimization: Computing the gradient with the adjoint model.** To minimize the cost function j, we used the quasi-Newton algorithm implemented in the $n1qn3$ Fortran subroutine of Gilbert and Lemaréchal [20]. The computation of the gradient of j with respect to control variables is required at each step of the minimization. This gradient results in one integration of the adjoint model. The adjoint code was obtained using the automatic differentiation program Odyssée [21, 22], which is an efficient tool for deriving adjoint codes since it enables the automatic production of adjoint instructions. However, codes produced by automatic differentiation do not usually use computer memory in a very efficient way. Saving the direct model trajectory is the major problem. A differentiation program has to follow systematic methods to provide the evaluation trajectory. Thus Odyssée systematically uses a local calculation and storage technique for the trajectory. Automatically differentiating a 3D model and using the adjoint code directly seems impossible for the moment. Thus the code generated by Odyssée had to be improved manually. A Taylor test was then conducted to compare the exact derivatives computed by the adjoint code to a finite difference approximation. Generally speaking, one aims at verifying that

$$r(\epsilon) = \frac{j(k + \epsilon\delta k) - j(k)}{\epsilon(\nabla j(k), \delta k)} \underset{\epsilon \to 0}{\longrightarrow} 0 \tag{46}$$

for any direction of perturbation δk. We present in Table 5 the result of such a test. As ϵ becomes smaller, one observes that the ratio $r(\epsilon)$ first tends linearly towards 1 up to $\epsilon = 10^{-6}$, which is the optimal value for a finite difference computation. Afterwards, the substraction of close floating-point numbers leads to a large cancellation error, which dominates the truncation error coming from the computation of

TABLE 5. Result of a Taylor test

ϵ	$r(\epsilon)$
10^{-1}	1.010756398
10^{-2}	1.001203422
10^{-3}	1.000115118
10^{-4}	1.000015861
10^{-5}	1.000001463
10^{-6}	1.000000518
10^{-7}	0.999993143
10^{-8}	0.999985731
10^{-9}	0.999847627
10^{-10}	0.997712868

the gradient by the finite difference method. A Taylor test with such a numerical behavior of the ratio $r(\epsilon)$ is said to be correct. It verifies that the adjoint code provides an exact computation of the gradient.

6. **Numerical results.**

6.1. **The simulation set-up.**

The standard run consists in a one-year simulation for the Indian Ocean. The spatial numerical grid used is shown on Fig. 1. Sizes of simulated skipjack tunas range from $S_0 = 0.4\ m$ to $S_1 = 1.2\ m$.

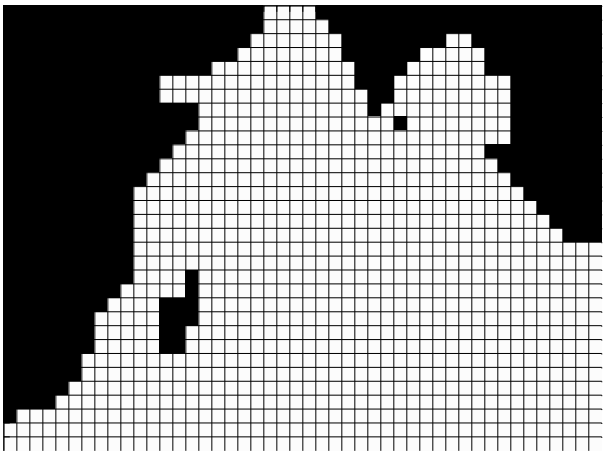

FIGURE 1. The 46×32 numerical grid used to integrate the population dynamics model on the Indian Ocean (earth in black, ocean in white).

Inititial conditions are chosen to be homogeneous over the space grid with a size distribution

$$p^0(x, y, s) = 0.1e^{-0.5s}. \tag{47}$$

This distribution assumes that the population is dominated by small organisms. Using these initial conditions a spin-up run of 6 years is conducted in order to reach an experimental and numerical fixed-point where mortality processes balance

the recruitment process and the total biomass slowly varies around a mean value during the year. The final distribution at the end of the spin up period provides the initial distribution for the standard run.

Since skipjack tunas inhabit the surface layer of the ocean, the model is forced with monthly velocity of oceanic surface currents, sea surface temperature and forage fields (Fig. 2). Velocity and temperature fields are outputs of the ocean general circulation model OPA,[1] whereas forage fields are outputs of a size structured model representing the energy flow in marine ecosystems from zooplankton to organisms of the size of tuna forage [23].

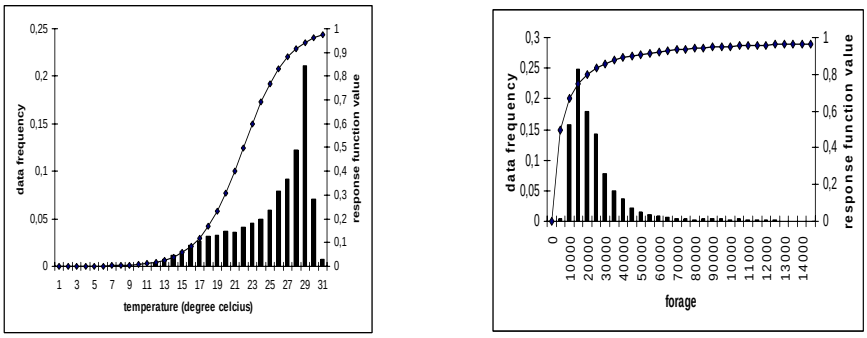

FIGURE 2. Response functions, f_T and f_F included in the hsi formulation corresponding to temperature (left) and forage (right). Frequency distributions over the whole grid and the whole year of temperature and forage values are also plotted.

Figure 3 shows a contour plot of the index hsi for the month of march. The strong south-north gradient of the index in the lower half of the map is typical of the area. Low temperatures are not suitable for tuna, which stay in the upper half of the map as shown on Fig. 4.

To test the possibility of estimating some parameters of the model from standard fishing data, we conduct in the following section numerical experiments with a synthetic data set computed by one simulation of the model. All parameters are set to their reference a priori values. Moreover for the sake of simplicity the two mortality parameters m and q are assumed to be size-independent; that is constants with values $m = 4.2438\ 10^{-8}\ s^{-1}$ [6] and $q = 6.43\ 10^{-8}\ s^{-1}$. Only one fleet is considered ($N_f = 1$), and the fishing effort is assumed to be constant and homogeneous during the year on an area roughly corresponding to the real fishing areas of the purse seine fleet in the Indian Ocean (see Fig. 5). With this configuration, a data set is computed following Eqs. (39) and (40).

6.2. **Sensitivity analysis.** We conduct a sensitivity analysis in order to identify the most important input parameters whose changes impact the most the outputs of the model (catches and length frequencies).

[1] http://www.lodyc.jussieu.fr/opa/

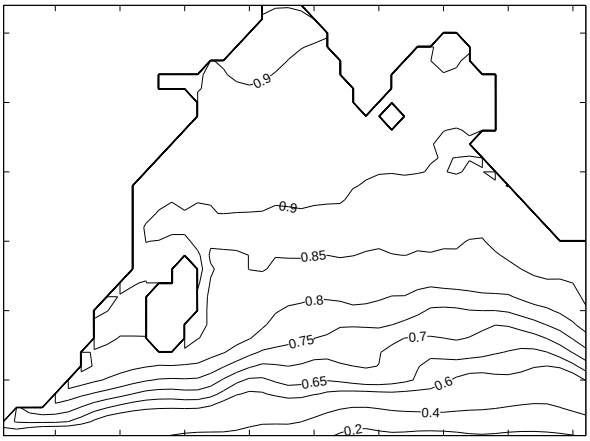

FIGURE 3. Contour plot of the index *hsi* for the month of March.

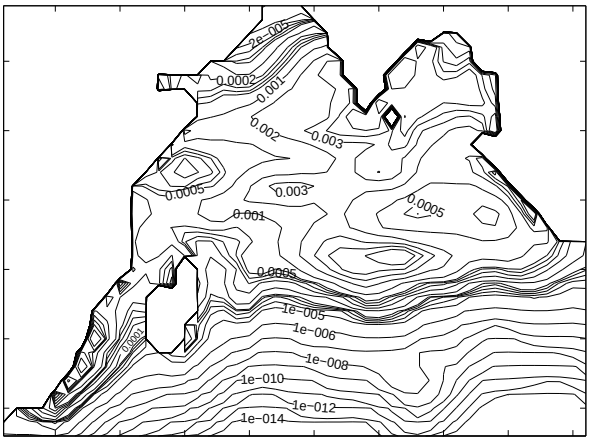

FIGURE 4. Contour plot of the population density function summed over size classes at the end of the month of March.

As a measure of the outputs of the model we consider the quantities

$$h_C(k) = \sum_{i,j} \sum_{m} (C_{i,j,m})^2 \tag{48}$$

and

$$h_Q(k) = \sum_{i,j} \sum_{l} \sum_{m} (Q_{i,j,l,m})^2. \tag{49}$$

Then the vector of relative sensitivities of these quantities to variations of the input parameters computed at the point k are

$$s_C(k) = \frac{\nabla h_C(k)}{h_C(k)} \tag{50}$$

and

$$s_Q(k) = \frac{\nabla h_Q(k)}{h_Q(k)}. \tag{51}$$

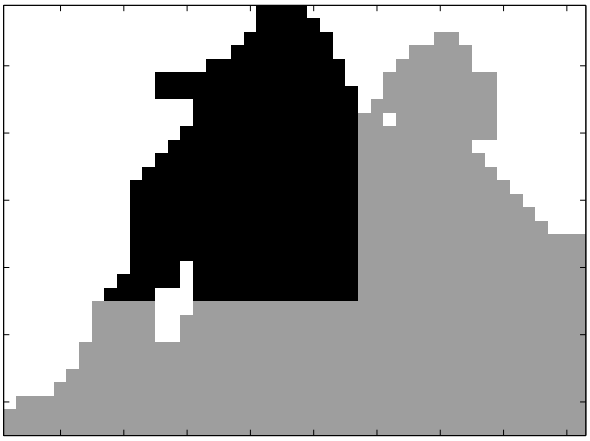

FIGURE 5. Distribution of the fishing efforts used in the simulations. $F = 0$ in the gray area and $F = 1$ in the black area.

Each of these relative sensitivity vector requires the computation of a gradient which is easily obtained with one integration of the adjoint model.

Table 6 shows the relative sensitivity vectors, $s_C(k^0)$ and $s_Q(k^0)$, computed with the initial a priori parameter vector k^0. Globally the different relative sensitivities are quite low, indicating that it may be difficult to estimate correctly all the parameters of the model using the two types of fishery data which are generally available. In the remaining part of this paper we will not try to estimate parameters which have the lowest sensitivities (10^{-3}, 10^{-2}). These parameters are

- D_{min}, D_{max} related to diffusion in space,
- d_1, d_2, γ_1 and γ_2 related to growth,
- a_f and s_{mat} related to recruitment.

Interestingly the relative sensitivities s_Q corresponding to variations in b_f, k_B and a_w (see Table 6) are exactly equal up to a sign. This comes from the formulation of recruitment in the model, which from Eqs. (13)-(16) can be rewritten with obvious notations as

$$\frac{b_0 B}{k_B + B} = \frac{b_0 b_f a_w \hat{B}}{k_B + b_f a_w \hat{B}} = \frac{b_0 \hat{B}}{\dfrac{k_B}{b_f a_w} + \hat{B}}. \tag{52}$$

An inconsistency in the formulation of the inverse parameter estimation problem appears clearly. The 3 parameters, b_f, k_B and a_w can not be determined independently, since for example an increase in k_B can also be interpreted as a decrease in b_f or in a_w. For this reason in our identification experiments we keep b_f and a_w fixed to their reference values and only try to estimate k_B. Moreover since the length/weight parameters a_w and b_w are well known we also do not select b_w for the parameter estimation formulation.

Although the 2 mortality parameters m and q do not correspond to very high sensitivities, we will try to estimate them since they really are badly known.

Finally the chosen formulation includes 11 parameters to be estimated:

- movements parameters : $k_{hsi}, p_T, p_{Food}, \alpha_T, T_0, K_{Food}, u_{hsi0}$
- recruitment parameters : b_0, k_B

- mortality parameters : m, q

TABLE 6. Relative sensitivities of catch and length frequency data

param.	$s_C(k^0)$	$s_Q(k^0)$	
D_{min}	$1.39\ 10^{-2}$	$-8.21\ 10^{-3}$	$-$
D_{max}	$7.81\ 10^{-2}$	$-1.12\ 10^{-2}$	$-$
k_{hsi}	$1.36\ 10^{-1}$	$-3.59\ 10^{-2}$	$+$
p_T	$-1.54\ 10^{-1}$	$4.75\ 10^{-2}$	$+$
p_{Food}	$-1.06\ 10^{-1}$	$1.50\ 10^{-2}$	$+$
α_T	-1.00	$3.12\ 10^{-1}$	$+$
T_0	$2.58\ 10^{-1}$	$-8.06\ 10^{-2}$	$+$
K_{Food}	$-1.01\ 10^{-1}$	$1.43\ 10^{-2}$	$+$
u_{hsi0}	$-1.68\ 10^{-1}$	$4.08\ 10^{-2}$	$+$
d_1	$5.53\ 10^{-3}$	$-3.83\ 10^{-2}$	$-$
d_2	$4.78\ 10^{-3}$	$-3.47\ 10^{-2}$	$-$
γ_1	$3.87\ 10^{-2}$	$-3.51\ 10^{-1}$	$-$
γ_2	$-4.93\ 10^{-3}$	$3.36\ 10^{-2}$	$-$
b_0	$1.93\ 10^{-1}$	$3.02\ 10^{-1}$	$+$
b_f	$-1.55\ 10^{-1}$	$-2.49\ 10^{-1}$	$-$
k_B	$1.55\ 10^{-1}$	$2.49\ 10^{-1}$	$+$
a_w	$3.19\ 10^{-1}$	$2.49\ 10^{-1}$	$-$
b_w	4.16	3.24	$-$
a_f	$-1.22\ 10^{-3}$	$-2.07\ 10^{-3}$	$-$
s_{mat}	$-3.90\ 10^{-2}$	$-6.27\ 10^{-2}$	$-$
m	$-9.11\ 10^{-2}$	$3.42\ 10^{-2}$	$+$
q	$7.47\ 10^{-2}$	$2.49\ 10^{-2}$	$+$

Note: In the last column, a + or − indicates whether or not the corresponding parameter is estimated.

6.3. **Identification experiments.** An essential validation step to perform before assimilation of real observed data is to conduct twin experiments. Synthetic data are produced by the model using the first guess parameter vector k^0. To fully test the possibility of recovering the selected parameters from the synthetic data, no penalty term is added and the cost function reduces to

$$j(k) = j_C(k) + j_Q(k). \tag{53}$$

The assumed variances are as follows: $\sigma_C = 0.1$ and $\sigma_Q = 0.01$. This provides a good balance between the two terms of j. In the experiments conducted, the convergence criterion is $\dfrac{||\nabla j(k)||}{||\nabla j(k^0)||} \leq \epsilon$, where ϵ is a small value fixed to 10^{-5}.

6.3.1. *Experiment 1.* A first numerical experiment was conducted to assess the capacity of the parameter estimation algorithm to distinguish between low (or high) recruitment and high (or low) mortality rates on the one hand and natural and fishing mortality rates on the other hand. Therefore in this optimization only the four parameters b_0, k_B, m, and q can vary, the others being fixed to their reference a priori value used to simulate the data. Different first guesses for the parameter vector were obtained by perturbing these four parameters within reasonable range (up to 50% of their reference value). All the corresponding optimizations converged

to the minimum of the cost function. The results of such an experiment are shown in Figs. 6, 7, and 8. The convergence criterion is satisfied after 20 iterations. The cost function value decreased from $2.9\ 10^5$ to $5.1\ 10^{-7}$ indicating that it has reached its global minimum and all 4 parameters have been recovered.

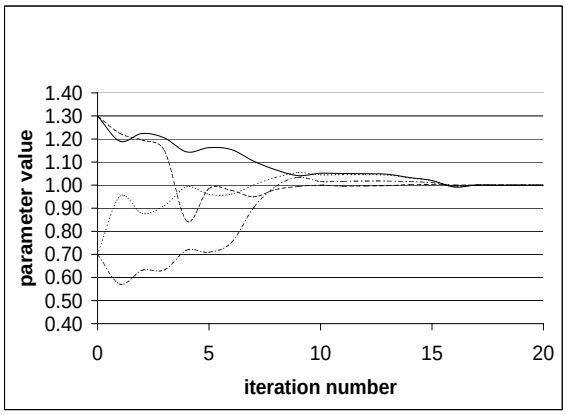

FIGURE 6. Convergence of the 4 selected parameters towards their reference value ($k^0 = 1$) during the optimization experiment 1.

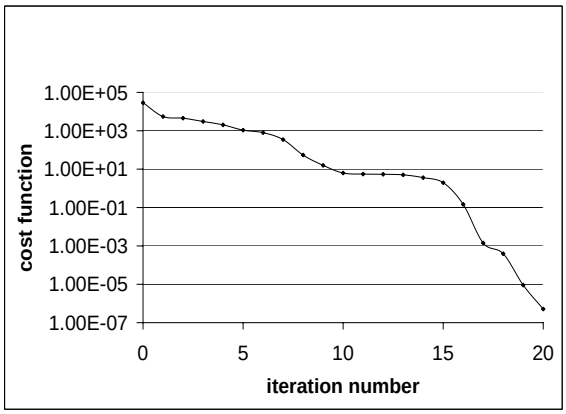

FIGURE 7. Evolution of the cost function $j(k)$ during the optimization experiment 1.

6.3.2. *Experiment 2.* A second numerical experiment was conducted to assess the capacity of the parameter estimation algorithm to recover all the 11 parameters at the same time. Therefore in this second optimization experiment all of 11 parameters can vary. Different first guesses for the parameter vector were obtained by perturbing these parameters within reasonable range (up to 20% of their reference

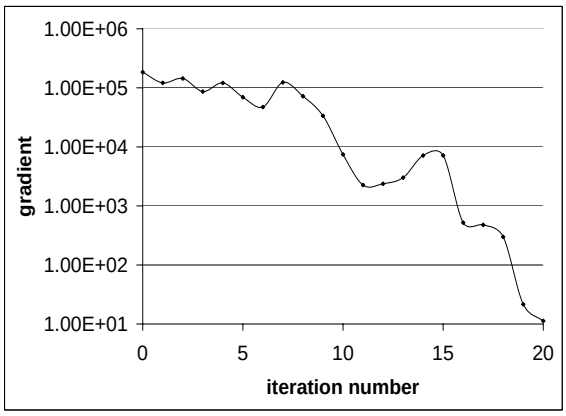

FIGURE 8. Evolution of the gradient $||\nabla j(k)||$ during the optimization experiment 1.

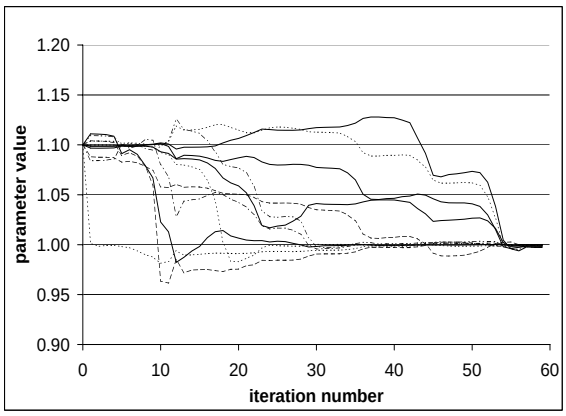

FIGURE 9. Convergence of the 11 selected parameters towards their reference value ($k^0 = 1$) during the optimization experiment 2.

value). All the corresponding optimizations converged to the minimum of the cost function. The results of such an experiment are shown in Figs. 9, 10, and 11.

The convergence criterion is satisfied after 59 iterations. The cost function value decreased from 6.4 $10^6$to 2.0 10^{-1}, indicating that it has reached its global minimum and all 11 parameters have been recovered.

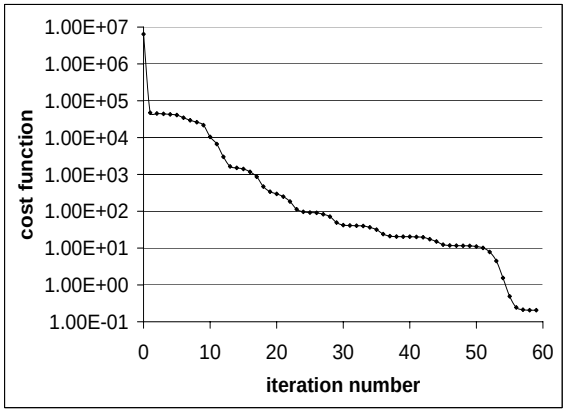

FIGURE 10. Evolution of the cost function $j(k)$ during the optimization experiment 2.

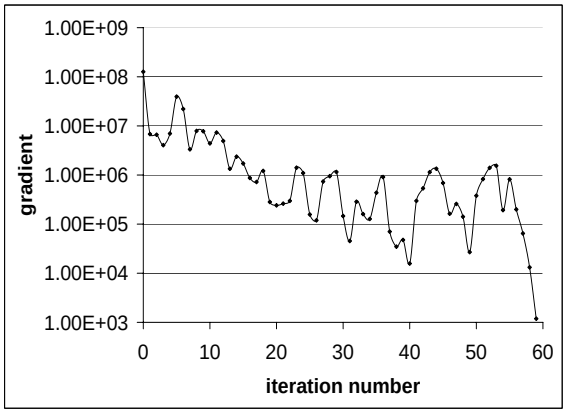

FIGURE 11. Evolution of the gradient $||\nabla j(k)||$ during the optimization experiment 2.

7. **Conclusion.** We developed an advection-diffusion size-structured fish population dynamics model and applied it to simulate the skipjack tuna population in the Indian Ocean. The model is fully spatialized and movements are parameterized with oceanographical and biological data. Thus the model naturally reacts to environmental and climatic changes. We have formulated an initial-boundary value problem and proved its mathematical well-posedness. We then discussed the numerical scheme chosen for the integration of the simulation model. From a modeling point of view, this study, is to our knowledge, the first one in which space and size structure of the population are fully taken into account and in which both mathematical and numerical difficulties were dealt with in a rigorous manner.

In a second step we addressed the parameter estimation problem for such a model. With the help of automatic differentiation we derived the adjoint code which enabled us to compute the exact gradient of a Bayesian cost function measuring the distance between the outputs of the model and catch and length frequency data. Thanks to the size structure of the modeled population the outputs of the model can be naturally compared to length frequency data. A sensitivity analysis showed that not all parameters could be estimated from the data. Finally twin experiments in which pertubated parameters were recovered from simulated data were successfully conducted. This point is particularly crucial since one limitation of the model lies in the choice to be made for different parameters value, or even in the choice to be made in the type of functions of temperature or forage parameterizing the habitat. The numerical experiments conducted demonstrate that fishing data can be used to estimate these parameters accurately.

This study is an important first step towards the assimilation of real observed fishing data in the model which is under progress. The mathematical and numerical tools which have been developed and validated will be extended to confront the model with tagging data which should bring more information and enable the estimation of several supplementary parameters such as growth and movements parameters. Developing a tool using tagging data is indeed especially timely, since no reliable stock assessment can be conducted at present for the skipjack tuna in the Indian Ocean and since a large-scale tuna tagging program in the Indian Ocean (IOTTP) has recently started and an important tag-recapture data set will be available in the coming months.

REFERENCES

[1] B. A. Megrey. Review and comparison of age-structured stock assessment models from theoritical and applied points of view. In E.F. Edwards and B.A Megrey, editors, *Mathematical analysis of fish stocks dynamics*, volume 6, pages 8–48. AM. Fish. Soc. Symp., 1989.

[2] D. A. Fournier, J. Hampton, and J. R. Sibert. MULTIFAN-CL: A length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. *Can. J. Fish. Aquat. Sci.*, 55:2105–2116, 1998.

[3] R. J. H. Beverton and S. J. Holt. *On the Dynamics of of Exploited Fish Populations*. Fish and Fisheries Series 11. Chapman & Hall, 1996.

[4] A. Pfister. Some consequences of size variability in juvenile prickly sculpin, *Cottus asper*. *Environmental Biology of Fishes*, 66:383–390, 2002.

[5] O. Maury and B. Faugeras. FASST: A fully age-size and space-time structured statistical model for the assessment of tuna populations. *ICCAT Coll. Vol. Sci. Pap.*, 57(1):206–217, 2005.

[6] J. R. Sibert, J. Hampton, D. A. Fournier, and P. J. Bills. An advection-diffusion-reaction model for the estimation of fish movement parameters from tagging data, with application to skipjack tuna (*Katsuwonus pelamis*). *Can. J. Fish. Aquat. Sci.*, 56:925–938, 1999.

[7] J. G. Skellam. Random dispersal in theoritical populations. *Biometrika*, 38:196–218, 1951.

[8] A. Okubo. *Diffusion and Ecological Problems: Mathematical Models*, volume 10 of *Biomathematics*. Springer-Verlag, 1980.
[9] E. E. Holmes, M. A. Lewis, J. E. Banks, and R. R. Veit. Partial differential equations in ecology: Spatial interactions and population dynamics. *Ecology*, 75(1):17–29, 1994.
[10] M. Bertignac, P. Lehodey, and J. Hampton. A spatial population dynamics simulation model of tropical tunas using a habitat index based on environmental parameters. *Fisheries Oceanography*, 7(3/4):326–334, 1998.
[11] O. Maury and D. Gascuel. SHADIS (Simulateur HAlieutique de DYnamiques Spatiales), a GIS based numerical model of fisheries. example application : The study of a marine protected area. *Aquat. Living Resour.*, 12(2):77–88, 1999.
[12] D. A. Fournier and J. R. Sibert. MULTIFAN a Likelihood-Based Method for Estimating Growth Parameters and Age Composition from Multiple Length Frequency Data Sets Illustrated using Data for Southern Bluefin Tuna (*Thunnus maccoyii*). *Can. J. Fish. Aquat. Sci.*, 47:301–317, 1990.
[13] B. Faugeras and O. Maury. A multi-region nonlinear age-size structured fish population model. *Nonlinear Analysis: Real World Appl.*, 6(3):447–460, 2005.
[14] J. Hampton. Natural mortality rates in tropical tunas: size really does matter. *Can. J. Fish. Aquat. Sci.*, 47:1002–1010, 2000.
[15] R. Dautray and J.-L. Lions. *Analyse mathématique et calcul numérique pour les sciences et les techniques*, volume 8. Masson, 1988b.
[16] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.
[17] G. I. Marchuk. Splitting and alternating direction methods. In *Handbook of numerical analysis*, volume I, pages 197–462. North-Holland, Amsterdam, 1990.
[18] B. Van Leer. Towards the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection. *J. Comput. Phys.*, 23:276–299, 1977.
[19] Y. Bard. *Nonlinear parameter estimation*. Academic Press, San Diego, CA, 1974.
[20] J. C. Gilbert and C. Lemaréchal. Some numerical experiments with variable storage quasi-newton algorithms. *Mathematical Programming*, 45:407–435, 1989.
[21] C. Faure and Y. Papegay. Odyssée Version 1.6, the language reference manual. Rapport Technique 211, INRIA, 1997.
[22] A. Griewank. *Evaluating Derivatives. Principles and Techniques of Algorithmic Differentiation*. Frontiers in applied mathematics. SIAM, Philadelphia, 2000.
[23] O. Maury, B. Faugeras, Y-J. Shin, T. Ben Ari, and F. Marsac. End to end modelling of the size-structured energy flow through marine ecosystems. Submitted to Journal of Theoretical Biology, 2005.

Received on ??, ??, 2005. Revised on ??, ??, 2005.

E-mail address: `Blaise.Faugeras@unice.fr or Blaise.Faugeras@ifremer.fr`
E-mail address: `Olivier.Maury@ird.fr`

Article H : [16] B. Faugeras et O. Maury. Modelling fish population movements : from an individual-based representation to an advection-diffusion equation. *J. Theor. Biol.* 247 (2007), p. 837–848

Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 247 (2007) 837–848
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Modeling fish population movements: From an individual-based representation to an advection–diffusion equation

Blaise Faugeras*, Olivier Maury

IRD, UR 109 THETIS, CRH, Avenue Jean Monnet, B.P. 171, 34203 Sète cedex, France

Received 7 February 2007; received in revised form 6 April 2007; accepted 10 April 2007
Available online 13 April 2007

Abstract

In this paper, we address the problem of modeling fish population movements. We first consider a description of movements at the level of individuals. An individual-based model is formulated as a biased random walk model in which the velocity of each fish has both a deterministic and a stochastic component. These components are function of a habitat suitability index, h, and its spatial gradient ∇h. We derive an advection–diffusion partial differential equation (PDE) which approximates this individual-based model (IBM). The approximation process enables us to obtain a mechanistic representation of the advection and diffusion coefficients which improves the heuristic approaches of former studies. Advection and diffusion are linked and exhibit antagonistic behaviors: strong advection goes with weak diffusion leading to a directed movement of fish. On the contrary weak advection goes with strong diffusion corresponding to a searching behavior. Simulations are conducted for both models which are compared by computing spatial statistics. It is shown that the PDE model is a good approximation to the IBM.

Keywords: Population dynamics; Biased random walk; Individual-based model; Partial differential equation

1. Introduction

Population dynamics models are essential to help to understand marine ecosystems dynamics and to provide assessment of fish abundance and fishery exploitation level. This is particularly true in the case of tuna fisheries, which are among the most valuable in the world and subject to increasing fishing pressure and to the effects of climate change. Although fish are mobile, models of population dynamics without any or with very crude representation of space are most of the time used for fisheries stock assessments. However, in order to understand the reasons and consequences of resource variability, many recent studies of ecological dynamics have emphasized the necessity to develop and use spatially explicit approaches.

Fish population dynamics can be represented with such partial differential equations (PDEs). Spatial advection–diffusion models have a long history in ecology (e.g. Skellam, 1951; Okubo, 1980; Holmes et al., 1994), but their use in fishery science has grown recently, particularly for tuna population modeling purposes (Bertignac et al., 1998; Maury and Gascuel, 1999; Sibert et al., 1999; Lehodey et al., 2003; Faugeras and Maury, 2005). Among the difficulties which arise with such models an important one is the choice that has to be made to express the time and space dependent advection and diffusion coefficients.

A first approach, used by Sibert et al. (1999), is to set these parameters to be constant over large spatial regions and temporal seasons and to try to estimate them by minimizing a cost function describing the distance between the outputs of the model and the available data. This is not completely satisfying since the spatio-temporal variability of advection and diffusion terms is roughly represented.

A second approach followed, for instance, by Bertignac et al. (1998) and Faugeras and Maury (2005) is to parameterize advection and diffusion terms as functions of an habitat suitability index. This approach has the advantage to fully take into account the spatio-temporal variability of the habitat of a fish population with a small

*Corresponding author. Tel.: +33 4 99 57 32 27.
E-mail address: Blaise.Faugeras@mpl.ird.fr (B. Faugeras).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.04.012

number of parameters. However, its main drawback is that the expressions chosen to parameterize advection and diffusion coefficients are arbitrary. The advection field $\mathbf{V}$ is usually considered to be proportional to the spatial gradient of the habitat suitability index h : $\mathbf{V} = c\nabla h$. The coefficient c is the taxis coefficient. It determines the rate of movements of fish up gradients of the habitat suitability index. This coefficient can be a constant (Bertignac et al., 1998; Maury, 2000), or a simple and empirical function of h and ∇h (Faugeras and Maury, 2005). The diffusion matrix is always supposed to be diagonal. Its diagonal elements are either assumed to be constant or simple arbitrary functions of h.

In this paper we provide a mechanistic approach to derive an advection–diffusion fish population dynamics model from individual fish behavior. Our approach is based on a biased random walk model. This type of model can also be viewed as simple individual-based models (IBMs). Such models are useful to describe movements at the level of individuals but cannot be easily used to treat large populations. Instead some level of approximation has to be made to reduce the problem to a state equation in which the variable is the spatial density of individuals. Related works, concerning the transformation of an individual-based or microscopic modeling into a population-based or macroscopic modeling, are Alt (1980) and Grünbaum (1999) in which the authors show that the solutions of an underlying differential–integral equation describing the movements of animals satisfy, under suitable assumptions, an advection–diffusion equation. One can also be interested in Flierl et al. (1999) where the authors analyze the processes by which organisms form groups and discuss the transformation of IBM into continuum models. In the present study, an advection–diffusion equation is obtained as a truncated Kramers–Moyall cumulant expansion (Risken, 1996) of the spatial density function of individuals. The parameters of the IBM are used in the expressions of the advection and diffusion terms. A consistent behavior is obtained concerning the dependence of these two terms on h and ∇h, and the balance between them. Advection and diffusion both are decreasing functions of the habitat index h. Moreover their dependence on ∇h implies that strong advection goes with weak diffusion leading to a directed movement of fish. On the contrary weak advection goes with strong diffusion corresponding to a searching behavior. This formalizes the heuristic approach of Faugeras and Maury (2005).

The paper is structured as follows. In Section 2 we describe the random walk model. It is viewed as a simple IBM and simulations are conducted in Section 4. In Section 3, starting from the random walk, a recursion equation is formulated for the spatial density of individuals. This equation is expanded with respect to two small parameters and finally approximated to give the advection–diffusion equation. Section 4 provides numerical simulations of both the IBM and PDE model. Spatial statistics are computed in order to compare the models. It is shown that the PDE model is a good approximation of the IBM despite of the simplifying assumptions that are made to derive the PDE model. The paper ends with a Conclusion section and two Appendices.

2. Individual-based model

In this section we propose an IBM describing movements of n independent but identical fish. We assume that there is no interaction between individuals and that more than one individual can occupy a given position. Only horizontal movements are modeled and individuals evolve in a domain $\Omega \in \mathbb{R}^2$ during a period of time $[0, T]$. Oceanographic currents are not considered in this paper where flow effects are not taken into account. The modeling focusses on the biological processes which drive individual movements. It is assumed that individuals assess their environment and that the decisions they make concerning their movements depend on an habitat suitability index function

$$h : \mathbb{R}^2 \times [0, T] \to [0, 1].$$

The function h is supposed to synthesize all the informations (water temperature, forage concentration and dissolved oxygen concentration for example) that individuals take into account to adjust the direction and velocity of their displacements. Individuals are assumed to search for and stay in regions corresponding to a high habitat suitability index. Therefore, their movements are considered to be induced by their need to maximize h.

Each individual is characterized by its position $\mathbf{x}$ and has a velocity

$$\mathbf{v} = v\mathbf{d}.$$

The norm, v, of the velocity is assumed to be deterministic, whereas $\mathbf{d}$ is a stochastic unit-norm direction vector. An individual trajectory follows

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{v}.$$

This equation is discretized using the explicit Euler method, assuming there exists a small mean time, τ, during which the velocity vector of an individual is constant. Therefore, an individual positioned at $\mathbf{x}$ at time t will move to $\mathbf{x} + \mathbf{v}(\mathbf{x}, t)\tau$ at time $t + \tau$.

The behavior of each individual is governed by the habitat suitability index h and its gradient ∇h. A simple linear relation is assumed between the norm of the velocity and the habitat suitability index

$$v(\mathbf{x}, t) = v_0(1 - h(\mathbf{x}, t)), \tag{1}$$

where v_0 is the maximum speed that a fish can reach. Hence fish located in regions where h is low have a higher velocity than those in regions where h is large.

The direction vector is given by

$$\mathbf{d} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}. \tag{2}$$

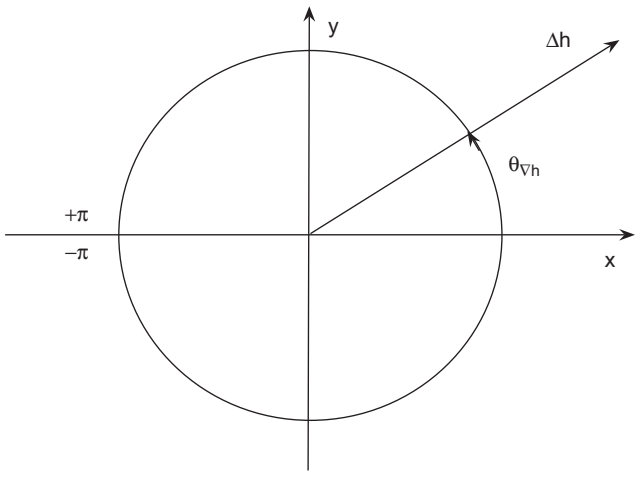

Fig. 1. The unit circle, ∇h and $\theta_{\nabla h} \in]-\pi, \pi]$.

The angle θ is a realization of a random variable Θ which follows a von Mises distribution, g defined by

$$g(\theta, \kappa, \theta_0) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta - \theta_0)),$$

where I_0 is the modified Bessel function of the first kind and of order 0 (see Appendix A). The distribution is centered around the mean angle $\theta_0 = \theta_{\nabla h} \in]-\pi, \pi]$ given by the direction of the gradient ∇h (see Fig. 1) and with a concentration parameter $\kappa = \alpha \|\nabla h\|$ proportional to the norm of the gradient. Hence, at time t, the angle of displacement of a fish located at $\mathbf{x}$ is drawn from the probability density

$$\begin{aligned} f(\theta, \mathbf{x}, t) &= g(\theta, \alpha\|\nabla h(\mathbf{x}, t)\|, \theta_{\nabla h}(\mathbf{x}, t)) \\ &= \frac{1}{2\pi I_0(\alpha\|\nabla h(\mathbf{x}, t)\|)} \exp(\alpha\|\nabla h(\mathbf{x}, t)\| \\ &\quad \times \cos(\theta - \theta_{\nabla h}(\mathbf{x}, t))). \end{aligned} \tag{3}$$

Since the mean movement direction is given by the direction of the gradient ∇h, fish tend to maximize the habitat suitability index h. The concentration parameter is proportional to $\|\nabla h\|$ and therefore high values of $\|\nabla h\|$ induce direction vectors that strongly follow the direction of the gradient, corresponding to a directed movement behavior. On the contrary low values of $\|\nabla h\|$ lead to less correlation between the direction vector and $\|\nabla h\|$. This corresponds to a searching behavior.

3. Approximation of the IBM: advection–diffusion equation

In this section, starting from the microscopic description of movements given by the IBM we formally derive a simplified macroscopic description in terms of an advection–diffusion PDE.

In order to achieve this task we have to use approximating hypothesis. The first one is to consider in a first step that the norm v of the velocity vector for each individual is a constant, that is to say independent of time and space. As a consequence we can suppose that at each time step τ an individual moves a distance δ in a direction θ with a probability which depends on space and time through the density $f(\theta, \mathbf{x}, t)$ of Eq. (3). The microscopic space and time scale parameters, δ and τ, are considered to be small with respect to the macroscopic space and time scales defined by the dimensions of the spatial domain Ω and the time domain $(0, T)$.

All individuals that can possibly reach position $\mathbf{x} = (x, y)$ at time $t + \tau$ lie at time t on a circle of radius δ centered on (x, y) (see Fig. 2). The density of individuals, $p(x, y, t + \tau)$ at position (x, y) and time $t + \tau$ can thus be expressed with the following recursion:

$$\begin{aligned} &p(x, y, t + \tau) \\ &\quad = \int_{-\pi}^{\pi} p(x + \delta\cos\theta, y + \delta\sin\theta, t) \\ &\qquad \times f(\theta + \pi, x + \delta\cos\theta, y + \delta\sin\theta, t)\,\mathrm{d}\theta \\ &\quad = \int_{-\pi}^{\pi} p(x - \delta\cos\theta, y - \delta\sin\theta, t) \\ &\qquad \times f(\theta, x - \delta\cos\theta, y - \delta\sin\theta, t)\,\mathrm{d}\theta. \end{aligned} \tag{4}$$

The remaining part of the derivation of the desired advection–diffusion equation from Eq. (4) relies on analytical computations which are fully detailed in Appendix B. It is based first of all on second order Taylor expansions with respect to the space variables for the right-hand side of Eq. (4) and with respect to the time variable for the left-hand side. Secondly the expansions are combined using recursive substitution and truncated neglecting high order terms. The results concerning the moments of the von Mises distribution given in Appendix A enable us to define a, b, c, d and e in the following way:

$$a = \int_{-\pi}^{\pi} f(\theta, x, y, t)\cos\theta\,\mathrm{d}\theta = \frac{I_1(\alpha\|\nabla h(\mathbf{x}, t)\|)}{I_0(\alpha\|\nabla h(\mathbf{x}, t)\|)} \cos\theta_{\nabla h}(\mathbf{x}, t),$$

$$b = \int_{-\pi}^{\pi} f(\theta, x, y, t)\sin\theta\,\mathrm{d}\theta = \frac{I_1(\alpha\|\nabla h(\mathbf{x}, t)\|)}{I_0(\alpha\|\nabla h(\mathbf{x}, t)\|)} \sin\theta_{\nabla h}(\mathbf{x}, t),$$

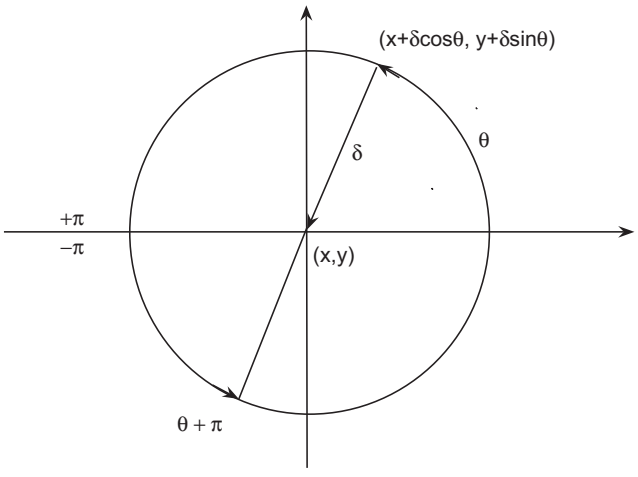

Fig. 2. All individuals that can possibly reach position (x, y) at time $t + \tau$ lie at time t on a circle of radius δ centered on (x, y).

$$c = \int_{-\pi}^{\pi} f(\theta, x, y, t)\cos^2\theta\,\mathrm{d}\theta = \frac{1}{2}\left(1 + \frac{I_2(\alpha\|\nabla h(\mathbf{x},t)\|)}{I_0(\alpha\|\nabla h(\mathbf{x},t)\|)}\cos 2\theta_{\nabla h}(\mathbf{x},t)\right),$$

$$d = \int_{-\pi}^{\pi} f(\theta, x, y, t)\,\sin\theta\,\cos\theta\,\mathrm{d}\theta = \frac{1}{2}\frac{I_2(\alpha\|\nabla h(\mathbf{x},t)\|)}{I_0(\alpha\|\nabla h(\mathbf{x},t)\|)}\sin 2\theta_{\nabla h}(\mathbf{x},t),$$

$$e = \int_{-\pi}^{\pi} f(\theta, x, y, t)\sin^2\theta\,\mathrm{d}\theta = \frac{1}{2}\left(1 - \frac{I_2(\alpha\|\nabla h(\mathbf{x},t)\|)}{I_0(\alpha\|\nabla h(\mathbf{x},t)\|)}\cos 2\theta_{\nabla h}(\mathbf{x},t)\right).$$

Eventually this leads to approximate Eq. (4) by the following advection–diffusion PDE:

$$\begin{aligned}\partial_t p = &-\left[\partial_x\left(\frac{\delta}{\tau}ap\right) + \partial_y\left(\frac{\delta}{\tau}bp\right)\right] \\ &+\left[\partial_x\left(\frac{\delta^2}{2\tau}(c-a^2)\partial_x p\right) + \partial_x\left(\frac{\delta^2}{2\tau}(d-ab)\partial_y p\right)\right. \\ &\left.+\partial_y\left(\frac{\delta^2}{2\tau}(d-ab)\partial_x p\right) + \partial_y\left(\frac{\delta^2}{2\tau}(e-b^2)\partial_y p\right)\right]. \qquad (5)\end{aligned}$$

Now, as was the case in the IBM, we assume that δ is not a constant but satisfies

$$\frac{\delta}{\tau} = v_0(1-h)$$

and hence we also have that

$$\frac{\delta^2}{2\tau} = \frac{\tau}{2}(v_0(1-h))^2.$$

Finally defining the advection velocity

$$\begin{aligned}\mathbf{V} &= v_0(1-h)\begin{pmatrix} a \\ b\end{pmatrix} \\ &= v_0(1-h)\frac{I_1(\alpha\|\nabla h\|)}{I_0(\alpha\|\nabla h\|)}\begin{pmatrix}\cos\theta_{\nabla h} \\ \sin\theta_{\nabla h}\end{pmatrix} \qquad (6)\end{aligned}$$

and the diffusion matrix

$$\begin{aligned}\mathbf{D} &= \frac{\tau}{2}(v_0(1-h))^2\begin{pmatrix} c-a^2 & d-ab \\ d-ab & e-b^2\end{pmatrix} \\ &= \frac{\tau}{2}(v_0(1-h))^2\left\{\frac{1}{2}\left(1 - \frac{I_2(\alpha\|\nabla h\|)}{I_0(\alpha\|\nabla h\|)}\right)\begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right. \\ &\quad + \left[\frac{I_2(\alpha\|\nabla h\|)}{I_0(\alpha\|\nabla h\|)} - \left(\frac{I_1(\alpha\|\nabla h\|)}{I_0(\alpha\|\nabla h\|)}\right)^2\right] \\ &\quad \left.\times\begin{pmatrix}\cos^2\theta_{\nabla h} & \sin\theta_{\nabla h}\cos\theta_{\nabla h} \\ \sin\theta_{\nabla h}\cos\theta_{\nabla h} & \sin^2\theta_{\nabla h}\end{pmatrix}\right\} \qquad (7)\end{aligned}$$

we obtain the final advection–diffusion equation approximating the IBM:

$$\partial_t p = \nabla\cdot(\mathbf{D}\nabla p - \mathbf{V}p). \qquad (8)$$

Note that the diffusion matrix, $\mathbf{D}$ has non-zero off-diagonal terms and its elements are the centered second order trigonometric moments of the von Mises distribution. $\mathbf{D}$ is also symmetric positive.

The advection velocity $\mathbf{V}$ is of chemotaxis type. It is oriented in the direction of ∇h and its amplitude is modulated by $I_1(\alpha\|\nabla h\|)/I_0(\alpha\|\nabla h\|)$ an increasing function of $\|\nabla h\|$. At a given level of habitat suitability index h, the balance between advection and diffusion only depends on the gradient $\|\nabla h\|$. Strong gradients impose strong advection and weak diffusion, whereas weak gradients induce weak advection and strong diffusion.

4. Numerical simulations and comparisons of the models

The IBM proposed in Section 2 is approximated by an advection–diffusion equation derived in Section 3. In this section, we conduct numerical simulations for both models and compute some spatial statistics in order to compare them.

An algorithm to simulate the IBM described in Section 2 is not difficult to program. We consider a rectangular spatial domain $\Omega = (0, L_x)\times(0, L_y)$ large enough for individuals never to reach its boundaries during the period of the simulation of length $T = K\tau$. For the sake of simplification we consider a time independent habitat suitability index h defined on Ω. As initial condition a set of $n = 10^4$ individuals are positioned at the same location: $\mathbf{x}_i^0 = \mathbf{x}^0$, $i = 1,\ldots,n$. At each time step (denoted by k), of length τ, and for each point $\mathbf{x}_i^k$, $h(\mathbf{x}_i^k)$ and $\nabla h(\mathbf{x}_i^k)$ are computed, and an angle θ_i^k is drawn from the von Mises distribution. Each individual then moves according to

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + v_0(1-h(\mathbf{x}_i^k))\tau\begin{pmatrix}\cos\theta_i^k \\ \sin\theta_i^k\end{pmatrix}.$$

All experiments were conducted with $\tau = 10^{-1}$ and $v_0 = 1$. The IBM simulation algorithm was programmed with Matlab. In order to generate random numbers from a von Mises distribution we used a Matlab code developed by A. Bar-Guy and A. Podgaetsky available on Matlab central website. It implements the method suggested in Yuan and Kalbleisch (2000) and described in Devroye (2002).

In the approximation procedure of the PDE model a finite difference discretization is used. Eq. (8) is solved on a grid with a spatial resolution of $\Delta x = \Delta y = 10^{-2}$ and a discrete time step $\Delta t = 10^{-2}$ is used. Since Ω is bounded, boundary conditions need to be added to Eq. (8). We have used Neumann boundary conditions. In order to be consistent with the simulation of the IBM we consider the following initial condition:

$$p^0 = \frac{n}{\Delta x\Delta y}\delta_{\mathbf{x}^0}.$$

The numerical scheme implemented is based on a splitting method (Strang, 1968; Marchuk, 1990). Diagonal diffusion terms are treated implicitly in time, whereas off-diagonal diffusion terms are treated explicitly. Concerning advection terms, the MUSCL scheme (monotonic upstream centered scheme for conservation laws (Van Leer, 1977)) is used. The choice of the advection scheme has an important part in the spatial statistics computed from the solution of the PDE model. The numerical diffusion introduced into the solution of the PDE by the use of a simple upwind or centered difference advection scheme leads to unreliable computed variances. The MUSCL scheme is more complicated to implement but far less diffusive.

In order to compare the population distributions generated by the IBM on the one hand and by the PDE approximating model on the other, we compute the first three centered spatial moments as functions of time. For the PDE model the mean x position is computed as

$$m_{x,PDE}(t) = \frac{\int_\Omega xp(x,y,t)\,\mathrm{d}x\,\mathrm{d}y}{\int_\Omega p(x,y,t)\,\mathrm{d}x\,\mathrm{d}y} \tag{9}$$

and for the IBM model it is computed as

$$m_{x,IBM}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t). \tag{10}$$

The variance about the mean x position is

$$\sigma^2_{x,PDE}(t) = \frac{\int_\Omega (x - m_{x,PDE}(t))^2 p(x,y,t)\,\mathrm{d}x\,\mathrm{d}y}{\int_\Omega p(x,y,t)\,\mathrm{d}x\,\mathrm{d}y} \tag{11}$$

for the PDE and

$$\sigma^2_{x,IBM}(t) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i(t) - m_{x,IBM}(t))^2 \tag{12}$$

for the IBM. Finally the third standardized centered moment or skewness in x is computed as

$$\gamma_{x,PDE}(t) = \frac{\int_\Omega (x - m_{x,PDE}(t))^3 p(x,y,t)\,\mathrm{d}x\,\mathrm{d}y}{\sigma^3_{x,PDE}(t)\int_\Omega p(x,y,t)\,\mathrm{d}x\,\mathrm{d}y} \tag{13}$$

for the PDE and as

$$\gamma_{x,IBM}(t) = \frac{\sqrt{n(n-1)}}{n-2}\frac{\sqrt{n}\sum_{i=1}^{n}(x_i(t) - m_{x,IBM}(t))^3}{(\sum_{i=1}^{n}(x_i(t) - m_{x,IBM}(t))^2)^{3/2}}. \tag{14}$$

Similar formulas are used to compute moments in the y direction.

4.1. Experiments 1 and 2

Both models are run on a time interval of length $T = 2$ and on a spatial domain $\Omega = (0,2)\times(0,2)$. The initial position of all individuals is $\mathbf{x}^0 = (0.6, 1)$. The habitat suitability index function is $h(x,y,t) = x/2$ and therefore the gradient is oriented along the x-axis. The only difference between both experiments is the value of the concentration parameter, $\alpha = 5$ in experiment 1 and $\alpha = 1$ in experiment 2.

Figs. 3 and 7 show the solutions of both models at $T = 2$. Due to the different values of the concentration parameter individuals are more scattered in experiment 2 than in experiment 1. As a consequence, advection in the direction of the gradient ∇h is stronger in experiment 1 than in experiment 2.

The mean positions, $m_{x,PDE}$, $m_{x,IBM}$ and $m_{y,PDE}$, $m_{y,IBM}$ are plotted in Fig. 4 as functions of time for experiment 1. The solution of the PDE model appears to follow closely the solution of the IBM. This is also true for experiment 2 as shown in Fig. 8 although, because of a strongest diffusion, the difference between $m_{x,PDE}$, $m_{x,IBM}$ at final time is highest in experiment 2 than in experiment 1.

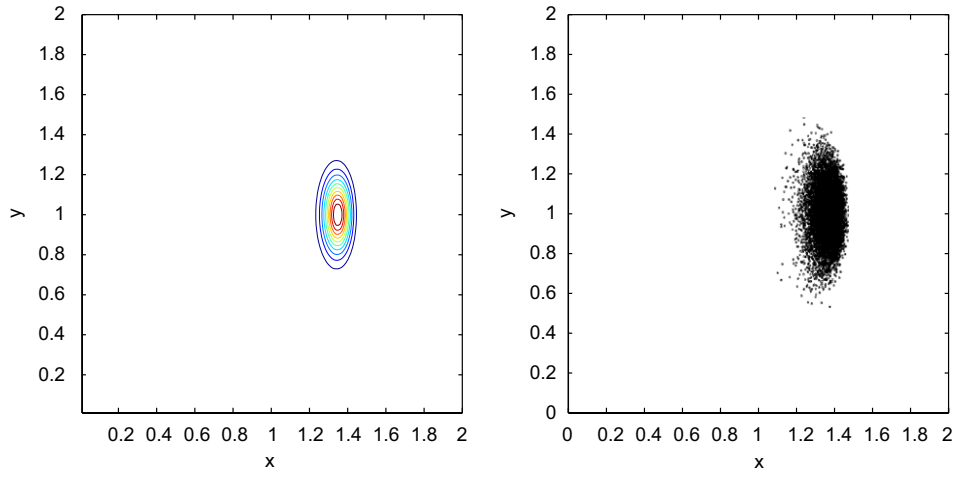

Fig. 3. Solution of the PDE model (left) and of the IBM (right) at final time $T = 2$ for experiment 1. The spatial domain is defined by $\Omega = (0,2)\times(0,2)$. The habitat suitability index function is $h(x,y,t) = x/2$. The initial position of all individuals is $\mathbf{x}^0 = (0.6, 1)$. The concentration parameter is $\alpha = 5$.

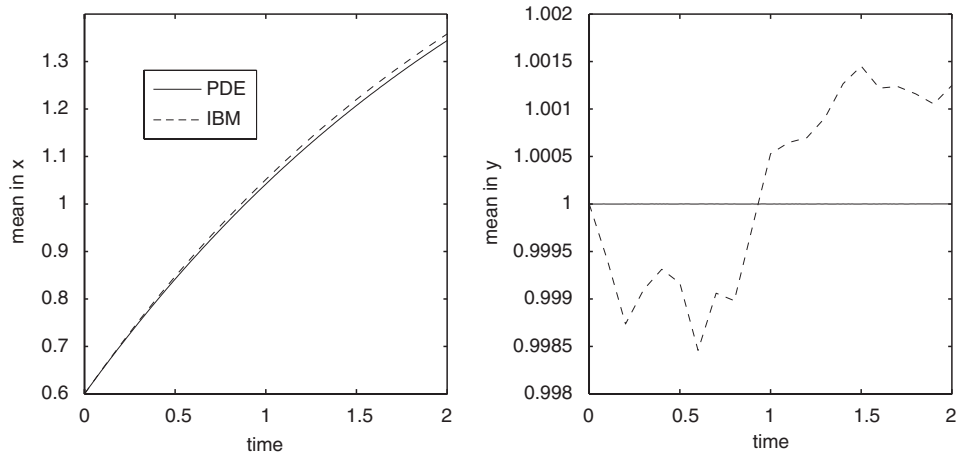

Fig. 4. Experiment 1. Mean position in x, $m_{x,PDE}(t)$ for the PDE model and $m_{x,IBM}(t)$ for the IBM (left) and mean position in y, $m_{y,PDE}(t)$ for the PDE and $m_{y,IBM}(t)$ for the IBM (right).

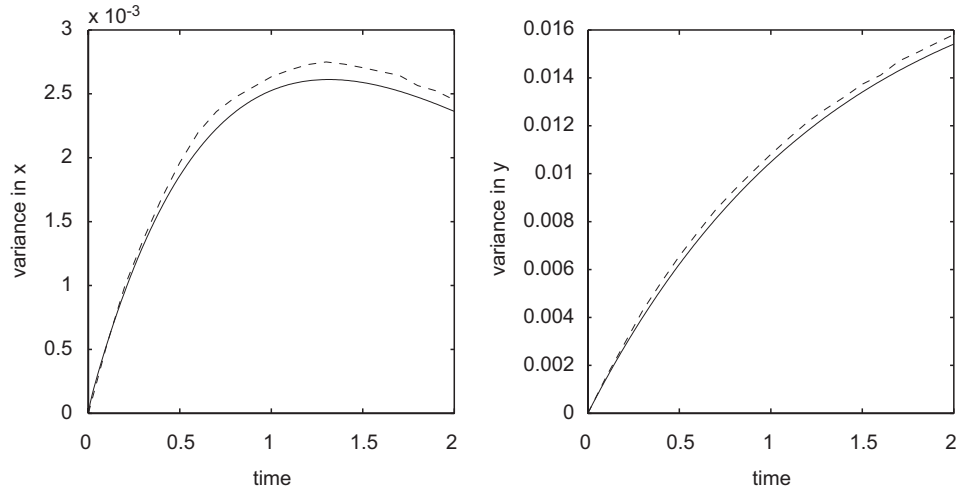

Fig. 5. Experiment 1. Variance in x, $\sigma^2_{x,PDE}(t)$ for the PDE model and $\sigma^2_{x,IBM}(t)$ for the IBM (left) and variance in y, $\sigma^2_{y,PDE}(t)$ for the PDE and $\sigma^2_{y,IBM}(t)$ for the IBM (right).

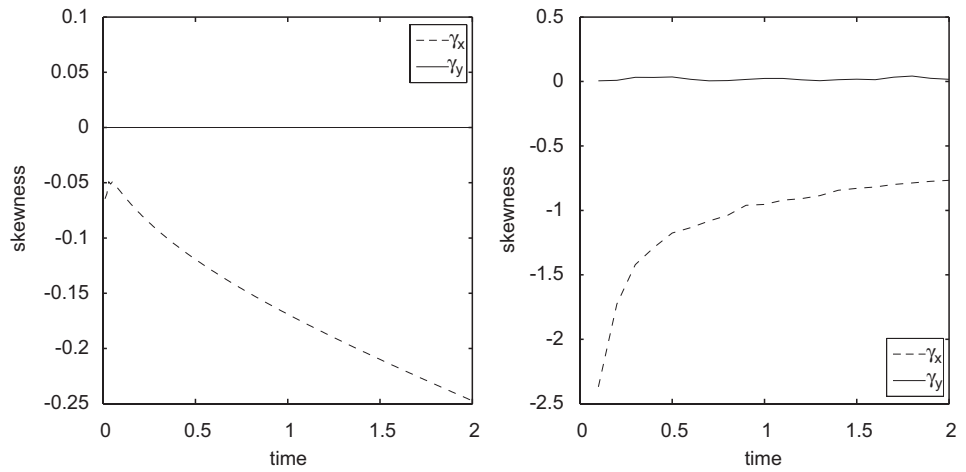

Fig. 6. Experiment 1. Skewness in x and y for the PDE model (left) and for the IBM (right).

The same type of conclusion, that is to say the solution of the PDE model follows closely the solution of the IBM, remains true for the second order centered moments as shown in Figs. 5 and 9. This is not surprising since in the PDE approximation of the IBM (described in Section 3) second order derivative terms

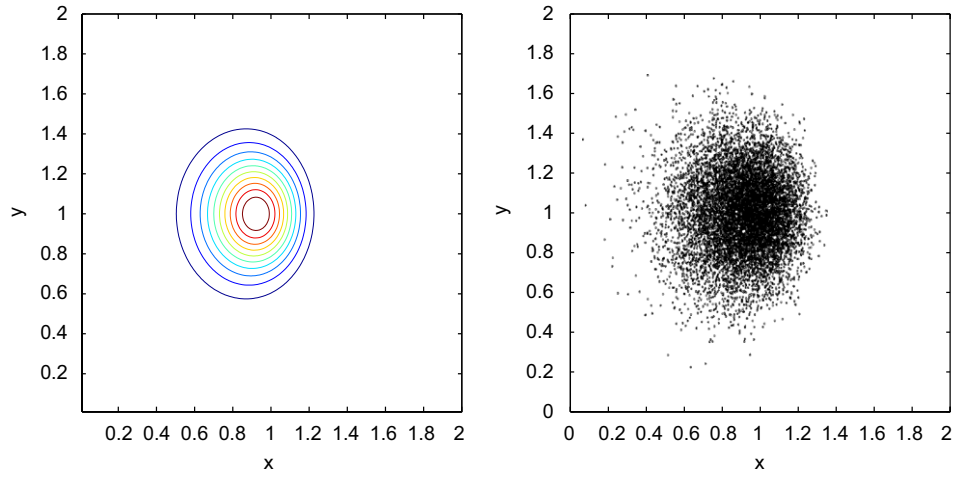

Fig. 7. Solution of the PDE model (left) and of the IBM (right) at final time $T = 2$ for experiment 2. The spatial domain is defined by $\Omega = (0, 2) \times (0, 2)$. The habitat suitability index function is $h(x, y, t) = x/2$. The initial position of all individuals is $\mathbf{x}^0 = (0.6, 1)$. The concentration parameter is $\alpha = 1$.

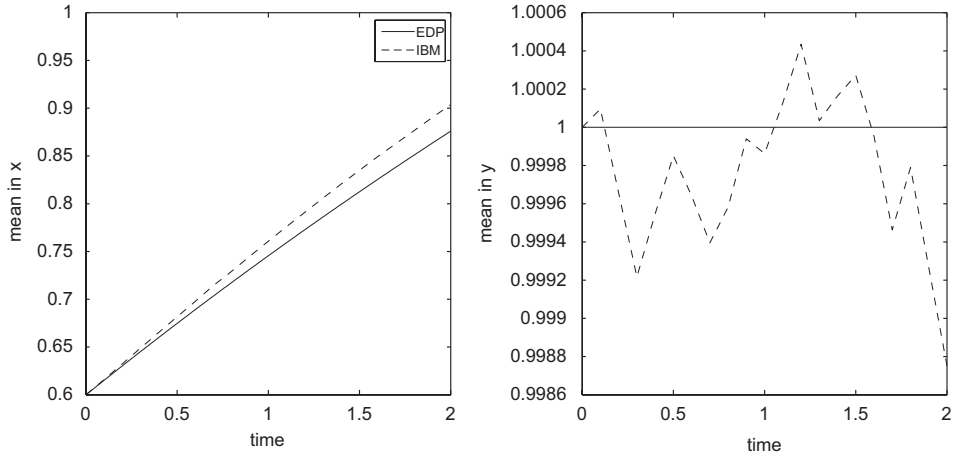

Fig. 8. Experiment 2. Mean position in x, $m_{x,PDE}(t)$ for the PDE model and $m_{x,IBM}(t)$ for the IBM (left) and mean position in y, $m_{y,PDE}(t)$ for the PDE and $m_{y,IBM}(t)$ for the IBM (right).

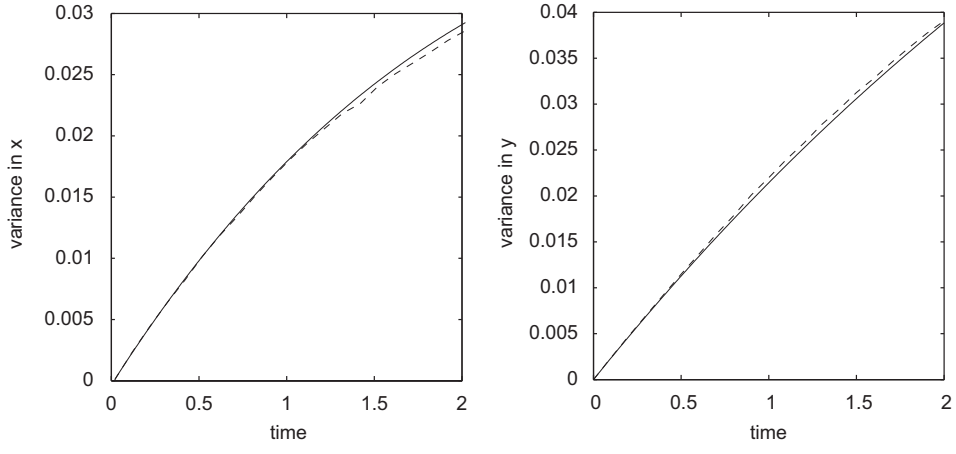

Fig. 9. Experiment 2. Variance in x, $\sigma^2_{x,PDE}(t)$ for the PDE model and $\sigma^2_{x,IBM}(t)$ for the IBM (left) and variance in y, $\sigma^2_{y,PDE}(t)$ for the PDE and $\sigma^2_{y,IBM}(t)$ for the IBM (right).

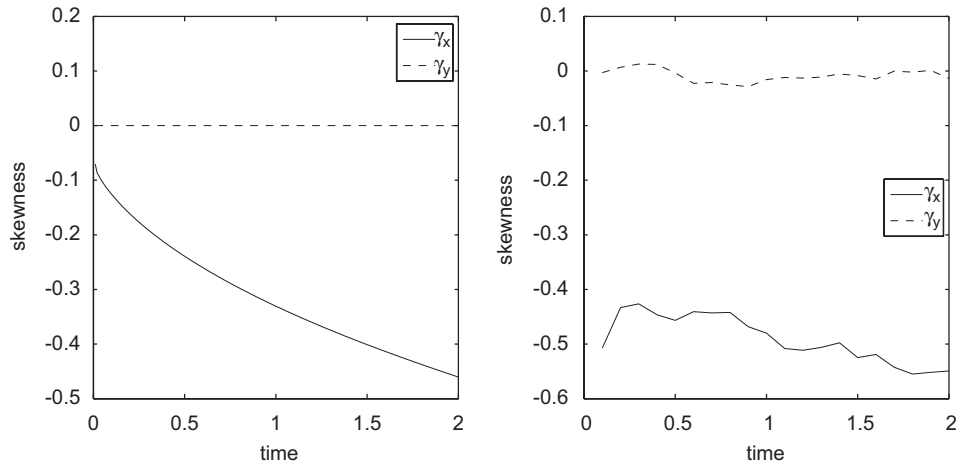

Fig. 10. Experiment 2. Skewness in x and y for the PDE model (left) and for the IBM (right).

are taken into account (although some of them are neglected).

As could be expected the numerical results for the third order moments, or skewness, show much less correlation between the PDE model and the IBM (Figs. 6 and 10) than for the first two moments. Nevertheless some features of the IBM still appear in the PDE solution. In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. A distribution has positive skew (right-skewed) if the right (higher value) tail is longer or fatter and negative skew (left-skewed) if the left (lower value) tail is longer or fatter. In both experiments the skewness in the y direction, γ_y, is null for both the IBM and the PDE, indicating that the distribution is symmetrical about the x-axis. Although values are different the skewness in the x direction, γ_x, is negative for both the IBM and the PDE indicating that the distributions have a longer "left" tail. This tail reflects the possibility for an individual not to move at each time step in the direction of the gradient with some probability depending on α. This probability is higher in experiment 2 than in experiment 1 and therefore the tail is bigger in experiment 2 than in experiment 1 (see also Figs. 7–10).

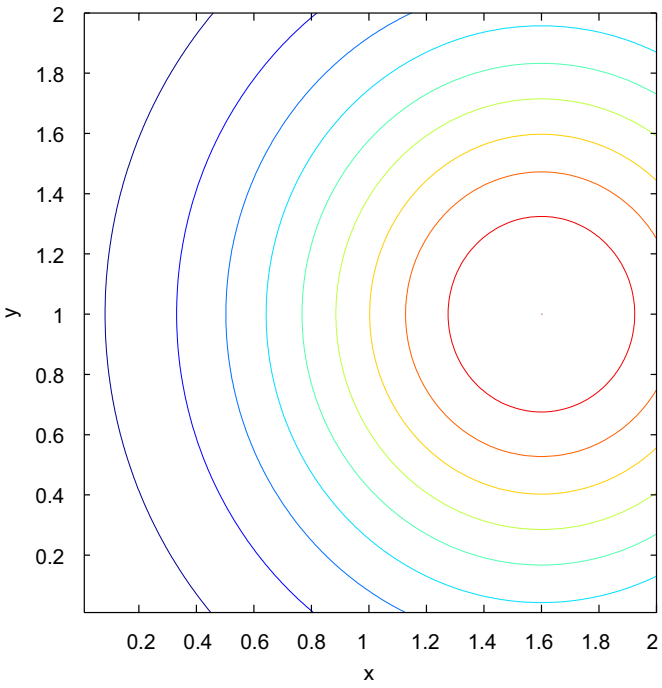

Fig. 11. Experiment 3. Contour plot of the habitat suitability index function $h(x,y) = \exp(-((x-1.6)^2 + (y-1)^2))$.

4.2. Experiment 3

In this experiment the habitat suitability function is (see Fig. 11)

$$h(x,y,t) = \exp(-((x-1.6)^2 + (y-1)^2)).$$

The initial position of all individuals is $\mathbf{x}^0 = (0.6, 0.6)$ and the concentration parameter is $\alpha = 2$. Fig. 12 shows the time evolution of the solution for both the PDE model and the IBM. This experiment illustrates the effect of non-zero off-diagonal terms in the diffusion matrix. The solution of the PDE model is not symmetrical with respect to the x- or y-axis. Figs. 13–15 show the time evolution of the first three moments for both the PDE model and the IBM. As for experiments 1 and 2 the first two moments of the solution of the approximated PDE model closely follow those of the IBM. Differences appear in the computation of the third moment.

5. Conclusions

In this paper, we provide a mechanistic approach to derive an advection–diffusion PDE modeling fish population movements. This PDE, Eqs. (6)–(8), describes the time and space evolution of the density of individuals. This study formalizes and improves the heuristic approaches of former papers dedicated to fish dynamics population modeling (Bertignac et al., 1998; Maury and Gascuel,

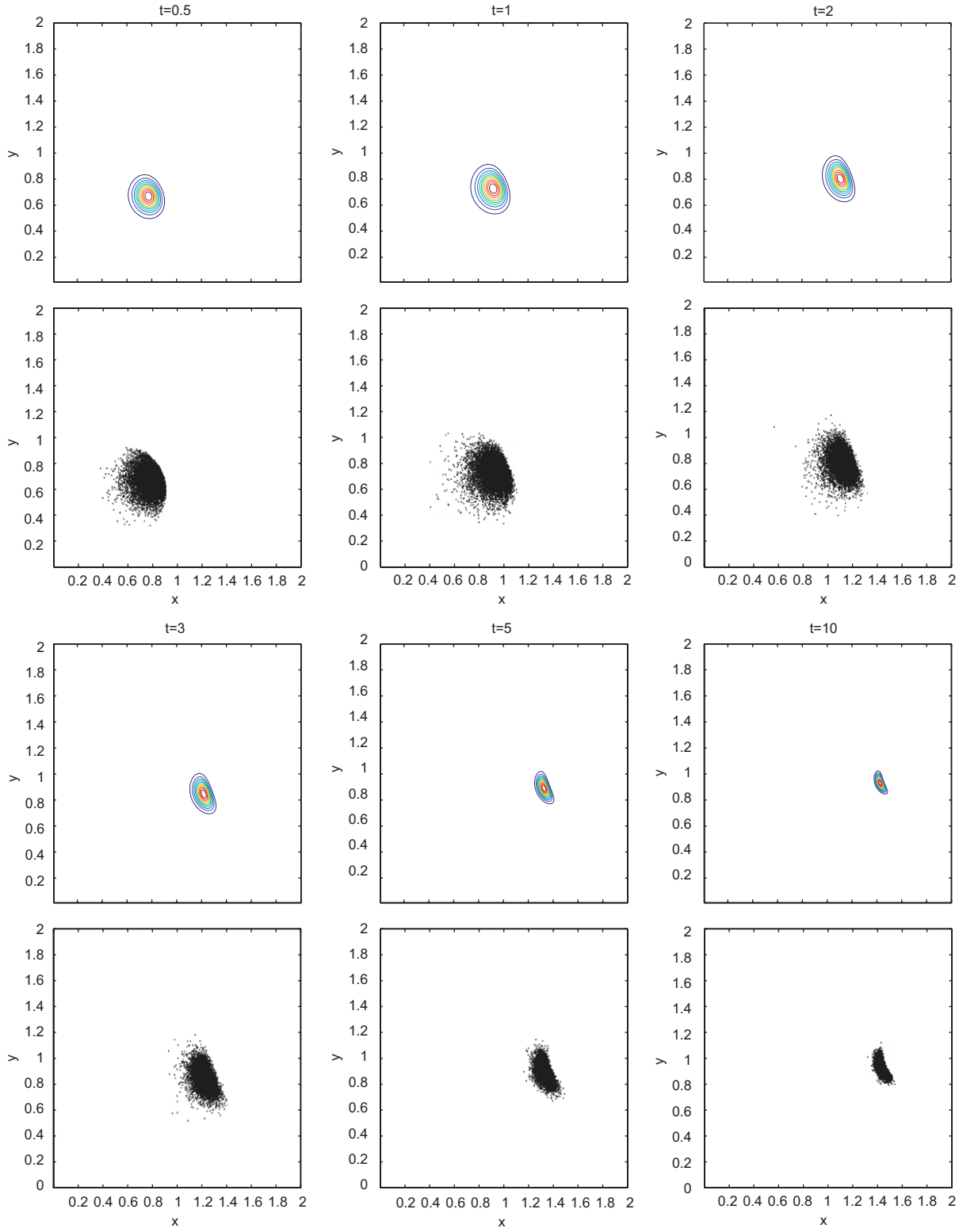

Fig. 12. Solution of the PDE model (rows 1 and 3) and of the IBM (rows 2 and 4) at time $t = 0.5$, 1, 2, 3, 5 and final time $T = 10$ for experiment 3. The spatial domain is defined by $\Omega = (0,2) \times (0,2)$. The habitat suitability index function is $h(x,y,t) = \exp(-((x-1.6)^2 + (y-1)^2))$ (see Fig. 11). The initial position of all individuals is $\mathbf{x}^0 = (0.6, 0.6)$. The concentration parameter is $\alpha = 2$.

1999; Maury, 2000; Maury et al., 2001; Sibert et al., 1999; Lehodey et al., 2003; Faugeras and Maury, 2005). The obtained formulation of advection and diffusion terms arises from a simple IBM, or biased random walk model, including hypotheses on individual fish movements. This formulation induces a balance between a directed movement behavior (strong advection and weak diffusion) and a searching behavior (weak advection

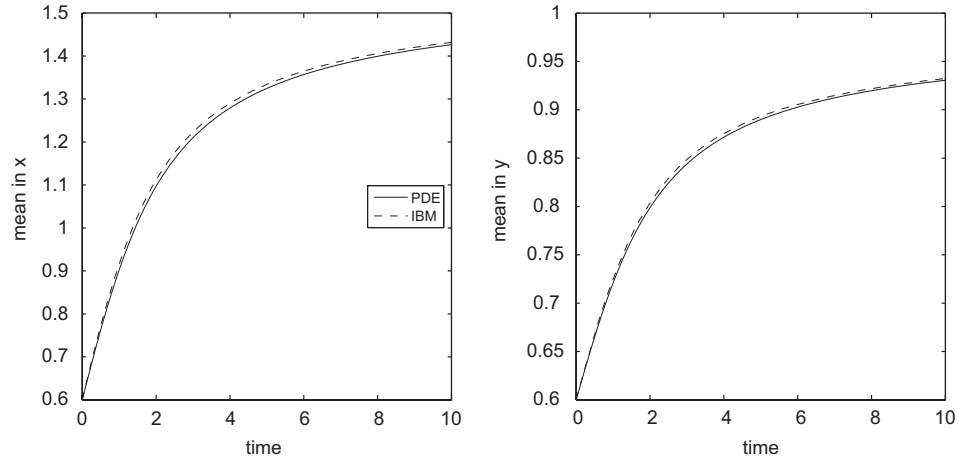

Fig. 13. Experiment 3. Mean position in x, $m_{x,PDE}(t)$ for the PDE model and $m_{x,IBM}(t)$ for the IBM (left) and mean position in y, $m_{y,PDE}(t)$ for the PDE and $m_{y,IBM}(t)$ for the IBM (right).

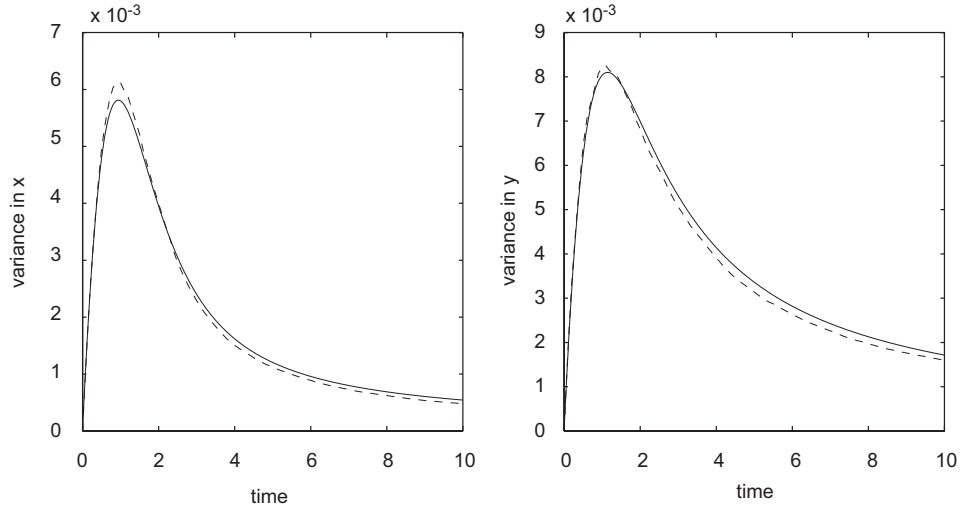

Fig. 14. Experiment 3. Variance in x, $\sigma^2_{x,PDE}(t)$ for the PDE model and $\sigma^2_{x,IBM}(t)$ for the IBM (left) and variance in y, $\sigma^2_{y,PDE}(t)$ for the PDE and $\sigma^2_{y,IBM}(t)$ for the IBM (right).

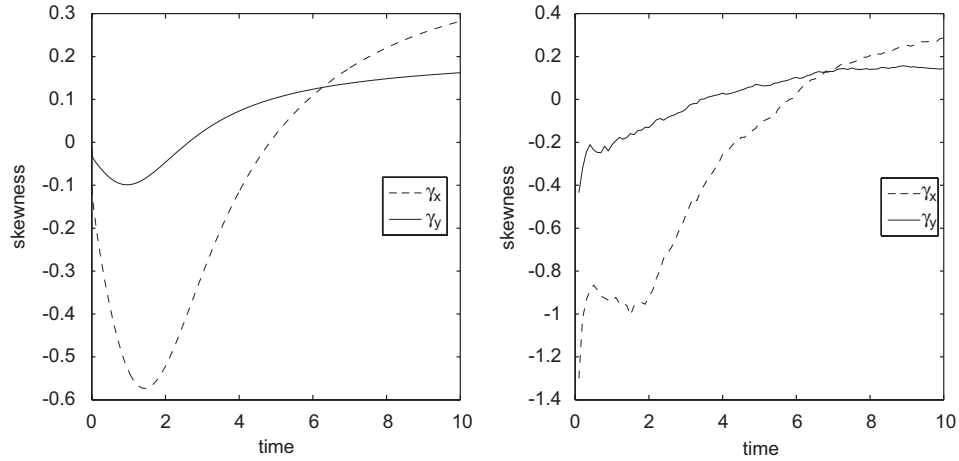

Fig. 15. Experiment 3. Skewness in x and y for the PDE model (left) and for the IBM (right).

and strong diffusion). We show through numerical experiments that the PDE model is a good approximation of the IBM.

We think that such a model, particularly thanks to the full diffusion matrix, will be able to improve the representation of the anisotropy of fish population move-

ments in an inhomogeneous and variable environment. This will be tested in an ongoing work in which a more complete version of the model, including oceanographic currents and a size structure of the population, will be confronted to fishing and tag-recapture data for tuna populations in the Indian ocean.

Appendix A. The von Mises distribution

The von Mises distribution or circular normal distribution is a continuous probability distribution describing the distribution of a random variable with period 2π. A reference for directional statistics is for example Mardia and Jupp (1999).

Its expression for an angle θ is

$$g(\theta, a, \theta_0) = \frac{1}{2\pi I_0(a)} \exp(a\cos(\theta - \theta_0)),$$

where I_0 denotes the modified Bessel function of the first kind and order 0. I_n the modified Bessel function of the first kind and order $n \geqslant 0$ is defined by

$$I_n(a) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{a\cos\theta}\cos n\theta \,\mathrm{d}\theta.$$

The parameter θ_0 is the mean angle and the parameter $a \geqslant 0$ is the concentration parameter. The distribution is unimodal and is symmetrical about $\theta = \theta_0$. The mode is at $\theta = \theta_0$. When $a = 0$ the von Mises distribution equals the uniform distribution and as $a \to \infty$ the distribution becomes sharply peaked about the mean angle θ_0.

The moments of the von Mises distribution are usually computed as the moments of $z = \mathrm{e}^{\mathrm{i}\theta}$ rather than the angle θ itself. These moments are referred to as circular moments and read

$$\langle z^n \rangle = \int_{-\pi}^{\pi} z^n g(\theta, a, \theta_0)\,\mathrm{d}\theta = \frac{I_n(a)}{I_0(a)}\mathrm{e}^{n\theta_0}.$$

Appendix B. Full derivation of the advection–diffusion equation

The density of individuals at position $\mathbf{x} = (x, y)$ and time $t + \tau$ satisfies:

$$\begin{aligned} p(x, y, t+\tau) = & \int_{-\pi}^{\pi} p(x + \delta\cos\theta, y + \delta\sin\theta, t) \\ & \times f(\theta + \pi, x + \delta\cos\theta, y + \delta\sin\theta, t)\,\mathrm{d}\theta \\ = & \int_{-\pi}^{\pi} p(x - \delta\cos\theta, y - \delta\sin\theta, t) \\ & \times f(\theta, x - \delta\cos\theta, y - \delta\sin\theta, t)\,\mathrm{d}\theta. \end{aligned} \tag{B.1}$$

A second order Taylor expansion of the integrand in Eq. (B.1) leads to

$$\begin{aligned} & p(x - \delta\cos\theta, y - \delta\sin\theta, t) f(\theta, x - \delta\cos\theta, y - \delta\sin\theta, t) \\ & \quad = p(x, y, t) f(\theta, x, y, t) \\ & \qquad - \delta[\partial_x(p(x, y, t) f(\theta, x, y, t))\cos\theta \\ & \qquad + \partial_y(p(x, y, t) f(\theta, x, y, t))\sin\theta] \\ & \qquad + \frac{\delta^2}{2}[\partial_x^2(p(x, y, t) f(\theta, x, y, t))\cos^2\theta \\ & \qquad + 2\partial_x\partial_y(p(x, y, t) f(\theta, x, y, t))\sin\theta\cos\theta \\ & \qquad + \partial_y^2(p(x, y, t) f(\theta, x, y, t))\sin^2\theta] + \mathcal{O}(\delta^3). \end{aligned} \tag{B.2}$$

The evolution equation (B.1) becomes integrating Eq. (B.2) over $(-\pi, \pi)$:

$$\begin{aligned} p(x, y, t+\tau) = & p(x, y, t) - \delta[\partial_x(ap)(x, y, t) + \partial_y(bp)(x, y, t)] \\ & + \frac{\delta^2}{2}[\partial_x^2(cp)(x, y, t) + 2\partial_x\partial_y(dp)(x, y, t) \\ & + \partial_y^2(ep)(x, y, t)] + \mathcal{O}(\delta^3). \end{aligned} \tag{B.3}$$

Eq. (B.3) is a Kramers–Moyall expansion. Its left-hand side can also be expanded to

$$\begin{aligned} p(x, y, t+\tau) = & p(x, y, t) + \tau\partial_t p(x, y, t) \\ & + \frac{\tau^2}{2}\partial_t^2 p(x, y, t) + \mathcal{O}(\tau^3). \end{aligned} \tag{B.4}$$

From Eq. (B.4) and (B.3) we obtain

$$\begin{aligned} \partial_t p(x, y, t) = & -\frac{\delta}{\tau}[\partial_x(ap)(x, y, t) + \partial_y(bp)(x, y, t)] \\ & + \frac{\delta^2}{2\tau}[\partial_x^2(cp)(x, y, t) + 2\partial_x\partial_y(dp)(x, y, t) \\ & + \partial_y^2(ep)(x, y, t)] + \mathcal{O}\left(\frac{\delta^3}{\tau}\right) + \mathcal{O}(\tau) \end{aligned} \tag{B.5}$$

and

$$\begin{aligned} \partial_t p(x, y, t) = & -\frac{\delta}{\tau}[\partial_x(ap)(x, y, t) + \partial_y(bp)(x, y, t)] \\ & + \frac{\delta^2}{2\tau}[\partial_x^2(cp)(x, y, t) + 2\partial_x\partial_y(dp)(x, y, t) \\ & + \partial_y^2(ep)(x, y, t)] - \frac{\tau}{2}\partial_t^2 f(x, y, t) + \mathcal{O}\left(\frac{\delta^3}{\tau}\right) \\ & + \mathcal{O}(\tau^2). \end{aligned} \tag{B.6}$$

We now use a recursive substitution method in the Kramers–Moyall expansion in order to rewrite the last term of Eq. (B.6). Differencing Eq. (B.5) with respect to t and multiplying by $\tau/2$ leads to

$$\begin{aligned} \frac{\tau}{2}\partial_t^2 p(x, y, t) = & -\frac{\delta}{2}[\partial_x\partial_t(ap)(x, y, t) + \partial_y\partial_t(bp)(x, y, t)] \\ & + \mathcal{O}(\delta^2) + \mathcal{O}(\tau^2). \end{aligned} \tag{B.7}$$

Using identities such as $\partial_t(uf) = u\partial_t f + f\partial_t u$, we can reinject Eq. (B.5) into Eq. (B.7) and obtain

$$\begin{aligned} \frac{\tau}{2}\partial_t^2 p = & -\frac{\delta}{2\tau}[\partial_x^2(a^2 p) + 2\partial_x\partial_y(abp) + \partial_y^2(b^2 p)] \\ & - \frac{\delta^2}{2\tau}[\partial_x((\partial_x a)ap) + \partial_x((\partial_y a)bp) \\ & + \partial_y((\partial_x b)ap) + \partial_y((\partial_y b)bp)] \end{aligned}$$

$$- \frac{\delta}{2}[\partial_x(p(\partial_t a)) + \partial_y(p(\partial_t b))] + \mathcal{O}\left(\frac{\delta^3}{\tau}\right) + \mathcal{O}(\delta\tau) + \mathcal{O}(\delta^2) + \mathcal{O}(\tau^2). \quad \text{(B.8)}$$

Reinjecting Eq. (B.8) into Eq. (B.6) leads to

$$\begin{aligned}\partial_t p = & -\frac{\delta}{\tau}[\partial_x(ap) + \partial_y(bp)] \\ & + \frac{\delta^2}{2\tau}[\partial_x^2((c-a^2)p) + 2\partial_x\partial_y((d-ab)p) + \partial_y^2((e-b^2)p)] \\ & + \frac{\delta^2}{2\tau}[\partial_x((\partial_x a)ap) + \partial_x((\partial_y a)bp) + \partial_y((\partial_x b)ap) \\ & + \partial_y((\partial_y b)bp)] + \frac{\delta}{2}[\partial_x(p(\partial_t a)) + \partial_y(p(\partial_t b))] \\ & + \mathcal{O}\left(\frac{\delta^3}{\tau}\right) + \mathcal{O}(\delta\tau) + \mathcal{O}(\delta^2) + \mathcal{O}(\tau^2) \end{aligned} \quad \text{(B.9)}$$

and finally to

$$\begin{aligned}\partial_t p = & -\frac{\delta}{\tau}[\partial_x(ap) + \partial_y(bp)] \\ & + \frac{\delta^2}{2\tau}[\partial_x((c-a^2)\partial_x p) + \partial_x((d-ab)\partial_y p) \\ & + \partial_y((d-ab)\partial_x p) + \partial_y((e-b^2)\partial_y p)] \\ & + \frac{\delta^2}{2\tau}[\partial_x((\partial_x c - a\partial_x a)p) + \partial_x((\partial_y e - a\partial_y b)p) \\ & + \partial_y((\partial_x e - b\partial_x a)p) + \partial_y((\partial_y d - b\partial_y b)p)] \\ & + \frac{\delta}{2}[\partial_x(p(\partial_t a)) + \partial_y(p(\partial_t b))] \\ & + \mathcal{O}\left(\frac{\delta^3}{\tau}\right) + \mathcal{O}(\delta\tau) + \mathcal{O}(\delta^2) + \mathcal{O}(\tau^2). \end{aligned} \quad \text{(B.10)}$$

At this stage we use two approximations. The first one concerns the terms of lines 4–6 in Eq. (B.10) which are advection terms. In the following we neglect them assuming that the derivatives $\partial_x a$, $\partial_y b$, $\partial_x c$, $\partial_y d$, $\partial_x e$, $\partial_y e$, $\partial_t a$ and $\partial_t b$ are small. As is seen from the dependence of a, b, c, d and e on ∇h, this is possible if we consider that the function h is smooth with small second order derivatives.

The second approximation is to neglect the $\mathcal{O}(\delta^3/\tau)$, $\mathcal{O}(\delta\tau)$, $\mathcal{O}(\delta^2)$ and $\mathcal{O}(\tau^2)$ terms in Eq. (B.10). This leads to

$$\begin{aligned}\partial_t p = & -\left[\partial_x\left(\frac{\delta}{\tau}ap\right) + \partial_y\left(\frac{\delta}{\tau}bp\right)\right] \\ & + \left[\partial_x\left(\frac{\delta^2}{2\tau}(c-a^2)\partial_x p\right) + \partial_x\left(\frac{\delta^2}{2\tau}(d-ab)\partial_y p\right)\right. \\ & \left. + \partial_y\left(\frac{\delta^2}{2\tau}(d-ab)\partial_x p\right) + \partial_y\left(\frac{\delta^2}{2\tau}(e-b^2)\partial_y p\right)\right] \end{aligned} \quad \text{(B.11)}$$

which exactly is Eq. (5) of the text.

References

Alt, W., 1980. Biased random walk models for chemotaxis and related diffusion approximations. J. Math. Biol. 9, 147–177.

Bertignac, M., Lehodey, P., Hampton, J., 1998. A spatial population dynamics simulation model of tropical tunas using a habitat index based on environmental parameters. Fish. Oceanogr. 7 (3/4), 326–334.

Devroye, L., 2002. Simulating bessel random variables. Stat. Probab. Lett. 57, 249–257.

Faugeras, B., Maury, O., 2005. An advection–diffusion-reaction size-structured fish population dynamics model combined with a statistical parameter estimation procedure: application to the Indian Ocean skipjack tuna fishery. Math. Biosci. Eng. 2 (4), 719–741.

Flierl, G., Grünbaum, D., Levin, S., Olson, D., 1999. From individuals to aggregations: the interplay between behavior and physics. J. Theor. Biol. 196, 397–454.

Grünbaum, D., 1999. Advection–diffusion equations for generalized tactic searching behaviors. J. Math. Biol. 38, 169–194.

Holmes, E.E., Lewis, M.A., Banks, J.E., Veit, R.R., 1994. Partial differential equations in ecology: spatial interactions and population dynamics. Ecology 75 (1), 17–29.

Lehodey, P., Chai, F., Hampton, J., 2003. Modelling climate-related variability of tuna populations from a coupled ocean-biogeochemical-populations dynamics model. Fish. Oceanogr. 12 (4/5), 483–494.

Marchuk, G.I., 1990. Splitting and alternating direction methods. In: Handbook of Numerical Analysis, vol. I. North-Holland, Amsterdam, pp. 197–462.

Mardia, K., Jupp, P., 1999. Directional Statistics. Wiley, New York.

Maury, O., 2000. A habitat-based simulation framework to design tag-recapture experiments for tunas in the indian ocean. Application to the skipjack (*Katsuwonus pelamis*) population. Working Party on Tagging, Indian Ocean Tuna Commission, Victoria, Seychelles.

Maury, O., Gascuel, D., 1999. SHADIS (Simulateur HAlieutique de DYnamiques Spatiales), a GIS based numerical model of fisheries. Example application: the study of a marine protected area. Aquat. Living Resour. 12 (2), 77–88.

Maury, O., Gascuel, D., Fonteneau, A., 2001. Spatial modeling of Atlantic yellowfin tuna population dynamics: a habitat based advection–diffusion-reaction approach with application to the local overfishing study. In: Dorn, M. (Ed.), 17th Lowell Wakefield Symposium on Spatial Processes and Management of Fish Populations, Alaska Sea Grant College Program, Anchorage, 27–30 October 1999.

Okubo, A., 1980. Diffusion and Ecological Problems: Mathematical Models. Biomathematics, vol. 10. Springer, Berlin.

Risken, H., 1996. The Fokker–Planck Equation: Methods of Solutions and Applications, second ed. Springer Series in Synergetics. Springer, Berlin.

Sibert, J.R., Hampton, J., Fournier, D.A., Bills, P.J., 1999. An advection–diffusion-reaction model for the estimation of fish movement parameters from tagging data, with application to skipjack tuna (*Katsuwonus pelamis*). Can. J. Fish. Aquat. Sci. 56, 925–938.

Skellam, J.G., 1951. Random dispersal in theoretical populations. Biometrika 38, 196–218.

Strang, G., 1968. On the construction and comparison of difference schemes. SIAM J. Numer. Anal. 5, 506–517.

Van Leer, B., 1977. Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. J. Comput. Phys. 23, 276–299.

Yuan, L., Kalbleisch, J., 2000. On the bessel distribution and related problems. Ann. Inst. Statist. Math. 52, 438–447.

Article I : [11] S. Dueri, B. Faugeras et O. Maury. Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E : Part 1. Model formulation. *Ecological Modelling* 245 (2012), p. 41–54

Ecological Modelling 245 (2012) 41–54

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E: Part 1. Model formulation

Sibylle Dueri[a,*], Blaise Faugeras[b], Olivier Maury[a]

[a] Institut de Recherche pour le Développement (IRD), UMR EME 212 (IRD/Ifremer/Université Montpellier 2), Centre de Recherche Halieutique Méditerranéenne et Tropicale, Avenue Jean Monnet BP 171, 34203 Sète Cedex, France
[b] CNRS, Laboratoire J.A. Dieudonné (UMR 7351), Université de Nice Sophia-Antipolis, Faculté des Sciences, Parc Valrose, 06108 Nice Cedex 02, France

ARTICLE INFO

Article history:
Available online 21 March 2012

Keywords:
Tropical tuna
Pelagic fishery
Dynamic Energy Budget
Marine ecosystem model

ABSTRACT

APECOSM-E (Apex-Predator-Ecosystem-Model-Estimation) is a deterministic model that represents the 3D distribution and population dynamics of tropical tuna under the joint effect of environmental conditions and exploitation by fisheries. It is a simplified version of the top predator component of the APECOSM framework, based on a single partial differential equation. The model is structured in 3D space and fish size and considers size dependent reproduction, growth, predation, natural mortality and fishing mortality. Processes are time, space and size-dependent and linked to the environment through mechanistic bioenergetic or behavioral parameterizations. Physiological rates such as growth, reproduction and ageing mortality are derived from the Dynamic Energy Budget (DEB) theory, while horizontal movements and vertical distribution obey a mechanistically derived advection–diffusion formulation driven by habitat gradients and oceanic currents. The effect of fishing is accounted for through the use of fleet-specific size and depth selectivity functions and time-dependent catchability coefficients which relate observed fishing effort to catches and size-frequencies.

In this paper we present the mathematical formulations of the physiological and behavioral components of the model, and an application to the skipjack tuna population in the Indian Ocean. The model is run with a daily time step on a $1^\circ \times 1^\circ$ horizontal grid and considers 20 vertical layers, reaching a maximal depth of 500 m. Results show the effects of spatial and temporal variability of environmental conditions on tuna physiology in terms of growth, reproduction and survival. Moreover, our results suggest that observed trends in reported catches are connected to environmental conditions by means of recruitment dynamics. In addition, the model allows representing the horizontal and vertical distribution of skipjack tuna and assessing the effect of accessibility of the resource to fisheries. The ability of the model to represent the distribution of biomass in accordance with the pattern given by the observed fishing activity was evaluated by comparing the spatial distribution of the simulated biomass with the observed distribution of commercial purse seiners and bait boats catches in the Indian Ocean.

The likelihood based method used for estimating the model parameters as well as an analysis of its sensitivity to their values is provided in a companion paper (Dueri et al., 2012).

1. Introduction

Skipjack tuna (*Katsuwonus pelamis)* is a widely distributed, pelagic fish commonly found in tropical waters and commercially harvested by industrial and artisanal surface fisheries using purse seine, gillnet and bait boat. In the Indian Ocean skipjack represents almost half of the tropical tuna catches. The exploitation has increased rapidly after the introduction of industrial purse seining in the early 1980s and the concurrent raise of bait boat and gillnet catches. In 2006 the annual catch of skipjack in the Indian Ocean peaked at 620,000 t and since then, catches have not exceeded 450,000 t (Indian Ocean Tuna Commission, 2010). A possible explanation for this trend can be found in the recent development of the Somalian piracy, which induced a decline of the nominal effort along the usually well exploited Somalian coast (UNOSAT, 2009). Nevertheless, the simultaneous decrease of catches reported by the Maldivian fishery (Adam, 2010), one of the leading skipjack tuna fisheries in the Indian Ocean which is not subjected to pirates' attacks, may indicate that the population is overfished.

Skipjack tuna is considered to be a highly migratory species, which does not show clear spawning or feeding migration patterns (Stéquert and Ramcharrun, 1996) but rather exhibits home range movements within areas of good habitat. The spatial

DOI of original article: 10.1016/j.ecolmodel.2012.02.008.
* *Corresponding author*. Tel.: +33 0 499 57 32 53; fax: +33 0 499 57 32 95.
E-mail address: sibylle.dueri@ird.fr (S. Dueri).

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2012.02.007

distribution, movements and vulnerability to fishing of skipjack are affected by their habitat preferences, which are mostly determined by prey availability, temperature and oxygen conditions (Barkley et al., 1978; Brill, 1994; Brill and Lutcavage, 2001). As a consequence, the spatial distribution of fishing effort directed to skipjack exhibits seasonal and inter-annual patterns that can be related to environmental conditions (Mugo et al., 2010).

The current knowledge of the skipjack physiology states that the species is characterized by a fast growth and a high spawning potential, implying that the population is likely to have a high resilience to exploitation. However, the recently observed trend of the Indian Ocean skipjack catches questions the resilience of the population under present conditions and emphasizes the need for tools capable to evaluate the state of the population and its future evolution (Indian Ocean Tuna Commission, 2010). For this purpose, we propose the APECOSM-E model (Apex Predator Ecosystem Model-Estimation) a deterministic model that represents the spatio-temporal variability of the population under variable environmental and fishing conditions. Our approach integrates the main biological, behavioral and exploitation processes in a single mathematical framework, based on a partial differential equation that explicitly represents 3D movements, growth and mortality and their dependency on environmental conditions. By integrating these processes the model allows to assess the population dynamics and the sustainability of its exploitation.

The APECOSM-E model is a simplified version of the more general APECOSM framework (Maury, 2010), which represents the global flow of energy through the marine ecosystem considering different communities of epipelagic and mesopelagic organisms. APECOSM-E is derived from APECOSM, but is focused on a single species and its main objective is to integrate fisheries data for parameter estimation. It describes the physiology and behavior of individuals in a population with a very high level of detail and represents the state of the art of our knowledge about the physiology and behavior of skipjack tuna. In this paper we present an application of the model to the skipjack tuna population of the Indian Ocean and we use environmental variables to define the habitat and constrain the physiological rates of the species and their spatio-temporal variability. The main goal of the present application is to investigate the joint effects of environmental variability and fishing on the spatio-temporal dynamics of skipjack tunas in the Indian Ocean and improve our understanding of environmental effects on the physiology and behavior of this top-predator.

A likelihood method used for estimating the model parameters related to fisheries as well as an analysis of its sensitivity to their value is provided in a companion paper (Dueri et al., 2012).

2. The model

The dynamics of the skipjack tuna population described in the APECOSM-E model is driven by the environment and by fisheries exploitation. Environmental factors such as temperature, oxygen, food and currents determine the movements of tunas and affect their physiological rates (growth, reproduction and mortality). On the other hand, spatialized fishing effort data determine the fishing mortality and are used to simulate monthly catches and size frequencies. A schematic overview of the model components in terms of forcing, processes and outputs is provided (Fig. 1). Parameters descriptions are summarized in Table 1.

2.1. Implementation of the Dynamic Energy Budget approach

In the APECOSM-E model, the main physiological processes such as growth, reproduction and ageing mortality, are represented using a Dynamic Energy Budget (DEB) based approach. The DEB theory (Kooijman, 2000) relies on a mechanistic bioenergetic representation of the organism that describes the individual in terms of biomass and energy fluxes. In the standard DEB model the energy of an organism is stored in three pools: reserve, structure and maturity. Energy is introduced into the organism through the ingestion of food which is assimilated and stocked in the reserves compartment. A fixed fraction κ of the energy utilized from the reserve compartment is allocated to growth of structure and somatic maintenance while the remaining part $(1-\kappa)$ is allocated to maturity development and reproduction and maturity maintenance. Total biomass can be expressed as the sum of structural biomass, reserves biomass and biomass of the reproductive buffer.

The APECOSM-E model adds two assumptions to the DEB theory that allow considerable simplifications:

(1) the dynamics of the reserve pool is fast compared to the dynamics of structure (see Maury and Poggiale, submitted, for the mathematical details about this assumption). This implies that, at the time scale relevant for population dynamics, the reserve density $[E]$ is at or near equilibrium and equals the scaled functional response to food f_F times the maximum energy density in the reserve $[E_m]$; $[E]^* = f_F[E_m]$.
(2) reproduction is supposed to be continuous without stocking of energy in the reproductive buffer so that the influence of the reproductive buffer on total biomass and energy budget is neglected. Therefore the total weight W_{tot} of an organism can be approximated as the sum of the structural biomass and the reserves biomass:

$$W_{tot} \approx d_V V + f_F V \frac{[E_m]}{\psi} \tag{1}$$

where d_V is the density [g m^{-3}], V is the structural volume [m^3] (or the volume of structural biomass), f_F is the functional response to food [–], $[E_m]$ the maximum energy density of reserves [J m^{-3}] and ψ is the energy content of reserves [J kg^{-1}]. In the model, according to the DEB theory, the representation of growth, reproduction and ageing mortality is based on the structural volume, while the calculation of catches is based on total weight (Eq. (1)).

Following the standard DEB model assumption, we consider that skipjack is an isomorphic organism and keeps the same shape while growing. This allows to link structural volume to length using a shape coefficient. Structural volume is calculated as the cube of the volumetric length L, $V = L^3$, and L is related to the physical length L_w through the shape coefficient δ_M, $L = \delta_M L_w$. Therefore the structural volume can be written as

$$V = (\delta_M L_w)^3 \tag{2}$$

The allometric length-weight conversion for skipjack tuna in the Indian Ocean (Indian Ocean Tuna Commission, 2005) can be calculated using following empirical relationship:

$$W_{tot} = a L_w^b \tag{3}$$

where L_w is the physical length and the coefficients a and b are equal to 5.32×10^{-6} and 3.34 respectively. By substituting V and W_{tot} in Eq. (1) we obtain the value of the shape coefficient δ_M.

2.2. General model equation and boundary conditions

The tuna population is described through a biomass density function $p(x,y,z,V,t)$ [kg m^{-3} m^{-3}], where position $(x,y,z) \in \Omega$, a bounded domain representing the Indian Ocean in 3D, structural volume $V \in (V_b, V_{max})$ with V_b being the structural volume at birth and time $t \in (0,T)$.

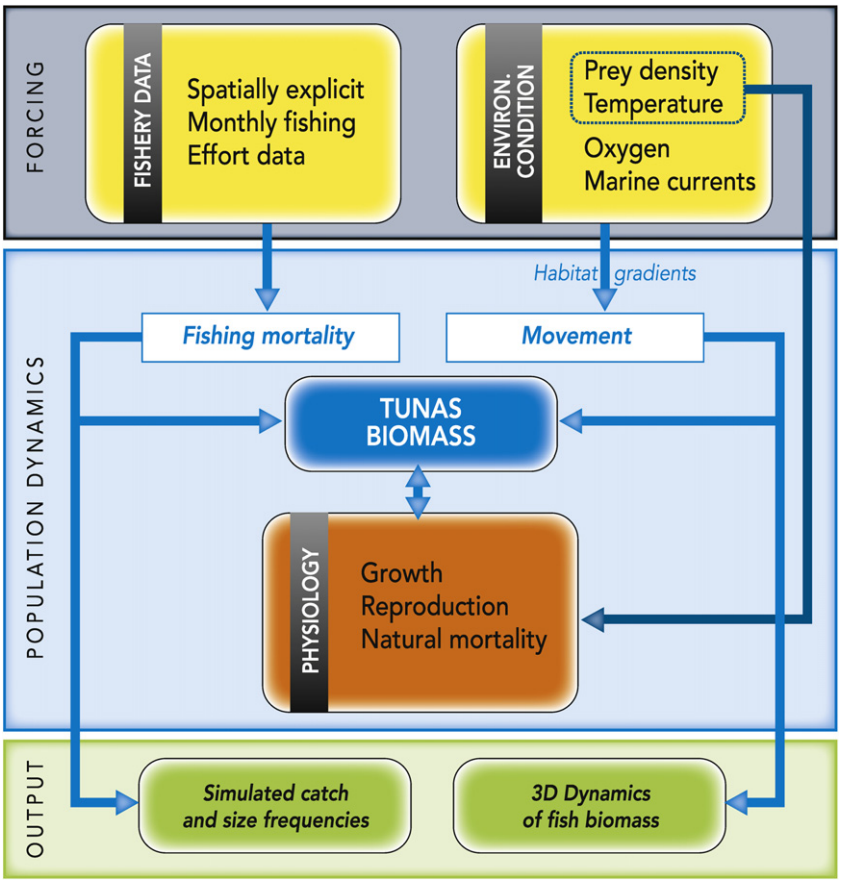

Fig. 1. Diagram showing the forcing, the processes affecting the population dynamics and the outputs of the APECOSM-E model.

The biomass of fish with structural volume comprised between V_1 and V_2 at time t in the domain $\Omega' \subset \Omega$ is given by the integral

$$\int_{\Omega'} \int_{V_1}^{V_2} p(x, y, z, V, t)\ dxdydzdV \tag{4}$$

The change of the population density function p as a function of time follows the mass balance equation below where ∇ and div are the usual differential operators. For technical reasons related to time-scale assumptions (see Section 2.6 for details) we distinguish horizontal movements from vertical movements.

$$\partial_t p = div(d\ \nabla p - vp) + \partial_z(d_z \partial_z p - v_z p) - \partial_V(gp) - (m + f)p \tag{5}$$

The four terms on the right side of Eq. (5) represent: (1) horizontal advection and diffusion, (2) vertical advection and diffusion, (3) growth and (4) natural and fishing mortality. Advection and diffusion are represented by the horizontal velocity $v(x,y,z,V,t)$ [m s^{-1}], the vertical velocity $v_z(x,y,z,V,t)$ [m s^{-1}], the horizontal diffusion $d(x,y,z,V,t)$ [m^2 s^{-1}] and the vertical diffusion $d_z(x,y,z,V,t)$ [m^2 s^{-1}]. Here we assume that there is no cross-diffusion term in z. Growth is represented as an advection of the biomass density in the size dimension and is characterized by the growth rate $g(x,y,z,V,t)$ [m^3 s^{-1}], while natural and fishing mortality rates are described by $m(x,y,z,V,t)$ [s^{-1}] and $f(x,y,z,V,t)$ [s^{-1}] respectively.

Initial and boundary conditions need to be prescribed to integrate Eq. (5). The initial population density distribution is given by:

$$p(x, y, z, V, 0) = p^0(x, y, z, V), \quad \forall (x, y, z, V) \in \Omega \times (V_b, V_{\max}) \tag{6}$$

The boundary conditions for the input of newborns into the systems $r(x,y,z,t;p)$ [g s^{-1}] is given by:

$$gp(x, y, z, V_b, t) = r(p), \quad \forall (x, y, z, t) \in \Omega \times (0, t_{\max}) \tag{7}$$

The mass conservation within the spatial domain is guaranteed by the following Neumann boundary condition:

$$\nabla p(x, y, z, V, t) \cdot n(x, y, z) = 0, \quad \text{on } \partial\Omega, \forall (V, t) \in (V_{\min}, V_{\max}) \times (0, t_{\max}) \tag{8}$$

where $n(x,y,z)$ is the unit normal vector pointing outside Ω.

The parameterization of the coefficients v, v_z, d, d_z, g, m, f and r and their biological and ecological basis are provided in the section below.

2.3. Distribution of forage and selectivity

APECOSM-E considers size-structured forage distribution. This allows accounting for the size selection of preys by predators (in this case skipjack tuna). Size-structured forage concentration $\varepsilon(V)$ is extrapolated from the mesozooplankton distribution of the NEMO-PISCES simulations. The concentration of mesozooplankton biomass [kg m^{-3}] is set as the first size-class (V_0) of the prey distribution $\varepsilon(V_0)$; then for size classes between [V_0, $V_{\max}$] $\varepsilon(V)$ is calculated assuming that the decrease of forage biomass follows a power law with a scaling exponent equal to −3 with respect to length, in accordance with size-distributions obtained in the general APECOSM model (Maury et al., 2007):

$$\varepsilon(V) = a \cdot L_W^{-3} = a \cdot \left(\frac{V^{1/3}}{\delta_M}\right)^{-3} \tag{9}$$

Table 1
Parameter description and parameter values used in the APECOSM-E model.

Parameter	Description	Value	Unit
κ	Fraction allocated to soma	0.8	–
δ_M	Shape coefficient	0.25	–
$L_{\max}$	Maximal length	1.1	m
$\{\dot{p}_{Am}\}$	Surface-area specific assimilation rate	$22.5 \times 10^6 \times L_{\max} \times \delta_M$,	J m^{-2} d^{-1}
$[E_m]$	Maximum energy density of reserves	$850 \times 10^8 \times L_{\max} \times \delta_M$,	J m^{-3}
$[E_G]$	Volume-specific energetic growth cost	2800×10^6	J m^{-3}
$[\dot{p}_M]$	Volume-specific maintenance cost	18×10^6	J m^3 d^{-1}
ψ_f	Energy content of reserves	38.8×10^6	J kg^{-1}
ψ_{str}	Energy content of structures	3.86×10^6	J kg^{-1}
l_{mat}	Length at maturity	0.4	m
κ_R	Fraction of reproduction energy fixed in eggs	0.95	–
ϕ	Sex ratio	0.5	–
$\ddot{h}_a$	Ageing acceleration	5×10^{-8}	d^{-2}
m_{p1}	Predation mortality coefficient 1	9.7×10^{-4}	d^{-1}
m_{p2}	Predation mortality coefficient 2	0.95	d^{-1}
m_{T1}	Temperature mortality coefficient	−1.	d^{-1}
T_0	Metabolic energy production/thermal capacity	0.2	°C s^{-1}
k_T	Thermic conductance/thermal capacity	0.12	m s^{-1}
T_a	Arrhenius temperature	5000	K
T_l	Lower boundary of tolerance range	299.15	K
T_h	Upper boundary of tolerance range	304.65	K
T_{al}	Lower boundary Arrhenius temperature	146,000	K
T_{ah}	Upper boundary Arrhenius temperature	38,000	K
T_1	Reference temperature	298.65	K
p_T	Weighting factor temperature	1	–
k_G	Half saturation constant for forage	4×10^{11}	kg m^3
p_F	Weighting factor forage	1	–
a_O	Steepness of oxygen limitation curve	10^5	
O_0	Half saturation constant for oxygen limitation	0.00014	mol L^{-1}
p_O	Weighting factor oxygen	1	–
a_{mdv}	Maximal attraction factor for Maldives	0.35	
v_{max}	Maximal horizontal speed	1	m s^{-1}
α	Concentration factor coefficient	1000	–
b	Maximal vertical speed	1	m s^{-1}
a	Behavioral diffusivity, vertical	0.15	m^2 s^{-1}
d^ϕ	Physical diffusivity, vertical	10^{-5}	m^2 s^{-1}
p_{ps1}	Catchability PS1	0.015	
p_{ps2}	Catchability PS2	0.015	
p_{ps3}	Catchability PS3	0.025	
p_{bb}	Catchability BB	0.005	
a_{ps1}	Increased efficiency due to technological development, PS1	0.200	
a_{ps2}	Increased efficiency due to technological development, PS2	0.200	
a_{ps3}	Increased efficiency due to technological development, PS3	0.200	
a_{bb}	Increased efficiency due to technological development, BB	0.100	
$l_{s,ps1}$	Length selectivity, PS1	0.5	m
$l_{s,ps2}$	Length selectivity, PS2	0.5	m
$l_{s,ps3}$	Length selectivity, PS3	0.5	m
$l_{s,bb}$	Length selectivity, BB	0.45	m
$k_{l,ps}$	Steepness length selectivity, PS	45	
$k_{l,bb}$	Steepness length selectivity, BB	45	
$z_{s,ps}$	Depth selectivity, PS	100	m
$z_{s,bb}$	Depth selectivity, BB	20	m
$k_{z,ps}$	Steepness depth selectivity, PS	0.3	
$k_{z,bb}$	Steepness depth selectivity, BB	0.3	

where $a = \varepsilon(V_0)/L_0^{-3}$.

The function $F(x,y,z,v,t)$ describes the biomass of forage "ingestible" by predators of size v. It is calculated by integrating the biomass of prey of size u, $\varepsilon(x,y,z,u,t)$ multiplied by the size dependent selectivity function $s(v,u)$ (Fig. 2) of a predator of size v on a prey of size u, over the size classes of the prey.

$$F(x,y,z,v,t) = \int_{V_{\min}}^{V_{\max}} s(v,u)\varepsilon(x,y,z,u,t)\,du \tag{10}$$

The selectivity function is calculated as the product of two sigmoid functions. It considers that predation occurs if the ratio of predator length over prey length is neither too small (prey too large to be ingested) nor too large (prey too small to be located and kept in the mouth). A detailed description of the selectivity function s is provided in Maury et al. (2007).

2.4. Growth, reproduction and mortality

In the DEB approach, physiological rates depend upon food availability and temperature (Kooijman, 2000) and the following sections describe the relation used in the model for this purpose.

2.4.1. Functional response for temperature

Physiological rates depend on body temperature and tuna are endothermic organisms, able to retain heat and maintain body temperature above that of ambient temperature (Block and Stevens, 2001). The APECOSM-E model uses a mechanistic size-dependent description of the body temperature as a function of external temperature (Maury, 2005).

$$T_b(x,y,z,V,t) = \frac{V^{(1/3)}}{\delta_M}\frac{T_0}{k_T} + T_b(x,y,z,t) \tag{11}$$

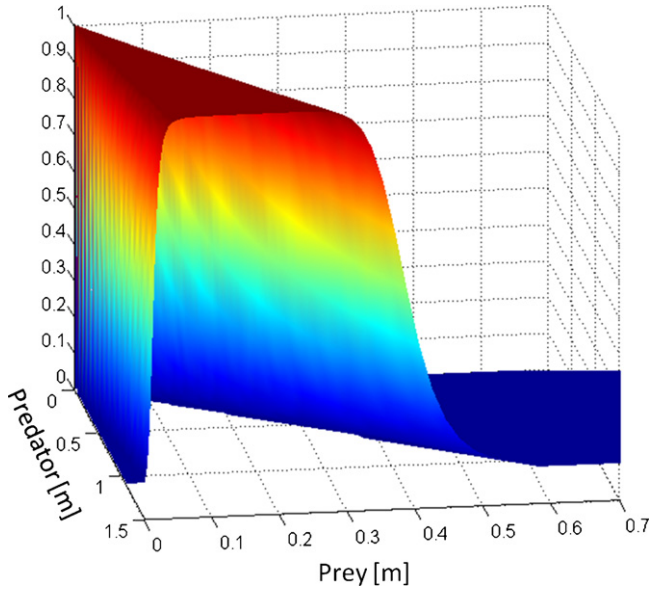

Fig. 2. Size selectivity function of predators on preys.

where k_T represents thermic conductance over thermal capacity and T_0 is metabolic energy production over thermal capacity. The values of k_T and T_0 for skipjack were estimated using data reported in Brill et al. (1994). This equation implies that, at a given external temperature, the steady state body temperature of a fish increases linearly with its length. As a result, to reach the same body temperature, a large fish will have to stay in cooler water than a small one.

The effects of body temperature on physiological rates can be represented as a product of several Arrhenius functions (Kooijman, 2000):

$$\hat{f}_T(T_b) = a_T(T_b) r_T(T_b) \tag{12}$$

where a_T describes the changes of any physiological rate with temperature:

$$a_T(T_b) = \exp\left(\frac{T_a}{T_1} - \frac{T_a}{T_b}\right) \tag{13}$$

and r_T describes the reduction of any physiological rates at low temperature due to congelation of phospholipidic cell membranes and subsequent inhibition of cellular metabolism and at high temperatures due to the loss of quaternary structure of protein catalytic enzymes and the subsequent inactivation of metabolic reactions:

$$r_T(T_b) = \left(1 + \exp\left(\frac{T_{al}}{T_b} - \frac{T_{al}}{T_l}\right) + \exp\left(\frac{T_{ah}}{T_h} - \frac{T_{ah}}{T_b}\right)\right)^{-1} \tag{14}$$

The final physiological response to temperature is normalized and reads

$$f_T(T_b) = \frac{\hat{f}_T(T_b)}{\hat{f}_{T_{\max}}} \tag{15}$$

The value of the Arrhenius temperature T_a was set in agreement with previous studies (Maury et al., 2007; Van der Veer et al., 2003) while the reference temperature T_1, the temperature at the lower and upper boundaries T_h and T_l and the Arrhenius temperature for the rate of decrease at the upper and lower boundaries T_{ah}, and T_{al} were determined according to reported habitat preferences of tropical tuna in the Indian Ocean, which are constrained between 20 and 32 °C (Stéquert and Marsac, 1989).

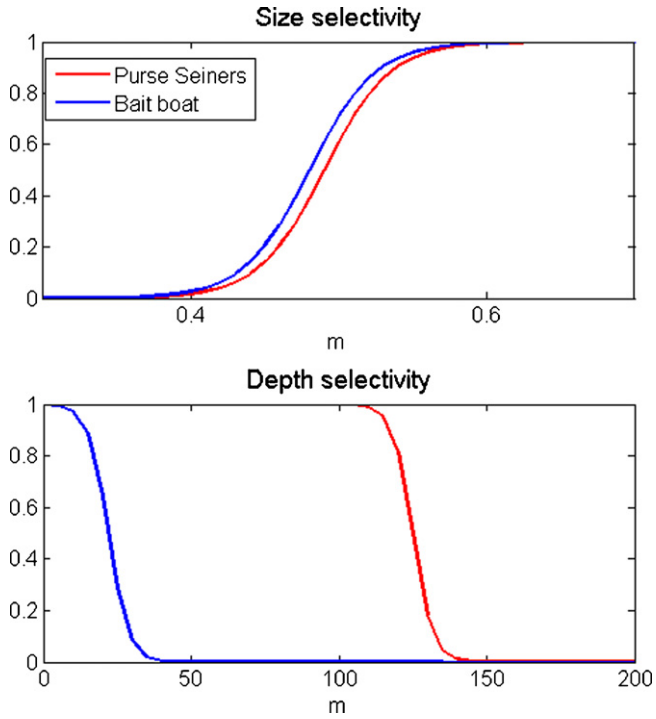

Fig. 3. Size and depth selectivity of purse seine and bait boat.

2.4.2. Functional response to food density

Physiological rates of tunas are affected by food intake, which in turn depends on the food availability. According to the DEB theory, APECOSM-E expresses this dependence with a Holling type 2 function:

$$f_F = \frac{\hat{F}}{k_G + \hat{F}} \tag{16}$$

where k_G is the half saturation constant and $\hat{F}$ is the biomass of accessible preys. The variable $\hat{F}$ accounts for the effect of density-dependence, so that when preys have to be shared between many predators, they become less accessible per capita. Since the rigorous computation of the abundance of predators sharing common preys (Maury et al., 2007) is computationally extremely demanding, we had to introduce an approximation. The biomass of accessible preys is therefore obtained by dividing the "ingestible" forage (Eq. (10)) at a given spatial location by the number of skipjack in the first size class V_1 at the same place.

$$\hat{F}(v) = \int_{V_{\min}}^{V_{\max}} s(v,u) \frac{\varepsilon(u)}{\int_{V_1}(p(V)/d_V(V))dV} du \tag{17}$$

2.4.3. Growth

Assuming that the reserve compartment is at equilibrium (cf. Section 2.1) and that heating costs are negligible in the energy budget, we can express the growth rate as follows (Kooijman, 2000)

$$g = \frac{dV}{dt} = f_T \left[\frac{\kappa f_F \{\dot{p}_{Am}\} V^{(2/3)} - [\dot{p}_M] V}{\kappa f_F [E_m] + [E_G]}\right]^+ \tag{18}$$

where $\{\dot{p}_{Am}\}$ is the surface-area specific assimilation rate [J m^{-2} s^{-1}], $[\dot{p}_M]$ is the volume-specific maintenance cost [J m^{-3} s^{-1}] and $[E_G]$ is the volume-specific energetic growth cost [J m^{-3}] and $[x]^+$ is the function defined by.

$$\begin{cases} [x]^+ = x & if \quad x \geq 0 \\ [x]^+ = 0 & if \quad x < 0 \end{cases}$$ The numerical values of $\{\dot{p}_{Am}\}$, $[\dot{p}_M]$ and $[E_G]$ were derived from Kooijman (2010) and their values are given in Table 1.

Following Kooijman (2000), under constant food availability conditions, the DEB growth equation is equivalent to the von Bertalanffy equation $L(t) = L_\infty - L_\infty e^{-r_B t}$, where L_∞ is the maximal length and r_B is the von Bertalanffy growth rate. As a consequence, r_B can be expressed as a function of DEB parameters:

$$r_B = \frac{1}{3}\frac{[\dot{p}_m]}{\kappa f_F[E_m] + [E_G]} \tag{19}$$

The von Bertalanffy growth rate estimated for skipjack tuna in the Indian Ocean ($r_B = 0.288$, Indian Ocean Tuna Commission, 2008) is therefore used to calculate the maximum energy density of reserves $[E_m]$, under the assumption that food and temperature conditions are constant and close to optimum ($f_F = 0.8$ and $f_T = 0.85$).

2.4.4. New born input

According to the standard DEB scheme, $(1-\kappa)$ of the energy mobilized from the reserve pool is allocated to reproduction and maturity maintenance. Reproduction rate is expressed using the same energetic parameters as growth plus two additional parameters: the mean proportion of females in the mature population ϕ and the fraction of the energy in the gonads which is turned into eggs κ_R. In our study ϕ is set at 0.5, which matches empirical observations regarding skipjack population of the Indian Ocean (Grande et al., 2010), while κ_R is set at 0.95 (Kooijman, 2000). In accordance with the DEB theory the total reproductive flux of the population $r(p)$ [g s^{-1}] is calculated as:

$$r(p) = f_T \frac{\phi\kappa_R(1-\kappa)}{d_V \psi_{str}} \int_{V_{mat}}^{V_{\max}} \left(p\left[\frac{f_F[E_m]}{[E_G] + \kappa f_F[E_m]} ([E_G]\dot{v}(V)^{(-1/3)} + [\dot{p}_M]) - \frac{[\dot{p}_M]V_{mat}}{\kappa V} \right]^+ \right) dV \tag{20}$$

where $\dot{v}$ is the energy conductance equal to $\{\dot{p}_{Am}\}/[E_m]$ (m s^{-1}) and V_{mat} is the structural volume at maturity. Grande et al. (2010) have estimated that 50% of the Indian Ocean skipjack females reach maturity at a length of 37.81 cm, while the results of Stéquert and Ramcharrun (1996) indicate slightly higher values of 41–42 cm. We therefore choose a mean length of sexual maturity at 40 cm. Moreover, we assume that reproduction occurs whenever the temperature is above 24 °C (Cayré and Farrugio, 1986).

The reproductive flux of the spawning population is used to calculate the size dependent population fecundity (g oocytes per kg of female per day)

$$Fec(V) = \frac{r(p(V))}{\phi p(V)} \tag{21}$$

where $r(p(V))$ is the size dependent reproductive flux [g s^{-1} m^{-3}] and the batch fecundity (number of oocytes per kg of female per spawning event):

$$BF = \frac{Fec(V) \times sf}{w_{oocyte}} \tag{22}$$

where sf is the spawning frequency and w_{oocyte} is the weight of an egg. A mean weight of 0.6 mg/oocytes is calculated by combing the mean dry weight, 0.042 mg/oocytes (Margulies et al., 2007) and the mean water content of tuna eggs, 93% (Ortega and Mourente, 2010).

2.4.5. Mortality

2.4.5.1. Natural mortality. The total natural mortality is represented as the sum of ageing, predation, starvation and temperature mortalities, i.e. $m = m_{ageing} + m_{pred} + m_{starv} + m_{temp}$.

The DEB theory relates ageing mortality to the amount of cellular damages that increase at a rate proportional to the respiration rate not associated to assimilation (Kooijman, 2000). As a consequence, a low metabolic rate corresponds to a longer life span. For the sake of simplicity we use the formula proposed by Maury and Poggiale (submitted) who calculate the mean size dependent ageing mortality rate by replacing the food functional response by its mean value $\bar{f}_F$. This leads to an explicit size-dependent expression of the ageing mortality based on DEB parameters only.

$$m_{ageing} = \frac{\ddot{h}_a}{V_t}\left\{ a^3 t_V - \frac{1}{c}\left(3a^2 d e^{ct_V} - \frac{3}{2}ad^2 e^{2ct_V} + \frac{1}{3}d^3 e^{3ct_V}\right) + \frac{1}{c}\left(3a^2 d - \frac{3}{2}ad^2 + \frac{1}{3}d^3\right) + V_{egg} t_V + \frac{[\dot{p}_M]}{[E_G]}\left[\frac{a^3}{2} t_V^2 - \frac{1}{c^2}\left(3a^2 d e^{ct_V} - \frac{3}{4}ad^2 e^{2ct_V} + \frac{1}{93}d^3 e^{3ct_V}\right) + \frac{1}{c}\left(3a^2 d - \frac{3}{2}ad^2 + \frac{1}{3}d^3\right) t_V + \frac{1}{c^2}\left(3a^2 d - \frac{3}{4}ad^2 + \frac{1}{9}d^3\right)\right]\right\} \tag{23}$$

With

$$a = \frac{\kappa\{\dot{p}_{Am}\}\bar{f}_F}{[\dot{p}_M]}$$

$$b = V_b^{(1/3)}$$

$$c = \frac{-[\dot{p}_M]}{3(\kappa\bar{f}_F[E_M] + [E_G])}$$

$$d = a - b$$

$$V_t = \left(\frac{\kappa\{\dot{p}_{Am}\}\bar{f}_F - (\kappa\{\dot{p}_{Am}\}\bar{f}_F - [\dot{p}_M]V_b^{(1/3)})([\dot{p}_M]/e^{3(\kappa\bar{f}_F[E_m]+[E_G])})t}{[\dot{p}_M]} \right)^3$$

$$t_V = \frac{-3(\kappa\bar{f}_F[E_m] + [E_G])}{[\dot{p}_M]} \ln\left(\frac{\kappa\{\dot{p}_{Am}\}\bar{f}_F - [\dot{p}_m]V^{(1/3)}}{\kappa\{\dot{p}_{Am}\}\bar{f}_F - [\dot{p}_m]V_b^{(1/3)}} \right)$$

where t_V is the time to reach size V with a mean food density of $\bar{f}_F$, V_b is the volume at birth and $\ddot{h}_a$ (s^{-2}) is the ageing acceleration.

Size dependent predation mortality is described by a power law relation with two mortality coefficients m_{p1} and m_{p2} that define the strength and the steepness of the function. Its value is maximal for small organisms (e.g. larvae) and decreases for larger organisms:

$$m_{pred} = m_{p1}\left(\frac{V^{(1/3)}}{\delta_M} \right)^{-m_{p2}} \tag{24}$$

The DEB theory states that the assimilated energy is used first for the maintenance of the organism before being allocated to growth and reproduction. If food availability is too low, the growth and reproduction ceases, and all the available energy is allocated to maintenance. When maintenance costs are not covered, the organism health declines and this threatens its survival. Therefore a starvation process is introduced. As in Maury et al. (2007) starvation mortality m_{st} is expressed as the energy which would be needed for maintenance but cannot be provided by the assimilation of food:

$$m_{st} = \frac{1}{(f_F[E_m] + d\psi)}\left\{ [E_G]\left[\frac{[\dot{p}_m] - \kappa\{\dot{p}_{Am}\}f_F V^{-(1/3)}}{[E_G] + \kappa f_F[E_m]} \right]^+ + (1-\kappa) f_T \left[\frac{[\dot{p}_m]V_{\max}}{\kappa V} - \frac{f_F[E_m]([E_G]\nu V^{-(1/3)} + [\dot{p}_m])}{[E_G] + \kappa f_F[E_m]} \right]^+ \right\} \tag{25}$$

Finally, we include a mortality term for organisms subject to temperatures too cold or too warm for their survival. This term is linked to the variable r_T of the Arrhenius relationship (see Eq. (14)) that describes the reduction of physiological rates at low and high

temperatures. The temperature mortality is computed only when physiological rates are lower than a given threshold.

$$m_{temp} = m_{T1}\ \log(m_T) \quad \text{if } m_T = 0.1 \tag{26}$$

where $m_T = r_T/r_{T\max}$ and m_{T1} is a negative parameter.

2.4.5.2. Fishing mortality. The fishing mortality is calculated using observed $1^\circ \times 1^\circ$ monthly fishing effort for four different fleets: French purse seiners "PS1", Spanish purse seiners "PS2", "World" purse seiners "PS3" grouping the fishing data of Mauritius, Seychelles and NEI-other and Maldivian bait boats "BB". These fleets represent the main skipjack fisheries of the Indian Ocean providing a time series of fishing data with a spatial resolution of $1^\circ \times 1^\circ$.

Fishing mortality of fleet k is calculated as the product of the observed fishing effort e_k by the catchability p_k at T_0 multiplied by an exponential function representing the increase in fishing efficiency at a rate a_k due to technological development in time, and two selectivity functions, one for size and the other for depth.

$$f_k(i,z,V,t) = e_k(i,t)p_k\ \exp(a_k t)\frac{1}{1+\exp(-k_1((V^{(1/3)}/\delta_M)-l_s))} \times \frac{1}{1+\exp(-k_z(z-z_s))} \tag{27}$$

For the purse seiners the fishing effort is expressed as the amount of time the fishermen spend at fishing while for bait boat it is expressed as the amount of time spent at sea. Technological development is a continuous process and includes the increase of the size and performance of the fishing vessels, the enhancement of the fishing gears, the progressive use of new electronic devices such as bird radar and other remote sensing tools and the deployment of more and more sophisticated fish aggregating devices (FADs) (Valdemarsen, 2001). The gear specific length and depth selectivity are represented using sigmoid functions where l_S and z_S are the length and depth leading to 50% selection while k_l and k_z characterize the steepness of the sigmoid curves (Fig. 3).

2.5. Habitat and movements

Water temperature, dissolved oxygen concentration and forage availability are the main factors affecting the physiological and behavioral responses of tuna to the environment. In the model, tunas are attracted to areas where the environmental conditions are favorable to their growth, reproduction and survival. For that purpose, structural environmental factors are translated into a synthetic functional habitat variable by means of functional responses that charaterize the habitat suitability. The spatial heterogeneity of the modeled functional habitat creates gradients that steer the movements of tunas.

2.5.1. Habitat suitability index

The model calculates the 3D habitat suitability index h by considering 3 factors: temperature, food and oxygen conditions. For each factor a functional response varying between 0 (highly unfavorable) and 1 (highly favorable) quantifies environmental suitability with respect to that factor (Fig. 4). The habitat suitability index is then expressed as the product of weighted functional responses to temperature f_T, food f_F and oxygen f_O, with p_T, p_F, p_O being the respective weighting factor.

$$h(x,y,z,V,t) = f_T(x,y,z,V,t)^{p_T} f_F(x,y,z,V,t)^{p_F} f_O(x,y,z,V,t)^{p_O} \tag{28}$$

The functional responses describing the habitat suitability in terms of temperature and food availability are consistent with the ones that represent the change in physiological rates (Eqs. (15) and (16)). The functional response for oxygen is presented in the next section.

2.5.2. Functional response for oxygen

The functional response to oxygen $f_O(x,y,z,V,t)$ is represented using a sigmoidal curve:

$$f_O = \frac{1}{1+\exp(a_0(O-O_0))} \tag{29}$$

where a_0 is the steepness of the curve and O_0 is the half saturation coefficient for oxygen limitation. The value of the steepness and half saturation coefficient have been determined according to previous studies which estimated that skipjack tuna need oxygen concentrations above 2 ml/L to survive (Gooding et al., 1981; Sharp, 1978) and usually prefer environments with oxygen levels higher than 3.0–3.5 ml/L (Barkley et al., 1978; Brill, 1994).

2.5.3. Maldivian Island attraction

The enhancement of primary production and associated aggregations of zooplankton, micronekton and fish arising in conjunction with ocean currents impinging on abrupt topographies is a known phenomenon and the mechanisms that drive this bio-physical process have been described by Genin (2004). The Maldives are a typical example of this enhanced productivity: the presence of narrow channels between the double chain of atolls and the mixing of the stratified equatorial water pumps nutrient rich subsurface water to the surface (Anderson et al., 2011). This so-called island mass effect (IME) has been observed in different channels of the Maldives Islands using chlorophyll-a field data of the Moderate Resolution Imaging Spectrometer (MODIS) (Sasamal, 2006).

Given the importance of the Maldivian fisheries in terms of tuna catches, it is essential to represent the IME in the model in order to explain the high productivity of this area; however the small-scale hydrodynamic and biogeochemical processes responsible for an increased productivity around the Maldives are missing in our model based the fact that the environmental forcing comes from an oceanographic simulation with a resolution of 0.5° (cf. 2.6), which is too coarse to capture the processes responsible for the enrichment. Therefore it was necessary to introduce an attraction factor β that increases the habitat quality around the Maldives. β can vary between 0 and a_{mdv} ($a_{mdv} \leq 1$) and is modeled using a two dimensional (x,y) Gaussian function centered on the islands.

$$\beta = a_{mdv} \cdot e^{-\left[\frac{(x-x_{mdv})^2}{2\sigma_x^2}+\frac{(y-y_{mdv})^2}{2\sigma_y^2}\right]} \tag{30}$$

where x_{mdv} and y_{mdv} are the coordinates of the center of the attraction basin which is located at 4.5°N and 74.5°E. To account for the north-south extent of the Maldives archipelago, the standard deviation of the Gaussian attraction function was set to 4° along the latitude and 2° along the longitude axis.

The Maldivian islands attraction is not supposed to be added to the normal habitat driven movements but to replace them so that the resulting habitat function $\hat{h}$ still varies between [0,1]:

$$\hat{h}(x,y,z,t,V) = (1-\beta)\cdot h(x,y,z,t,V) + \beta \tag{31}$$

2.5.4. Horizontal and vertical movements

Horizontal advection and diffusion have both a physical component due to passive transport by marine currents and a biological component due to active movement of fish. In APECOSM-E the biological advection depends on the habitat gradient: the velocity and direction of tuna movement are locally affected by temperature, oxygen, forage fields and the Maldivian island attraction. Advection is oriented in the direction of the habitat gradient and the balance between advection and diffusion depends on the gradient intensity. While strong gradients impose strong advection and weak diffusion, weak gradients induce weak advection and strong diffusion. Moreover, active swimming of the fish is assumed to decrease when habitat quality increases so that both advection and diffusion

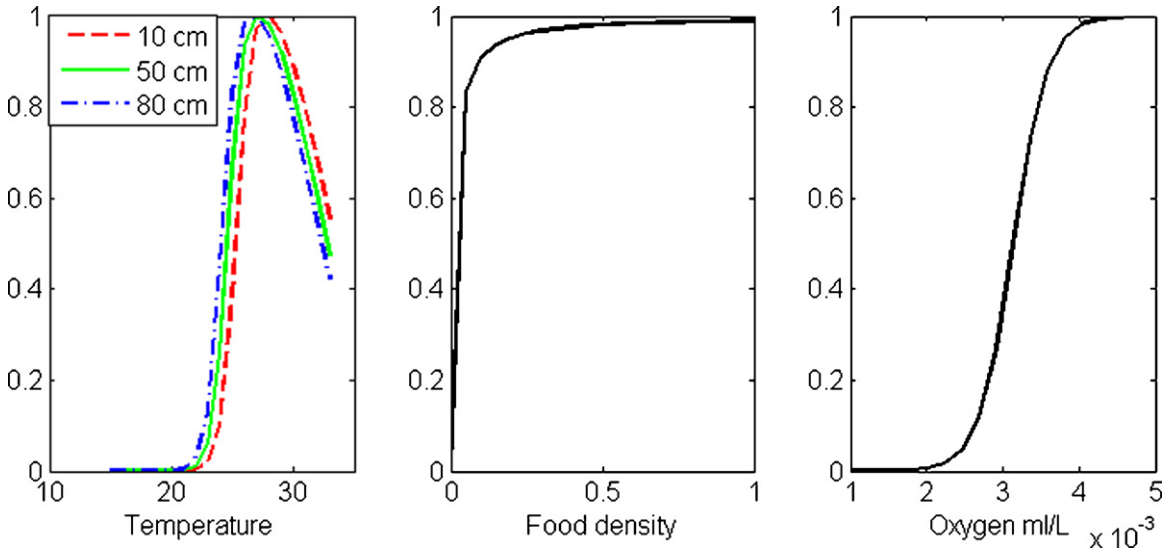

Fig. 4. Functional responses for temperature, food density and oxygen conditions.

decrease simultaneously. This implies that the better the habitat, the lower the interest in heading toward better habitats.

Accordingly, horizontal movement in APECOSM-E are expressed using the mechanistically derived advection–diffusion equation presented in Faugeras and Maury (2007) which allows to conserve the total size and habitat dependent distance traveled per time by a fish when advection and diffusion change. The horizontal advection vector is represented as the sum of biological advection due to swimming (left term) and physical advection due to marine currents (right term):

$$v = v_{\max}\frac{V^{(1/3)}}{\delta_M}(1-h)\frac{I_1(\alpha||\nabla h||)}{I_0(\alpha||\nabla h||)}\begin{pmatrix}\cos\theta_{\nabla h}\\ \sin\theta_{\nabla h}\end{pmatrix} + v_{phy} \tag{32}$$

where v_{max} is the maximal speed that a 1 m fish can reach, α is the concentration factor, $||\nabla h||$ is the norm of the habitat gradient, I is the modified Bessel function at order 0 and 1, $\theta_{\nabla h}$ is the angle of the gradient and v_{phy} is the physical velocity determined by the current forcing field.

The horizontal diffusion matrix is given by a physical diffusion term (D_{min}) that is added to the behavioral term:

$$d = D_{\min}\begin{pmatrix}1&0\\0&1\end{pmatrix} + \frac{\tau}{2}\left(v_{\max}\frac{V^{(1/3)}}{\delta_M}(1-h)\right)^2 \times \left\{\frac{1}{2}\left(1-\frac{I_2(\alpha||\nabla h||)}{I_0(\alpha||\nabla h||)}\right)\begin{pmatrix}1&0\\0&1\end{pmatrix} + \left[\frac{I_2(\alpha||\nabla h||)}{I_0(\alpha||\nabla h||)} - \left(\frac{I_2(\alpha||\nabla h||)}{I_0(\alpha||\nabla h||)}\right)^2\right]\begin{pmatrix}\cos^2\theta_{\nabla h} & \sin\theta_{\nabla h}\cos\theta_{\nabla h}\\ \sin\theta_{\nabla h}\cos\theta_{\nabla h} & \sin^2\theta_{\nabla h}\end{pmatrix}\right\} \tag{33}$$

where τ is a small mean time during which the velocity vector of an individual is constant. Vertical advection relies only on active movements such as bounce-dive foraging behavior (Schaefer et al., 2009) while passive vertical transport due to vertical marine currents is neglected. In the model the vertical advection velocity decreases with habitat quality and is proportional to the maximal vertical speed b, the size of the organism and the vertical gradient of the habitat function.

$$v_z = b(1-h)\left(\frac{V}{V_{\max}}\right)^{(1/3)}\partial_z h \tag{34}$$

The vertical diffusion d_z has two components: a behavioral and a physical one.

$$d_z = ah\left(\frac{V}{V_{\max}}\right)^{(2/3)} + d_z^{\phi} \tag{35}$$

The first term describes the size-dependent diffusion emerging from random foraging vertical movements. It depends on the behavioral diffusivity coefficient a ($m^2\,s^{-1}$), the size of the organisms and increases linearly with the habitat index h meaning that the tuna spend more time randomly looking for food when the habitat is good. The second term of the equation is a physical vertical diffusivity term d^{ϕ} ($m^2\,s^{-1}$) that accounts for purely physical vertical mixing which is especially important for small organisms.

2.6. Numerical integration and forcing

In order to speed up the calculation and reduce the memory needs, we simplify Eq. (5) using time-scale assumptions. Assuming that the vertical movements are fast processes compared to horizontal movements, mortalities and newborns input, vertical movements can be partially decoupled from the other processes and integrated analytically. This avoids a costly numerical solving of the full 3D + size system and reduced the numerical procedure to a lighter 2D + size problem. The mathematical details of this simplification are provided in the Annex.

The numerical integration of APECOSM-E uses a $1^\circ \times 1^\circ$ horizontal grid covering the Indian Ocean (20–130° East, 40° South to 30° North). There are 20 vertical layers from 0 to 500 m, with a 10 m interval in the first 150 m. This vertical grid allows having a good resolution of the water column between 0 and 150 m, which owing to temperature and oxygen conditions corresponds to the depth usually occupied by skipjack. Organisms considered in the model range between 1 mm and 1 m length. In order to accurately and effectively account for growth and predation of organism having very different sizes, we define 83 size classes using a logarithmic scale (see Maury et al., 2007 for details). Therefore the size interval is very small for small organisms and becomes progressively larger. This allows reducing the number of size classes and ensures that all processes are considered at the proper resolution.

The non-dispersive MUSCL (monotonic upstream centered scheme for conservation laws) is used for integrating the spatial advection terms (Van Leer, 1977), while for diffusion we use a three-point finite difference scheme explicit in time. A first order upwind finite difference scheme is used to integrate the growth term of Eq. (5). All integrals are evaluated using first order centered approximations.

Simulations are run with a daily time step for the time period 1958–2001, with the industrial fishery exploitation starting in

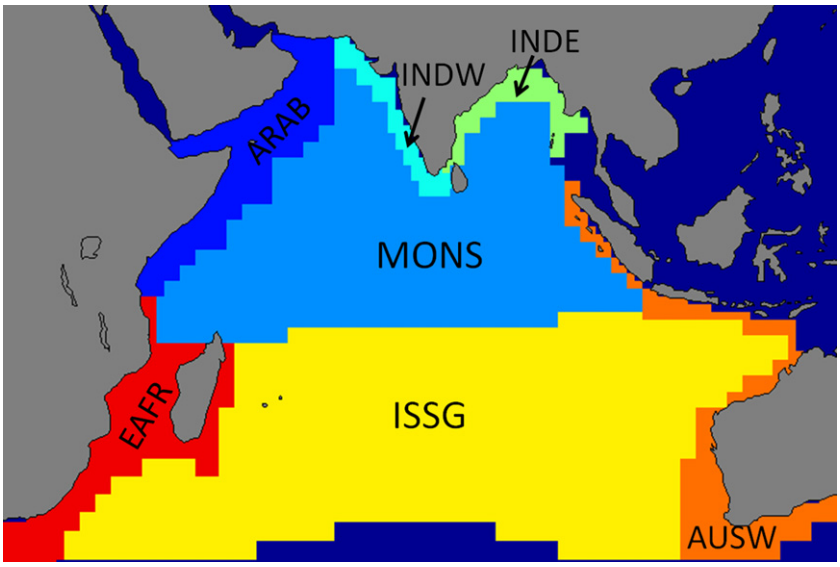

Fig. 5. Longhurst provinces in the Indian Ocean.

1984. The first 10–15 years of simulation are considered to be a spin-up phase of the model. The environmental conditions that determine the tuna habitat are provided by 3D temperature, oxygen, mesozooplankton and marine current fields generated by the NEMO-PISCES model (Aumont and Bopp, 2006), a coupled physical–biogeochemical model that is run for the global ocean. NEMO-PISCES was run at a 0.5° resolution using the ERA40 reanalysis, a re-analysis of the global atmosphere and surface conditions from 1957 to 2002 performed by the European Centre for Medium-Range Weather Forecast (ECMWF). This environmental forcing is read by the APECOSM-E model every 10 simulated days.

The effect of fishing activities is introduced by means of observed effort data obtained from the Indian Ocean Tuna Commission in the standardized form available on the CLIOTOP website (http://vmmdst-proto.mpl.ird.fr/MDST/). Fishing efforts is spatially aggregated on a $1^\circ \times 1^\circ$ grid and on a monthly basis for the four fleets considered. They are used to simulate catches and size frequencies.

3. Results and discussion

3.1. Spatio-temporal variability of growth and reproductive flux

Skipjack tuna are highly mobile species and during their life span they encounter a wide range of environmental conditions that affect their physiological rates. While all metabolic rates are regulated by temperature according to Eq. (12), the energy allocated to growth and reproduction depends upon the food assimilation rate, which is proportional to ingestion and which in turn is determined by the food availability. The model allows to explicitly account for this spatial variability of food and temperature conditions and to represent the range of predicted growth and reproduction rates anywhere in the Indian Ocean, from food and temperature fields.

Following Longhurst (1998), the Indian Ocean can be subdivided into seven biogeochemical provinces with homogeneous conditions for temperature and productivity (Fig. 5): Northwestern Arabian Upwelling (ARAB), Australia-Indonesia Coastal (AUSW), Eastern Africa Coastal (EAFR), Eastern India Coastal (INDE), Western

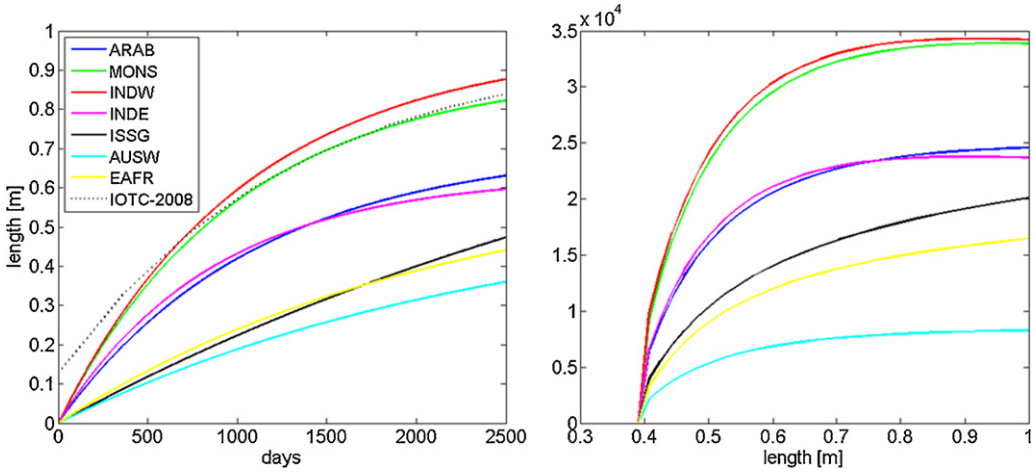

Fig. 6. Yearly average of Von Bertalanffy growth function (left) and daily fecundity [oocytes per kg biomass and day] (right) for the Longhurst provinces of the Indian Ocean.

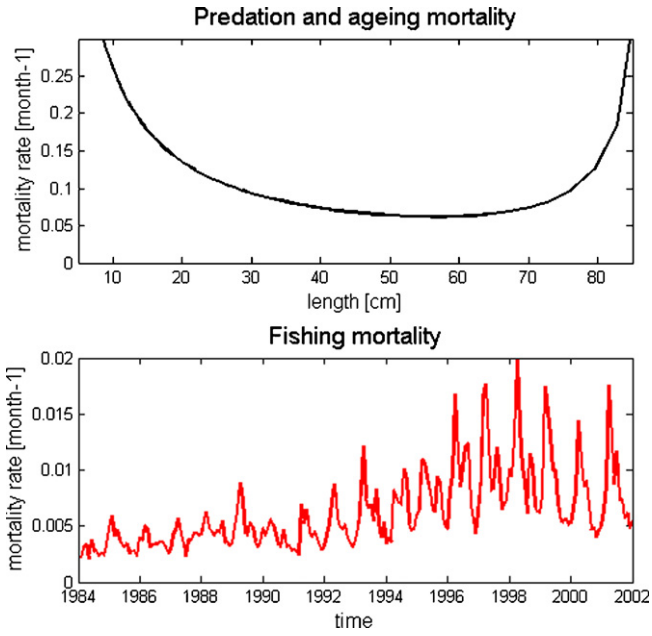

Fig. 7. Size dependent natural fish mortality (predation and ageing) compared to fishing mortality (simulated) as a function of time (1984–2001).

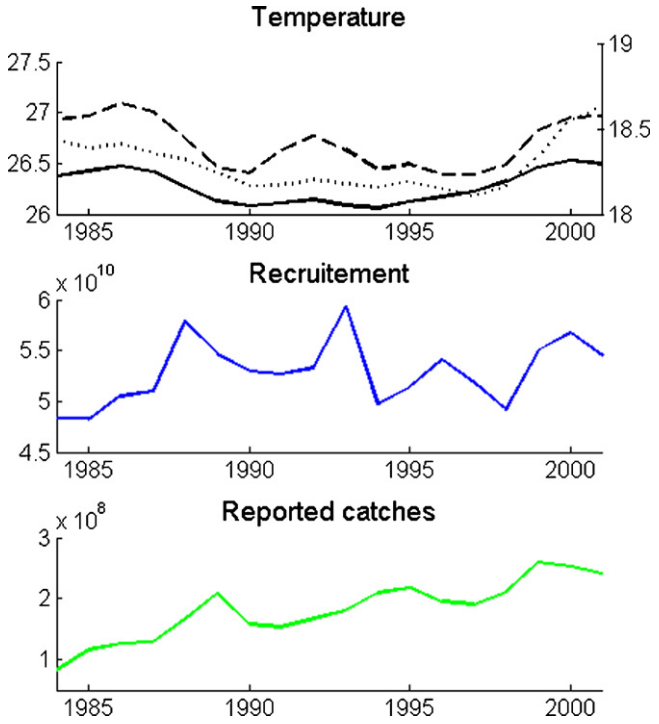

Fig. 8. Annual trends in mean temperature in the MONS (solid line), ARAB (dashed line) and INDW (dotted line) provinces (top); computed recruitment of skipjack in number of fishes reaching a size of 30 cm (center) and reported total catches (bottom).

India Coastal (INDW), Indian South Subtropical Gyre (ISSG), Indian Monsoon Gyres (MONS). Here, we use this subdivision to quantify the spatial variability of the mean yearly growth rates and fecundity as a function of size at the basin scale (Fig. 6).

The results point out that the provinces of the southern hemisphere (ISSG, EAFR, AUSW) are characterized by a slower growth rate and a lower fecundity, as a consequence of the lower temperature and generally lower food availability in those areas. On the contrary, the MONS and the INDW province appear to be the ones with the most suitable conditions for growth and reproduction of the population. Moreover, the comparison of the growth rates between provinces implies an important variability of the size-at-age of tunas exposed to different environmental conditions. It is important to account for this variability since environmental conditions change in time and space following seasonal cycles and tuna might be exposed to more or less favorable conditions during their migration and their physiological development might be therefore affected. By representing explicitly the spatio-temporal variations of physiological rates and their dependency on environmental conditions, the model provides a valuable tool to investigate the complex interactions between population and environment and explains the wide range of observed growth rates and the dispersion of sizes at age.

The comparison of the predicted and observed reproduction and growth rates allows us to evaluate the model's performance. The comparison is based on the MONS province, which covers a large portion of the area exploited by industrial surface fisheries and where most of the data were collected. The mean observed batch fecundity of skipjack having a size between 40 and 60 cm ranges from 40 to 150 oocytes per g of biomass (Grande et al., 2010; Stéquert and Ramcharrun, 1996). These values are in the order of magnitude of the predicted batch fecundity (Eq. (22)), which is in the range of 10–70 oocytes per g biomass for fishes of comparable size and considering a spawning frequency of 2 days. Growth rates predicted by the model for the MONS province are also in good agreement with the IOTC growth curve (Fig. 6).

3.2. Population dynamics, environmental variability and fisheries

The monthly fishing mortality applied to the stock was calculated from the ratio catches/exploitable biomass, using values of simulated catches and defining the exploitable biomass as the biomass >30 cm. The fishing mortality was then compared to the size-dependent natural mortality due to predation and ageing (Eqs. (23) and (24)) (Fig. 7). At the beginning of industrial fisheries computed fishing mortalities were in the range of 0.003–0.005 month^{-1} and they have increased to 0.005–0.02 month^{-1} between 1996 and 2001. In comparison, total natural mortality of skipjack tuna ranges

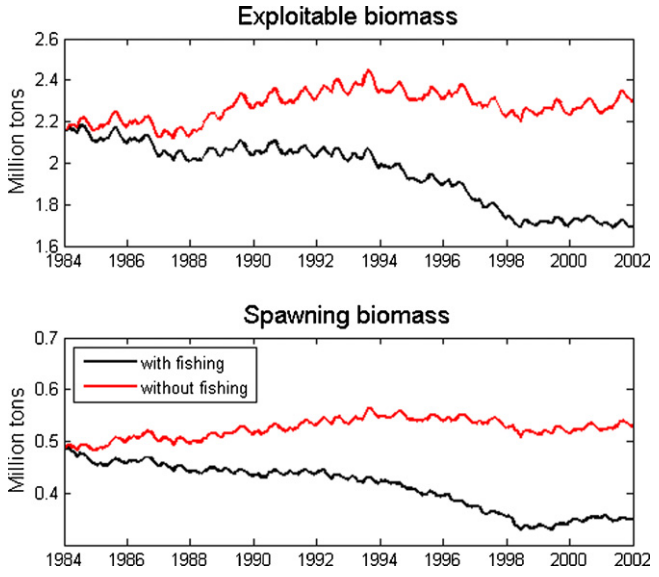

Fig. 9. Simulated exploitable and spawning skipjack biomass, with and without the effect of fishing mortality.

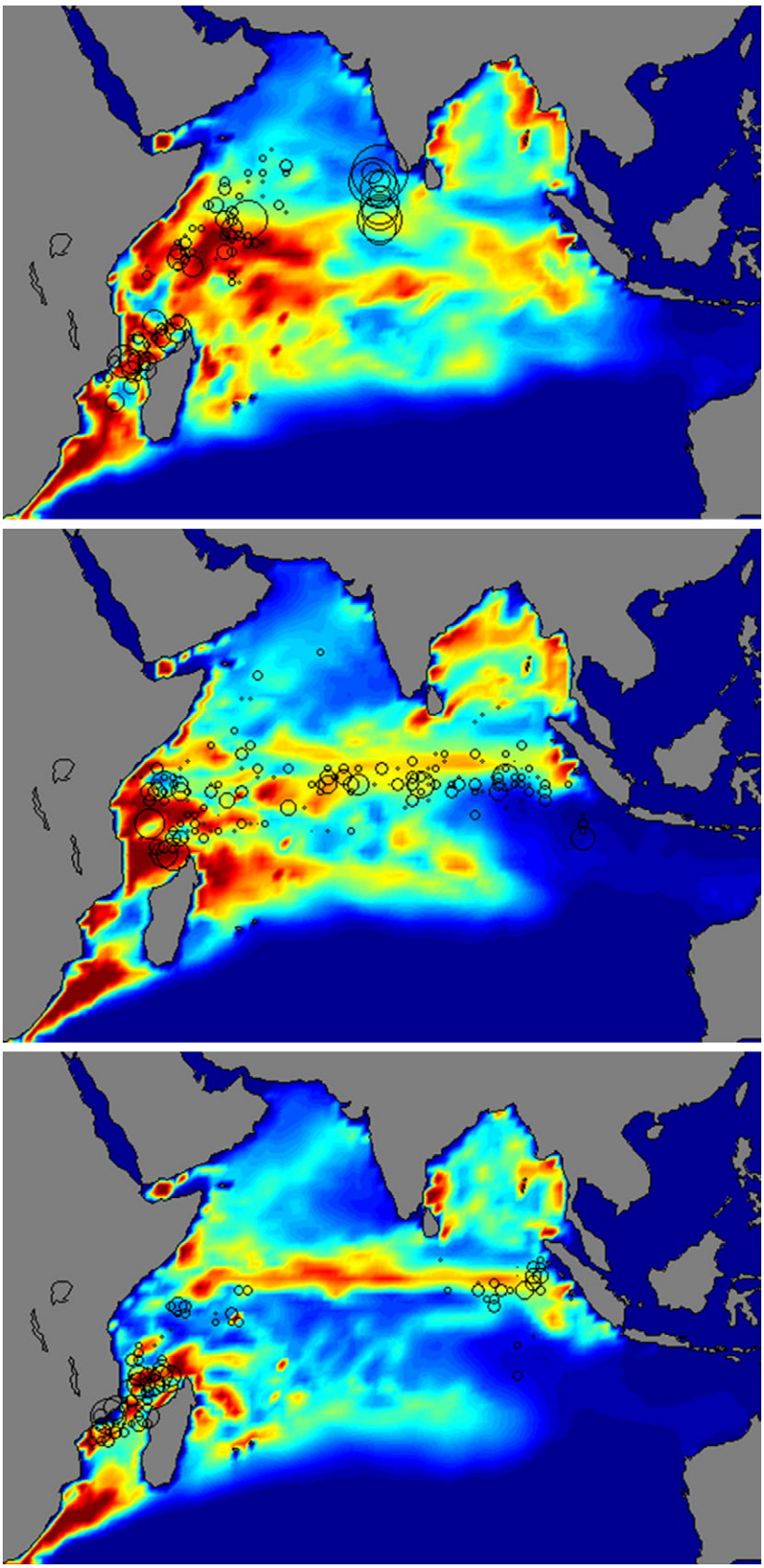

Fig. 10. Computed vertically integrated exploitable skipjack population vs observed catches (circles) in the Indian Ocean: April 1993 (top), February 1998 (center), April 1998 (bottom).

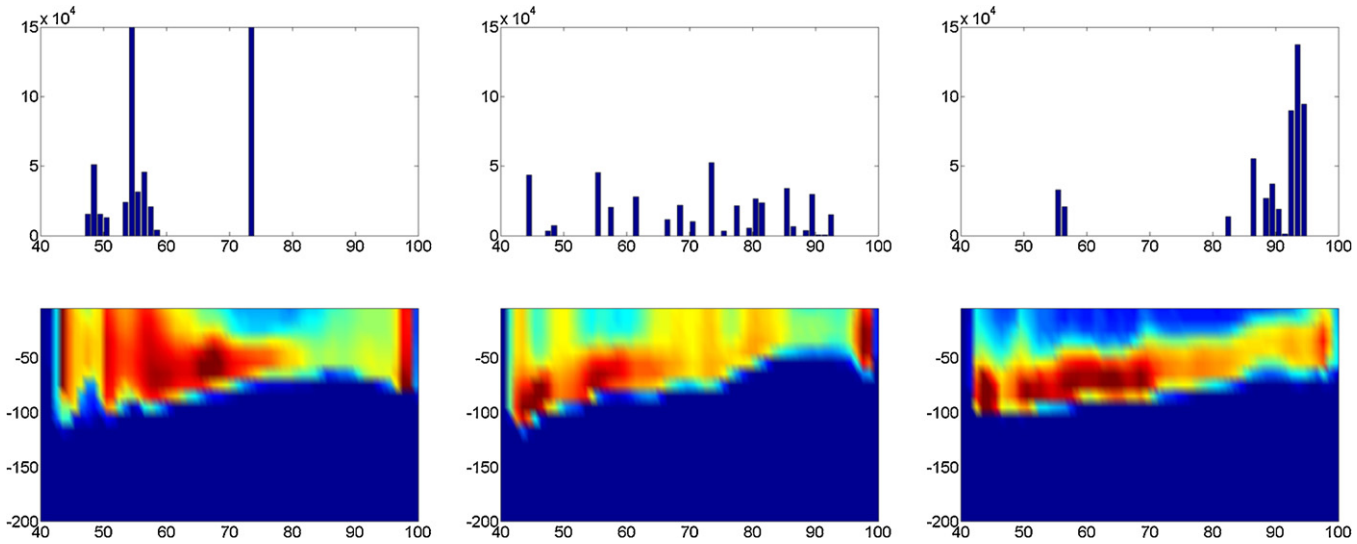

Fig. 11. Observed catches in the West-East transect on the Equator in April 1993 (top left), February 1998 (top center), April 1998 (top right) and corresponding computed vertical distribution of skipjack biomass on the same transect in April 1993 (bottom left), February 1998 (bottom center), April 1998 (bottom right).

between 0.085 and 0.125 month^{-1} for the fish sizes between 30 cm and 80 cm, corresponding to the exploited sizes. The model thus suggests that fishing mortality added approximately 20% of mortality to the natural mortality in the period between 1996 and 2001. Since catches have further increased in recent years this percentage has presumably risen.

The yearly recruitment, defined as the number of surviving fish entering the fishery, was calculated as the number of fishes reaching the exploitable size of 30 cm. The computed temporal dynamics of the yearly recruitment for the Indian Ocean is marked by four periods with increased recruitment: 1988, 1993, 1996 and 1999–2000 (Fig. 8). If we compare these peaks with the mean annual temperature in the top 100 m of the water column of the 3 biogeochemical provinces that cover most of the areas exploited by industrial fisheries in the Indian Ocean (MONS, ARAB, and INDW), we observe that the years with increased recruitment occur in general with a delay of one year with respect to the periods with higher temperature. The warmer periods accelerate the growth of larvae and juveniles and allow them to escape more quickly the small size domain where the predation mortality is the highest, thus improving survival and enhancing recruitment. This illustrates well how the dynamics of the population depends on the environment. In the model the recruitment is linked to the environmental factors by the bioenergetic representation of growth, reproduction and survival. Furthermore, we explore the relation between the predicted recruitment trends and the yearly total observed catches of the four fleets considered in the model. The temporal dynamics of reported catches shows three years of increased productivity that occur in 1989, 1995 and 1999, between 0 and 2 years after the recruitment peaks. The model suggests a strong link between environmental factors, recruitment and observed trends in fisheries. However, it has to be noted that other factors influence the temporal dynamics of catches such as the accessibility to the resource, which is also affected by the environment and by the technological development of fisheries.

In order to assess the impacts of exploitation on the population we compare simulations with and without the effects of fisheries. For this purpose, the model was run for the period 1958–2001 with the observed fishing effort of purse seine and bait boat fisheries starting after 1984. The starting year 1984 corresponds to the year when a simultaneous and important rise in both, industrial and artisanal fisheries begun in the Indian Ocean, leading to an increasing level of catches. The results indicate that the exploitation of the resource by industrial fisheries induced a marked decrease of both exploitable and spawning biomasses (Fig. 9). Compared to the simulation without the effect of fisheries, exploitation induces a reduction of 30% in spawning biomass and of 25% in exploitable biomass. Moreover the model indicates that the decrease of the population is not steady: periods of nearly stable biomass alternate with periods of steeper decrease. An important decrease is observed between 1991 and 1999 and is followed by a stable period. This stability is presumably induced by an increase of the recruitment during the years 1999–2000 due to environmental factors (the 1998 El Niño year, see next section), which have partially compensated the loss of biomass due to fishing mortality.

3.3. Spatial dynamics of the skipjack population compared to catches

The ability of the model to represent the spatial dynamics of the skipjack biomass under variable environmental conditions was also explored. Considering that the distribution of observed catches must be linked to the distribution of the resource, we compared the horizontal distribution of the accessible exploitable biomass (size > 30 cm, depth < 50 m) to the spatial distribution of observed catches (Fig. 10). In general, most of the skipjack catches are localized in the central-western part of the Indian Ocean and the main fishery areas are spread between the Somalian upwelling, the Mozambique Channel and the Maldivian Islands. Comparing the vertically integrated exploitable biomass to the catches distribution shows that the model is able to represent the main features of the horizontal distribution of the biomass at different periods of the year and under different environmental condition.

An examination of the vertical distribution of the simulated biomass (Fig. 11) shows that the biomass is not always located near the surface, where it is more accessible to fishers, but that habitat conditions can attract the skipjack biomass in deeper waters, usually between 50 m and 100 m depth. This limits the ability of the surface fleets to detect and catch the resource and implies that skipjack tunas might be present in certain zones but not fished.

The conjunction of a dipole mode (IODM) and a strong El Niño-Southern Oscillation (ENSO) event in 1997–1998 lead to important environmental anomalies in the Indian Ocean which had important consequences on fishing activities (Marsac and Le Blanc, 1998;

Ménard et al., 2007). The abnormal easterly wind stress along the equator caused the reversal of the E–W thermocline slope. The inversion of the normal thermocline E–W slope depth increased the catchability of tropical tuna for purse seine gears in the east and decreased it in the West where they are normally fished. This lead to massive and very unusual movement of the fishing fleets to the eastern area while the usual fishing grounds in the western part of the ocean were deserted by fishers. This exceptional event was well captured by the model as important changes in both the horizontal and vertical distribution of skipjack are clearly visible in the simulation in 1998 (Figs. 10 and 11). The simulations indicate that a substantial part of the biomass was still present in the western basin but since skipjack habitat was deeper than usual, the resource likely remained in deeper waters (50–100 m) thus less detectable and accessible to surface fisheries. On the contrary, biomass was closer to the surface in the central and eastern parts of the basin, therefore increasing the catchability in these sectors.

4. Conclusion

The comparison of the model and the data has reinforced the model and has highlighted the importance of the joint effect of environmental factors and exploitation by fisheries for the assessment of population dynamics. The model allowed integrating the spatio-temporal variability of temperature, dissolved oxygen and food conditions and evaluating their effect on the population. It has shown how environmental factors produce differences in growth and reproduction at the basin scale and how this can affect the size-at-age of skipjack tuna. Moreover the model allowed computing the temporal dynamics of mortality and recruitment and therefore provides a means to bridge the gap between environmental variability and observed temporal dynamics of total catches. This kind of analysis is essential to increase our understanding of the ecosystem's dynamics, since it allows a better interpretation of the data.

The model was also able to represent the environmentally driven spatial variability of the skipjack tuna population in the Indian Ocean and a good overlap of the simulated spatial distribution of biomass with observed fishing data distribution was observed, even during the extreme ENSO event that occurred in 1997–1998. Results have highlighted the influence of the environmental conditions on the horizontal and vertical distribution of skipjack tuna, and their effect on the accessibility of the resource to fisheries.

Acknowledgements

This work was supported by the Pelagic Fisheries Research Programme (PFRP) of the University of Hawaii, the AMPED project (www.amped.ird.fr) through a grant from the French National Research Agency (ANR), Systerra Programme, grant number ANR-08-STRA-03 and the MACROES project (http://www.macroes.ird.fr/) through a grant from the French National Research Agency (ANR), CEP Programme, grant number ANR-09-CEP-003.

The fisheries data analyzed in this publication were obtained from the IOTC (http://www.iotc.org). We wish to acknowledge the contribution of the staff of the 'Observatoire Thonier' of the Mixed Research Unit 212 'Exploited Marine Ecosystems' (IRD) for data processing and management.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2012.02.007.

References

Adam, M.S., 2010. Declining catches of skipjack in the Indian Ocean—observation from the Maldives. In: Proceedings of the 10th Meeting of the Working Party on Tropical Tuna, Indian Ocean Tuna Commission, IOTC-2010-WPTT-09.

Anderson, R.C., Adam, M.S., Goes, J.I., 2011. From monsoon to mantas: seasonal distribution of Manta alfredi in the Maldives. Fisheries Oceanography 20 (2), 104–113.

Aumont, O., Bopp, L., 2006. Globalizing results from ocean in situ iron fertilization studies. Global Biogeochemical Cycles 20, GB2017, doi:10.1029/2005GB002591.

Barkley, R., Neill, W.H., Gooding, R.M.G., 1978. Skipjack tuna, Katsuwonus Pelamis, habitat based on temperature and oxygen requirements. Fishery Bulletin 76, 653–662.

Block, B.A., Stevens, E.D., 2001. Tuna: Physiology, Ecology and Evolution. Academic Press.

Brill, R.W., 1994. A review of temperature and oxygen tolerance studies of tunas pertinent to fisheries oceanography, movement models and stock assessments. Fisheries Oceanography 3, 204–216.

Brill, R.W., Dewar, H., Graham, J.B., 1994. Basic concepts relevant to heat transfer in fishes, and their use in measuring the physiological thermoregulatory abilities of tunas. Environmental Biology of Fishes 40, 109–124.

Brill, R.W., Lutcavage, M.E., 2001. Understanding environmental influences on movements and depth distributions of tunas and billfishes can significantly improve population assessment. American Fisheries Society Symposium 25, 179–198.

Cayré, P., Farrugio, H., 1986. Biologie de la reproduction du listao (Katsuwonus pelamis) de l'Océan Atlantique. In: Symons, P.E., Miyake, P.M., Sakagawa, G.T. (Eds.), Proc. ICCAT Conf. Intern. Skipjack Year Prog, pp. 253–272.

Dueri, S., Faugeras, B., Maury, O., 2012. Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E: Part 2. Parameter estimation and sensitivity analysis. Ecological Modelling 245, 55–64.

Faugeras, B., Maury, O., 2007. Modeling fish population movements: from an individual-based representation to an advection–diffusion equation. Journal of Theoretical Biology 247, 837–848.

Genin, A., 2004. Bio-physical coupling in the formation of zooplankton and fish aggregation over abrupt topographies. Journal of Marine Systems 50, 3–20.

Grande, M., Murua, H., Zudaire, I., Korta, M., 2010. Spawning activity and batch fecundity of skipjack Kastuwonus pelamis, in the Western Indian Ocean. In: Proceedings of the 10th Meeting of the Working Party on Tropical Tuna, Indian Ocean Tuna Commission, IOTC-2010-WPTT-47.

Gooding, R.M., Neill, W.H., Dizon, A.E., 1981. Respiration rates and low-oxygen tolerance limits in skipjack tuna, Katsuwonus pelamis. Fishery Bulletin 79, 31–48.

Indian Ocean Tuna Commission, 2005. Biological data on tuna and tuna-like species gathered at the IOTC Secretariat: Status Report. IOTC-2005-WPPT-05.

Indian Ocean Tuna Commission, 2008. Report of the tenth session of the IOTC Working Party on Tropical Tunas. Bangkok, Thailand, 23–31 October 2008. IOTC-2008-WPPT-R[E].

Indian Ocean Tuna Commission, 2010. Report of the twelfth session of the IOTC Working Party on Tropical Tunas. Victoria, Seychelles, 18–25 October 2010. IOTC-2010-WPTT-R[E].

Kooijman, S.A.L.M., 2000. Dynamic Energy and Mass Budgets in Biological Systems. Cambridge University Press.

Kooijman, S.A.L.M., 2010. Dynamic Energy Theory for Metabolic Organisation, third edition. Cambridge University Press.

Longhurst, A., 1998. Ecological Geography of the Sea. Academic Press, London.

Margulies, D., Suter, J.M., Hunt, S.L., Olson, R.J., Scholey, V.P., Wexler, J.B., Nakazawa, A., 2007. Spawning and early development of captive yellowfin tuna (Thunnus albacares). Fishery Bulletin 105, 249–265.

Marsac, F., Le Blanc, J.L., 1998. Interannual and ENSO-associated variability of the coupled ocean–atmosphere system with possible impacts on the yellowfin tuna fisheries in the Indian and Atlantic oceans. In: ICCAT Tuna Symposium, ICCAT Coll. Vol. Sci. Pap., L(1), pp. 345–377.

Maury, O., 2005. How to model the size-dependent vertical behaviour of bigeye (Thunnus obesus) tuna in its environment. Collect. Vol. Sci. Pap. ICCAT 57, 115–126.

Maury, O., Faugeras, B., Shin, Y., Poggiale, J., Ari, T.B., Marsac, F., 2007. Modeling environmental effects on the size-structured energy flow through marine ecosystems. Part 1: The model. Progress in Oceanography 74, 479–499.

Maury, O., 2010. An overview of APECOSM, a spatialized mass balanced "Apex Predators ECOSystem Model" to study physiologically structured tuna population dynamics in their ecosystem. In: Parameterisation of Trophic Interactions in Ecosystem Modelling, M.St. John, P. Monfray editors. Progress in Oceanography 84, 113–117.

Maury, O., Poggiale, J.C., submitted. From individuals to populations to communities: a dynamic DEB-based model of marine ecosystem size-spectrum including life-history diversity.

Ménard, F., Marsac, F., Bellier, E., Cazelles, B., 2007. Climatic oscillations and tuna catch rates in the Indian Ocean: a wavelet approach to time series analysis. Fisheries Oceanography 16 (1), 95–104.

Mugo, R., Saitoh, S.-I., Nihira, A., Kuroyama, T., 2010. Habitat characteristics of skipjack tuna (Katsuwonus pelamis) in the western North Pacific: a remote sensing perspective. Fisheries Oceanography 19 (5), 382–396.

Ortega, A., Mourente, G., 2010. Comparison of the lipid profiles from wild caught eggs and unfed larvae of two scombrid fish: northern bluefin tuna (Thunnus thynnus L., 1758) and Atlantic bonito (Sarda sarda Bloch, 1793). Fish Physiology and Biochemistry 36, 461–471.

Sasamal, S.K., 2006. Island mass effect around the Maldives during the winter months of 2003 and 2004. International Journal of Remote Sensing 27 (20), 5087–5093.

Schaefer, K.M., Fuller, D.W., Block, B.A., 2009. Vertical movements and habitat utilization of skipjack (Katsuwonus pelamis), yellowfin (Thunnus albacares), and bigeye (Thunnus obesus) tunas in the equatorial eastern Pacific Ocean, ascertained through archival tag. In: Nielsen, J.L., Arrizabalaga, H., Fragoso, N., Hobday, A., Lutcavage, M., Sibert, J. (Eds.), Reviews: Methods and Technologies in Fish Biology and Fisheries, Volume 9: Tagging and Tracking of Marine Animals with Electronic Devices. Springer, Netherlands, pp. 121–144.

Sharp, G.D., 1978. Behavioral and physiological properties of tunas and their effects on vulnerability to fishing gear. In: Sharp, G.D., Dizon, A.E. (Eds.), The Physiological Ecology of Tunas. Academic Press, New York, pp. 397–450.

Stéquert, B., Marsac, F., 1989. Tropical tuna—surface fisheries in the Indian Ocean. FAO Fisheries technical paper no. 282 Rome, Italy. 238 pp.

Stéquert, B., Ramcharrun, B., 1996. La reproduction du listao (Katsuwonus pelamis) dans le bassin ouest de l'océan Indien. Aquatic and Living Resources 9, 235–247.

UNOSAT, 2009. Analysis of Somali Pirate Activity in 2009, Available: http://unosat.web.cern.ch/unosat/freeproducts/somalia/Piracy/2009/UNOSAT_Somalia_Pirates_Analysis_Q1_2009_23April09_v1.pdf.

Valdemarsen, J.W., 2001. Technological trend in capture fisheries. Ocean & Coastal Management 44, 635–651.

Van Leer, B., 1977. Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. Journal of Computational Physics 23, 276–299.

Van der Veer, H.W., Kooijman, S.A.L.M., van der Meer, J., 2003. Body size scaling relationships in flatfish as predicted by Dynamic Energy Budgets (DEB theory): implications for recruitment. Journal of Sea Research 50 (2–3), 255–270.

Appendix

The reduced model

In this section we define characteristic quantities and use them to put the model in an adimensional form. Through this fast and slow dynamics appear which are characterized by a small parameter ε. We then derive a reduced model passing to the limit ε=0.

Adimensionalization

In what follows capital letters refer to characteristic scales (e.g. L_h for the horizontal length scale). For a variable with dimension, e.g. x, the notation $\tilde{x}$ refers to the adimentionalized variable.

Let us define the characteristic scales $x = L_h\tilde{x}$ and $y = L_h\tilde{y}$ in the horizontal dimension, $z = L_z\tilde{z}$ in vertical dimension, $V=S\tilde{V}$ in the structural volume dimension, and $t=T\tilde{t}$ for time. Let us also define for the biomass density state variable $p(x,y,z,V,t) = P\tilde{p}(\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t})$

Concerning the different processes involved in the dynamic of the system let us define:

- Horizontal velocity: $v = V_h\tilde{v}$
- Vertical velocity: $v_z = V_z v_z = \frac{B}{L_z}\tilde{b}\,\partial_{\tilde{z}}\tilde{h}$
- Horizontal diffusion: $d = D\tilde{d} = L_h V_h \tilde{d}$
- Vertical diffusion: $d_z = L_z V_z \tilde{d}_z$
- Growth: $g = G\tilde{g}$
- Mortality: $m + f = M(\tilde{m} + \tilde{f})$
- Newborn input: $r(p) = R\,\mathrm{P\tilde{r}}(\tilde{p})$

The density function $\tilde{p}$ follows:

$$\begin{cases} \partial_{\tilde{t}}\tilde{p} = \dfrac{TV_h}{L_h} d\tilde{i}v\left(\tilde{d}\tilde{\nabla}\tilde{p} - \tilde{v}\tilde{p}\right) + \dfrac{TV_z}{L_z}\partial_{\tilde{z}}\left(\tilde{d}_z\partial_{\tilde{z}}\tilde{p} - \tilde{v}_{\tilde{z}}\tilde{p}\right) \\ -\dfrac{TG}{S}\partial_{\tilde{V}}\left(\tilde{g}\tilde{p}\right) - TM\left(\tilde{m}+\tilde{f}\right)\tilde{p}, \quad in\ \tilde{\Omega}\times\left(\tilde{V}_{\min},\tilde{V}_{\max}\right)\times\left(0,\tilde{t}_{\max}\right), \\ p_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},0} = \tilde{p}^0{}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V}}, \quad \forall\left(\tilde{x},\tilde{y},\tilde{z},\tilde{V}\right)\in\tilde{\Omega}\times\left(\tilde{V}_{\min},\tilde{V}_{\max}\right) \\ \tilde{g}\tilde{p}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V}_{\min},\tilde{t}} = \dfrac{R}{G}\tilde{r}\left(\tilde{p}\right), \quad \forall\left(\tilde{x},\tilde{y},\tilde{z},\tilde{t}\right)\in\tilde{\Omega}\times\left(0,\tilde{t}_{\max}\right) \\ \nabla\tilde{p}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}}\cdot n\left(\tilde{x},\tilde{y},\tilde{z}\right) = 0, \quad in\ \partial\tilde{\Omega}, \quad \forall\left(\tilde{V},\tilde{t}\right)\in\left(\tilde{V}_{\min},\tilde{V}_{\max}\right)\times\left(0,\tilde{t}_{\max}\right) \end{cases} \tag{A.1}$$

where $\tilde{\nabla}$ and $d\tilde{i}v$ are the usual differential operators on $\tilde{\Omega}$.

Let us define a small parameter $\varepsilon = \dfrac{L_z}{L_h}$ and make the following assumptions:

- $V_z = \dfrac{V_h}{\varepsilon}$
- $T = \dfrac{L_h}{V_h}$
- $G = \varepsilon^4\dfrac{S}{T}$
- $TM = \varepsilon^2$
- $\dfrac{R}{G} = \varepsilon^2$

The main evolution equation reads

$$\begin{aligned} &\partial_{\tilde{t}}\tilde{p} = d\tilde{i}v\left(\tilde{d}\tilde{\nabla}\tilde{p} - \tilde{v}\tilde{p}\right) + \frac{1}{\varepsilon^2}\partial_{\tilde{z}}\left(\tilde{d}_z\partial_{\tilde{z}}\tilde{p} - \tilde{v}_{\tilde{z}}\tilde{p}\right) \\ &-\varepsilon^4\partial_{\tilde{V}}\left(\tilde{g}\tilde{p}\right) - \varepsilon^2\left(\tilde{m}+\tilde{f}\right)\tilde{p}, \quad in\ \tilde{\Omega}\times\left(\tilde{V}_{\min},\tilde{V}_{\max}\right)\times\left(0,\tilde{t}_{\max}\right) \end{aligned} \tag{A.2}$$

Reduction

Multiplying both sides of Eq. (A.2) by ε^2 and taking ε=0 leads to

$$\partial_{\tilde{z}}\left[\left(\tilde{a}\tilde{h}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}}\tilde{\omega}^{2/3}+\tilde{d}_{\phi}\right)\partial_{\tilde{z}}\tilde{p}-\tilde{b}\left(1-\tilde{h}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}}\right)\tilde{\omega}^{1/3}\partial_{\tilde{z}}\tilde{h}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}}\tilde{p}\right]=0 \tag{A.3}$$

With $\tilde{\chi}=\left(\frac{\tilde{V}}{\tilde{V}_{\max}}\right)$

The Neuman boundary condition imposed at the surface and at the maximum boundary depth $\left(\partial_{\tilde{z}}\tilde{p}_{\tilde{x},\tilde{y},0,\tilde{V},\tilde{t}}=\partial_{\tilde{z}}\tilde{p}_{\tilde{x},\tilde{y},\tilde{z}_{\max},\tilde{V},\tilde{t}}=0\right)$ leads to

$$\begin{cases}-\tilde{b}\left(1-\tilde{h}_{\tilde{x},\tilde{y},0,\tilde{V},\tilde{t}}\right)\tilde{\chi}^{1/3}\partial_{\tilde{z}}\tilde{h}_{\tilde{x},\tilde{y},0,\tilde{V},\tilde{t}}\tilde{p}_{\tilde{x},\tilde{y},0,\tilde{V},\tilde{t}}=c_{\tilde{x},\tilde{y},\tilde{V},\tilde{t}}\\ -\tilde{b}\left(1-\tilde{h}_{\tilde{x},\tilde{y},\tilde{z}_{\max},\tilde{V},\tilde{t}}\right)\tilde{\chi}^{1/3}\partial_{\tilde{z}}\tilde{h}_{\tilde{x},\tilde{y},\tilde{z}_{\max},\tilde{V},\tilde{t}}\tilde{p}_{\tilde{x},\tilde{y},\tilde{z}_{\max},\tilde{V},\tilde{t}}=c_{\tilde{x},\tilde{y},\tilde{V},\tilde{t}}\end{cases} \tag{A.4}$$

Where c is a constant independent of z

We assume that

$$\partial_{\tilde{z}}\tilde{h}_{\tilde{x},\tilde{y},0,\tilde{V},\tilde{t}}=\partial_{\tilde{z}}\tilde{h}_{\tilde{x},\tilde{y},\tilde{z}_{\max},\tilde{V},\tilde{t}}=0 \tag{A.5}$$

This assumption on $\tilde{h}$ is not a necessity at this stage but is needed for the integration of Eq. (A.2) at the end of the reduction process.

It follows that *c*=0 and

$$\partial_{\tilde{z}}\tilde{p}-\frac{\tilde{b}\left(1-\tilde{h}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}}\right)\tilde{\chi}^{1/3}}{\tilde{a}\ \tilde{h}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}}\tilde{\chi}^{2/3}+\tilde{d}_{\phi}}\partial_{\tilde{z}}\tilde{h}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}}\tilde{p}=0 \tag{A.6}$$

Integrating this first order differential equation on $(0,\tilde{z})$ leads to

47

$$
\begin{aligned}
&\widetilde{p}_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}} \\
&= \widetilde{p}_{\widetilde{x},\widetilde{y},0,\widetilde{V},\widetilde{t}} \exp\left(\int_0^{\widetilde{z}} \frac{\widetilde{b}\left(1-\widetilde{h}_{\widetilde{x},\widetilde{y},s,\widetilde{V},\widetilde{t}}\right)\widetilde{\chi}^{1/3}}{\widetilde{a}\widetilde{h}_{\widetilde{x},\widetilde{y},s,\widetilde{V},\widetilde{t}}\widetilde{\chi}^{2/3}+\widetilde{d}_{\phi}} \partial_{\widetilde{z}}\widetilde{h}_{\widetilde{x},\widetilde{y},s,\widetilde{V},\widetilde{t}}\, ds \right) \\
&= \widetilde{p}_{\widetilde{x},\widetilde{y},0,\widetilde{V},\widetilde{t}} \exp\left(\frac{\widetilde{b}\left(\widetilde{a}\widetilde{\chi}^{2/3}+\widetilde{d}_{\phi}\right)}{\widetilde{a}^2\widetilde{\chi}} \log\left(\frac{\widetilde{a}\widetilde{h}_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}}\widetilde{\chi}^{2/3}+\widetilde{d}_{\phi}}{\widetilde{a}\widetilde{h}_{\widetilde{x},\widetilde{y},0,\widetilde{V},\widetilde{t}}\widetilde{\chi}^{2/3}+\widetilde{d}_{\phi}} \right) - \frac{\widetilde{b}}{\widetilde{a}\widetilde{\chi}^{1/3}}\left(h_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}} - \widetilde{h}_{\widetilde{x},\widetilde{y},0,\widetilde{V},\widetilde{t}} \right) \right) \\
&= \widetilde{p}_{\widetilde{x},\widetilde{y},0,\widetilde{V},\widetilde{t}}\, \widetilde{e}_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}}
\end{aligned}
$$

48

49 (A.7)

50 where we have defined

$$
\widetilde{e}_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}} = \exp\left(\frac{\widetilde{b}\left(\widetilde{a}\widetilde{\chi}^{2/3}+\widetilde{d}_{\phi}\right)}{\widetilde{a}^2\widetilde{\chi}} \log\left(\frac{\widetilde{a}\widetilde{h}_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}}\widetilde{\chi}^{2/3}+\widetilde{d}_{\phi}}{\widetilde{a}\widetilde{h}_{\widetilde{x},\widetilde{y},0,\widetilde{V},\widetilde{t}}\widetilde{\chi}^{2/3}+\widetilde{d}_{\phi}} \right) - \frac{\widetilde{b}}{\widetilde{a}\ \widetilde{\chi}^{1/3}}\left(\widetilde{h}_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}} - \widetilde{h}_{\widetilde{x},\widetilde{y},0,\widetilde{V},\widetilde{t}} \right) \right)
$$

51

52

53 Let us define the reduced model state variable $\overline{p}_{\widetilde{x},\widetilde{y},\widetilde{V},\widetilde{t}}$, by integrating Eq. (A.7) on $\left(0,\widetilde{z}_{\max}\right)$

$$
\begin{aligned}
\overline{p}_{\widetilde{x},\widetilde{y},\widetilde{V},\widetilde{t}} &= \int_0^{\widetilde{z}_{\max}} \widetilde{p}_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}}\ d\widetilde{z} \\
&= \widetilde{p}_{\widetilde{x},\widetilde{y},0,\widetilde{V},\widetilde{t}} \int_0^{\widetilde{z}_{\max}} \widetilde{e}_{\widetilde{x},\widetilde{y},\widetilde{z},\widetilde{V},\widetilde{t}}\ d\widetilde{z}
\end{aligned}
$$

54 (A.8)

55 Given any variable, e.g. $\bar{g}$, the average along the profile $\tilde{e}$ can be defined as

$$
\overline{g} = \frac{\int_0^{\widetilde{z}_{\max}} \widetilde{g}\widetilde{e}\, d\widetilde{z}}{\int_0^{\widetilde{z}_{\max}} \widetilde{e}\, d\widetilde{z}}
$$

56 (A.9)

57

58 Using this notation let us integrate Eq. (A.2) over $\left(0,\widetilde{z}_{\max}\right)$ in order to derive an evolution
59 equation for $\overline{p}$

60 • Time derivative term:

61 $$\int_0^{\tilde{z}_{\max}} \partial_t \tilde{p}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}} d\tilde{z} = \partial_t \tilde{p}_{\tilde{x},\tilde{y},0,\tilde{V},\tilde{t}} \int_0^{\tilde{z}_{\max}} \tilde{e}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}} d\tilde{z} = \partial_t \bar{p} \tag{A.10}$$

62 • Horizontal advection term:

$$\begin{aligned} \int_0^{\tilde{z}_{\max}} \tilde{\nabla} \cdot \left(\tilde{v} \tilde{p}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}} \right) d\tilde{z} &= \tilde{\nabla} \cdot \int_0^{\tilde{z}_{\max}} \tilde{v} \tilde{p}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}} d\tilde{z} \\ &= \tilde{\nabla} \cdot \left(\frac{\int_0^{\tilde{z}_{\max}} \tilde{v} \tilde{e} d\tilde{z}}{\int_0^{\tilde{z}_{\max}} \tilde{e} d\tilde{z}} \tilde{p}_{\tilde{x},\tilde{y},0,\tilde{V},\tilde{t}} \int_0^{\tilde{z}_{\max}} \tilde{e} d\tilde{z} \right) \\ &= \tilde{\nabla} \cdot \left(\overline{vp} \right) \end{aligned} \tag{A.11}$$

63

64 • Horizontal diffusion term. A difficulty arises because the diffusion matrix is not a
65 constant. A small computation leads to:

66 $$\int_0^{\tilde{z}_{\max}} \tilde{\nabla} \cdot \left(\tilde{d} \tilde{\nabla} \tilde{p} \right) d\tilde{z} = \tilde{\nabla} \cdot \left(\bar{d} \tilde{\nabla} \bar{p} \right) + \tilde{\nabla} \cdot \left(\left(\tilde{\nabla} \bar{d} - \overline{\nabla \tilde{d}} \right) \bar{p} \right) \tag{A.12}$$

67 which we approximate neglecting the new advection term

68 $$\int_0^{\tilde{z}_{\max}} \tilde{\nabla} \cdot \left(\tilde{d} \tilde{\nabla} \tilde{p} \right) d\tilde{z} \approx \tilde{\nabla} \cdot \left(\bar{d} \tilde{\nabla} \bar{p} \right) \tag{A.13}$$

69 • Computations for growth and mortality terms, as well as for the newborn input term,
70 are straightforward

71 • Thanks to the Neuman boundary condition on $\tilde{p}$ and to the assumption of Eq. (A.1)
72 the advection and diffusion terms in *z* vanish.

73 The reduced model for $\bar{p}$ finally reads

74 $$\begin{cases} \partial_{\tilde{t}} \bar{p} = d\tilde{i} v \left(\bar{d} \tilde{\nabla} \bar{p} - \overline{vp} \right) - \varepsilon^4 \partial_{\tilde{V}} \left(\overline{gp} \right) - \varepsilon^2 \left(\bar{m} + \bar{f} \right) \bar{p}, \quad in\ \tilde{\omega} \times \left(\tilde{V}_{\min}, \tilde{V}_{\max} \right) \times \left(0, \tilde{t}_{\max} \right), \\ \bar{p}_{\tilde{x},\tilde{y},\tilde{V},0} = \bar{p}^{0}_{\tilde{x},\tilde{y},\tilde{V}}, \quad \forall \left(\tilde{x}, \tilde{y}, \tilde{V} \right) \in \tilde{\omega} \times \left(\tilde{V}_{\min}, \tilde{V}_{\max} \right) \\ \overline{gp}_{\tilde{x},\tilde{y},\tilde{V}_{\min},\tilde{t}} = \varepsilon^2 \tilde{r}(\bar{p}), \quad \forall \left(\tilde{x}, \tilde{y}, \tilde{t} \right) \in \tilde{\omega} \times \left(0, \tilde{t}_{\max} \right) \\ \nabla \bar{p}_{\tilde{x},\tilde{y},\tilde{V},\tilde{t}} \cdot n(\tilde{x}, \tilde{y}) = 0, \quad \forall \left(\tilde{V}, \tilde{t} \right) \in \tilde{\omega} \in \left(\tilde{V}_{\min}, \tilde{V}_{\max} \right) \times \left(0, \tilde{t}_{\max} \right) \end{cases} \tag{A.14}$$

where position $(\tilde{x},\tilde{y}) \in \tilde{\omega}$ is a bounded domain representing the Indian Ocean in 2D. One can notice that the newborn input function $\tilde{r}$ is kept unchanged by the averaging process. The 3D form of the reduced state variable can be recalculated at any time by combining Eq. A.7 and A.8, and obtain:

$$\tilde{p}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}} = \frac{\bar{p}_{\tilde{x},\tilde{y},\tilde{V},\tilde{t}}}{\int_0^{\tilde{z}_{\max}} \tilde{e}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}}\, d\tilde{z}} \tilde{e}_{\tilde{x},\tilde{y},\tilde{z},\tilde{V},\tilde{t}} \tag{A.16}$$

The reduced model (Eq. A.14) can be put back into dimensional form resulting in the following model.

$$\begin{cases} \partial_t \bar{p} = div\left(\bar{d}\nabla\bar{p} - \overline{vp}\right) - \varepsilon^4 \partial_V\left(\overline{gp}\right) - \varepsilon^2\left(\bar{m} + \bar{f}\right)\bar{p}, \quad in\ \omega \times \left(V_{\min}, V_{\max}\right) \times \left(0, t_{\max}\right), \\ \bar{p}_{x,y,V,0} = \bar{p}^0{}_{x,y,V}, \quad \forall (x,y,V) \in \omega \times \left(V_{\min}, V_{\max}\right) \\ \overline{gp}_{x,y,V_{\min},t} = \varepsilon^2 r\left(\bar{p}\right), \quad \forall (x,y,t) \in \omega \times \left(0, t_{\max}\right) \\ \nabla\bar{p}_{x,y,V,t} \cdot n(x,y) = 0, \qquad \forall (V,t) \in \left(V_{\min}, V_{\max}\right) \times \left(0, t_{\max}\right) \end{cases} \tag{A.15}$$

Article J : [12] S. Dueri, B. Faugeras et O. Maury. Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E : Part 2. Parameter estimation and sensitivity analysis. *Ecological Modelling* 245 (2012), p. 55 –64

Ecological Modelling 245 (2012) 55–64

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E – Part 2: Parameter estimation and sensitivity analysis

Sibylle Dueri[a,*], Blaise Faugeras[b], Olivier Maury[a]

[a] Institut de Recherche pour le Développement (IRD), UMR EME 212 (IRD/Ifremer/Université Montpellier 2), Centre de Recherche Halieutique Méditerranéenne et Tropicale, Avenue Jean Monnet BP 171, 34203 Sète CEDEX, France
[b] CNRS, Laboratoire J.A. Dieudonné (UMR 7351), Université de Nice Sophia-Antipolis, Faculté des Sciences, Parc Valrose, 06108 Nice Cedex 02, France

ARTICLE INFO

Article history:
Available online 21 March 2012

Keywords:
Tropical tuna
Pelagic fisheries
Optimization
Automatic differentiation
Parameter identifiability

ABSTRACT

This paper presents the parameter estimation and sensitivity analysis of the APECOSM-E model, which describes the basin scale 3D size-structured population dynamics of the skipjack tuna population in the Indian Ocean under the joint effect of environmental variability and fisheries exploitation. The model is presented in detail in the companion paper (Dueri et al., 2012). A common methodology based on the evaluation of a cost function that combines the negative log-likelihoods of commercial catches and size frequencies is used for both tasks. A Bayesian term representing the a priori probabilities about the model parameters is added to the cost function used for parameter estimation. The partial derivatives of the cost function with respect to the parameters are obtained by deriving the tangent linear code of the model by automatic differentiation of the direct code. A minimization algorithm is used to estimate the parameters related to fisheries and maximise the fitness of the model to the available observations. In a second step, we evaluate the local sensitivity of the non-estimated parameters and identify the model parameters that have an important effect on the output of the model and that would deserve better measurements in order to reduce the level of uncertainty in the model outputs. The comparison between the optimized simulation and the observations allows identifying the model's strengths and limitations, in the perspective of using the model to test scenarios concerning the resilience of the population and the sustainability of its exploitation in the Indian Ocean.

1. Introduction

Artisanal fisheries have exploited skipjack tuna (*Katsuwonus pelamis*) for several centuries (Marine Research Section, 1996) in the Indian Ocean. However, since the early 1980s the level of exploitation has dramatically increased due to the introduction of industrial purse seining and the concurrent raise of bait boat and gillnet catches. Currently, the catches of this highly migratory tropical fish represent almost half of the tropical tuna catches in the Indian Ocean. Although skipjack tuna are considered to have a high resilience against overfishing due to their fast growth rate and their year round spawning, the decrease of catches recently reported by the Maldivian fishery, one of the leading skipjack fisheries in the Indian Ocean, raises concern about the sustainability of present levels of exploitation (Adam, 2010).

Mechanistic mathematical models are useful tools for the evaluation of trends in marine ecosystems. These models allow improving our understanding of the vulnerability and forecasting the ecosystem's evolution under future scenarios of climate change and increased exploitation. As a result, they are a source of valuable information supporting the development of policies for the sustainable management of marine resources. In the field of marine science the development and validation of such models is particularly challenging given the difficulties and cost of obtaining direct field observations (Blackford et al., 2010). However, given the increasing anthropogenic pressures that affect marine ecosystems the development of reliable mathematical models for marine ecosystems based on a mechanistic representation of processes is becoming a high priority issue (Brierley and Kingsford, 2009; Jackson, 2010).

Ecosystems are complex and non stationary systems and recent technological improvements in the field of computation have supported the development of numerical models that explicitly account for this complexity. The rigorous representation of the important mechanisms is certainly a fundamental requisite of every model that aims at deepening the understanding of processes and describing the future evolution of the ecosystem. However, increased complexity of the models is often associated with a larger number of parameters that are difficult to constrain and this can increase the uncertainty of the simulated results (Anderson, 2010).

DOI of original article: 10.1016/j.ecolmodel.2012.02.007.
* Corresponding author. Tel.: +33 0 499 57 32 53; fax: +33 0 499 57 32 95.
E-mail address: sibylle.dueri@ird.fr (S. Dueri).

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2012.02.008

Parameters are characterised by an inherent uncertainty linked to the paucity of comprehensive and accurate data which would be needed to constrain the model. Assessing the effects of these uncertainties in the model parameter values is an essential step for model evaluation. Sensitivity analysis provides a tool for this purpose as it allows evaluating the relative importance of the model parameters on the response of the model. Furthermore, it can help to investigate the way uncertainty propagates through the model to its outputs and it allows distinguishing between factors that would deserve better measurements in order to reduce the uncertainty in the model outputs and factors that are less influential (Cariboni et al., 2007).

The optimization of model parameters against available data is essential for improving the realism of the model and identifying the domains where it is not performing well. Optimizing the model parameters consists in finding the parameters that allow the best fit to the data. It can be achieved using a parameter estimation algorithm to minimize the distance between the simulation and the observed data. However, before starting the optimization we have to ensure that the inverse problem is well-posed and that the parameters can be estimated independently and accurately from the available data. Therefore, parameter estimation has to be preceded by the assessment of the parameters identifiability.

In this paper we present the parameter estimation and the sensitivity analysis of the APECOSM-E model, a deterministic model that represents the basin scale 3D population dynamics of skipjack tuna under the joint effect of environmental conditions and exploitation by fisheries (Dueri et al., 2012.). A common methodology based on the evaluation of the partial derivatives of the cost function with respect to the model parameters is used for sensitivity analysis and parameter estimation. It relies on the tangent linear code of the model derived by automatic differentiation of the direct code. The first objective is to estimate the parameter values that maximise the fitness of the model to the available observations. However, the high number of model parameters makes it difficult to estimate all the parameters at once. Since the available observations are related to fisheries (catch and size frequency data), we estimate only the fisheries-related parameters. In a second step, we characterize which of the non-optimized parameters are sensitive and influence the model output. The comparison between the optimized simulation and the observations allows identifying the strengths and limitations of the model and constitutes the basis for improvement, in view of using the model to test scenarios concerning the resilience of the population and the sustainability of its exploitation in the Indian Ocean.

2. Methodology

2.1. Model description

APECOSM-E (Apex-Predator-Ecosystem-Model – Estimation) is a deterministic model that represents the basin scale 3D population dynamics of skipjack tuna under the joint effect of environmental conditions and exploitation by fisheries (Dueri et al., 2012). This model is a simplified version of the more general APECOSM framework (Maury, 2010), which represents the flow of energy through the global marine ecosystem considering different communities of epipelagic and mesopelagic organisms. APECOSM-E is a single-species version devoted to parameter estimation. In the model, temperature, oxygen, food and oceanic currents affect the physiology of tuna (growth, reproduction and mortality), their movements and the spatial distribution of the population. Observed spatial fishing effort data are linked to fishing mortality and are used to simulate monthly catches and size frequencies.

The model is structured in 3D space and fish size and considers size dependent reproduction, growth, predation, natural mortality and fishing mortality. It is based on a single partial differential equation describing the change of the population density function p as a function of time:

$$\partial_t p = div(d\nabla p - vp) + \partial_z(d_z\partial_z p - v_z p) - \partial_V(gp) - (m+f)p \quad (1)$$

in $\Omega \times (w_{\min}, w_{\max}) \times (0, t_{\max})$. The four terms on the right side of Eq. (1) represent: (1) horizontal advection and diffusion, (2) vertical advection and diffusion, (3) growth and (4) natural and fishing mortality. Advection and diffusion are represented by the horizontal velocity $v(x, y, z, V, t)$ [m s^{-1}], the vertical velocity $v_z(x, y, z, V, t)$ [m s^{-1}], the horizontal diffusion $d(x, y, z, V, t)$ [m^2 s^{-1}] and the vertical diffusion $d_z(x, y, z, V, t)$ [m^2 s^{-1}]. Here we assume that there is no cross-diffusion term in z. Growth is represented as an advection of the biomass density in the size dimension and is characterised by the growth rate $g(x, y, z, V, t)$ [m^3 s^{-1}], while natural and fishing mortality rates are described by $m(x, y, z, V, t)$ [s^{-1}] and $f(x, y, z, V, t)$ [s^{-1}] respectively.

Processes are all time, space and size-dependent and linked to the environment through mechanistic bioenergetic or behavioural parameterizations. Physiological rates such as growth, reproduction and ageing mortality are described consistently with the Dynamic Energy Budget (DEB) theory (Kooijman, 2000). Both horizontal and vertical movements are driven by habitat gradients, oceanographic currents and physical diffusion. Horizontal processes are modelled using the mechanistic approach developed in Faugeras and Maury (2007) which enables to consistently relate advection and diffusion.

The model has 48 parameters (Table 1), 18 of which are related to fisheries and define length and depth selectivity of the fishing gears, catchability and increasing efficiency due to technological development of different fleets, 8 are DEB parameters describing the kinetic of growth, reproduction and ageing mortality and 22 are ecological parameters describing the interaction between the environment and the population.

APECOSM-E is integrated numerically on a 1° by 1° horizontal grid covering the Indian Ocean and 20 vertical layers reaching 500 m depth, with a 10 m interval in the first 150 m. A more detailed description of the model is presented in a companion paper (Dueri et al., 2012).

2.2. Fishery data

The model requires three types of fishery time series: fishing effort, catch and size frequency data. Observed effort data are used to impose the effect of fishing activities on the population: spatially explicit fishing effort is applied to the simulated biomass to produce simulated catches and size frequencies. On the other hand, observed catch and size frequency data are required for the computation of the cost function that is used for parameter estimation. These datasets are obtained from the Indian Ocean Tuna Commission in the standardized form which is available on the CLIOTOP MDST website (http://vmmdst-proto.mpl.ird.fr/MDST/). Observed monthly catch and effort data are spatially aggregated over a 1° by 1° grid, while monthly size frequencies are spatially aggregated over a 5° by 5° grid.

The model considers four different fleets: French purse seiners "PS1", Spanish purse seiners "PS2", "World" purse seiners "PS3" (combining the fishing data of Mauritius, Seychelles and NEI-other) and Maldivian bait boats "BB". These four fleets represent the main skipjack fisheries of the Indian Ocean providing a reliable time series of fishing data from the beginning of industrial fisheries in 1984. Other fleets, such as the gillnet fleet, had to be excluded due to the uncertain quality of their data.

Table 1
Model parameters.

Parameter	Description	Type
κ	Fraction allocated to soma	DEB
$\{\dot{p}_{Am}\}$	Surface-area specific assimilation rate	DEB
$[E_m]$	Maximum energy density of reserves	DEB
$[E_G]$	Volume-specific energetic growth cost	DEB
$[\dot{p}M]$	Volume-specific maintenance cost	DEB
κ_R	Fraction reproduction energy fixed in eggs	DEB
l_{mat}	Length at maturity	DEB
$\ddot{h}_a$	Ageing acceleration	DEB
a_v	Maximal horizontal speed	Ecology
α	Concentration factor coefficient	Ecology
$a_{v,z}$	Maximal speed, vertical	Ecology
d^{ϕ}	Physical diffusivity, vertical	Ecology
T_0	Metabolic energy production/thermal capacity	Ecology
k_{T}	Thermic conductance/thermal capacity	Ecology
T_{a}	Arrhenius temperature	Ecology
T_{l}	Reference temperature	Ecology
T_{l}	Lower boundary of tolerance range	Ecology
T_{h}	Upper boundary of tolerance range	Ecology
T_{al}	Arrhenius temperature for lower boundary	Ecology
T_{ah}	Arrhenius temperature for upper boundary	Ecology
k_{G}	Half saturation constant for forage	Ecology
a_{O}	Steepness of oxygen limitation curve	Ecology
O_0	Half saturation constant for oxygen limitation	Ecology
p_{T}	Weighting factor temperature	Ecology
p_{F}	Weighting factor forage	Ecology
p_0	Weighting factor oxygen	Ecology
m_{p1}	Predation mortality coefficient 1	Ecology
m_{p2}	Predation mortality coefficient 2	Ecology
m_{T1}	Temperature mortality coefficient	Ecology
a_{mdv}	Maximal attraction factor for Maldives	Ecology
$l_{l,\mathrm{ps1}}$	Length selectivity, PS1	Fishery
$l_{s,\mathrm{ps2}}$	Length selectivity, PS2	Fishery
$l_{s,\mathrm{ps3}}$	Length selectivity, PS3	Fishery
$l_{s,\mathrm{bb}}$	Length selectivity, BB	Fishery
$k_{l,\mathrm{ps}}$	Steepness length selectivity, PS	Fishery
$k_{l,\mathrm{bb}}$	Steepness length selectivity, BB	Fishery
$z_{s,\mathrm{ps1}}$	Depth selectivity, PS	Fishery
$z_{s,\mathrm{bb}}$	Depth selectivity, BB	Fishery
$k_{z,\mathrm{ps}}$	Steepness depth selectivity, PS	Fishery
$k_{z,\mathrm{bb}}$	Steepness depth selectivity, BB	Fishery
p_{ps1}	Catchability PS1	Fishery
p_{ps2}	Catchability PS2	Fishery
p_{ps3}	Catchability PS3	Fishery
p_{bb}	Catchability BB	Fishery
a_{ps1}	Increased efficiency, PS1	Fishery
a_{ps2}	Increased efficiency, PS2	Fishery
a_{ps3}	Increased efficiency, PS3	Fishery
a_{bb}	Increased efficiency, BB	Fishery

2.3. Cost function components

The quantitative comparison of the model results with available observations is essential for the assessment of the model strengths and weaknesses. In the present study we used time series of catches and size frequencies of different fishing fleets for the computation of a cost function that quantifies the discrepancy between simulated and observed data using the method developed by Faugeras and Maury (2005).

In order to calculate the cost function, we first have to generate model outputs that are formally consistent with the available observations, i.e. monthly 1 degree square catches and 5 degree square size frequencies per fleet k. For this purpose, in each cell i of the grid where the observed fishing effort is positive, we calculate the daily catch by multiplying the size- and depth dependent fishing mortality f by the corresponding total biomass density p. Total monthly catches per 1 degree square and per fleet, C_k, are computed by integrating daily catches over the vertical domain Δz, over all fished size classes ΔV and over the number of days in a month $\Delta t(m)$. The resulting value comes with the unit of weight and is calculated per grid cell i and per month m.

$$C_k(i,m) = \sum_{z=1}^{nz}\sum_{V=1}^{nV}\sum_{t=1}^{nt(m)} f_k(i,z,V,t(m))\, p(i,z,V,t(m))\Delta x\Delta y\Delta z\Delta V\Delta t(m) \quad (2)$$

where the nz corresponds to the number of layer of the water column, nV to the number of size classes and $nt(m)$ to the number of days per month.

Fishing mortality exerted by given fleet k is calculated as the product of the observed fishing effort e_k by the catchability p_k at t_0 multiplied by an exponential function representing the increase of fishing efficiency at a rate a_k due to technological development in time, a size selectivity function and a depth selectivity function. Technological development is assumed to be continuous and includes the raise of the size and performance of the fishing vessels, the enhancement of the fishing gears, the progressive use of new electronic devices such as bird radar and other remote sensing tools and the deployment of more and more sophisticated fish aggregating devices (FADs) (Valdemarsen, 2001). The length and depth selectivity of the different gears are represented using two sigmoid functions where l_S and z_S are the length and depth leading to 50% selection while k_l and k_z define the steepness of the sigmoid curves.

$$f_k(i,z,V,t) = e_k(i,t)p_k\,\exp(a_k t) \times \frac{1}{1+\exp(-k_l(V^{1/3}/\delta_M - l_s))}\,\frac{1}{1+\exp(k_z(z-z_s))} \quad (3)$$

Similarly, for each cell of the grid with positive fishing effort, size frequencies per fleet are calculated per grid cell, month and size.

$$Q_k(i,m,V) = \frac{\sum_{z=1}^{nz}\sum_{t=1}^{nt} f_k(i,z,V,t)p(i,z,V,t)\Delta x\Delta y\Delta z\Delta V\Delta t}{\sum_{z=1}^{nz}\sum_{V=1}^{nV}\sum_{t=1}^{nt} f_k(i,z,V,t)p(i,z,V,t)\Delta x\Delta y\Delta z\Delta V\Delta t} \quad (4)$$

Since the observed size frequencies are given on a 5° by 5° grid while the model runs with a 1° by 1° grid, we need to further integrate Q_k in order to obtain the same level of horizontal resolution as the data.

We now define the vector of the model parameters K and the components of the cost function related to catch $J_C(K)$, size frequencies $J_Q(K)$ and parameters $J_P(K)$. For estimating the parameters the total cost is obtained by summing the three components, while for sensitivity analysis the cost function is determined by the sum of the first two components (excluding the priors on the parameter values). The components of the cost function are expressed as the negative log-likelihoods, which measure the distance between observed and simulated responses. Likelihoods are then summed over all fleets, months and horizontal grid squares. Assuming that the observation error for catch data follows a log-normal distribution, we calculate the total cost of catches as:

$$J_C(K) = \sum_k \frac{1}{2\sigma_{C,k}^2}\sum_i\sum_m \left(\log(C_k(i,m)) - \log(C_k^{obs}(i,m))\right)^2 \quad (5)$$

where $\sigma_{C,k}$ is the fleet dependent standard deviation for catches.

Length frequencies are assumed to exhibit a normally distributed observation error. Their contribution to the cost function is therefore expressed as:

$$J_Q(K) = \sum_k \frac{1}{2\sigma_{Q,k}^2}\sum_i\sum_l\sum_m \left(Q_k(i,m,l) - Q_k^{obs}(i,m,l)\right)^2 \quad (6)$$

where $\sigma_{Q,k}$ is the fleet dependant standard deviation for size frequency.

The Bayesian component of the cost function accounts for the differences between the initial values of the parameters and the new value. This allows us to estimate the parameters in a Bayesian context by computing the mode of the posterior density function of the parameters knowing the data. Assuming that the a priori distribution of parameters is normally distributed, the contribution of the n values of the parameter vector K to the cost function is

$$J_p(K) = \sum_n \frac{1}{2\sigma_n^2}(K_n - K_n^0)^2 \tag{7}$$

where K^0 is the initial guess of the parameter. Since parameters have different units and magnitude we need to adimensionalize the vector by dividing each parameter by the corresponding initial guess.

2.4. Parameter identifiability

Before attempting to solve an inverse problem that involves constraining parameters to fit observations, it is important to ensure that the problem is well posed and that the parameters can be estimated accurately and independently from the available dataset. For this purpose, we use an approach based on the computation of the Hessian matrix that allows assessing the identifiability of the parameters given the conceptual representation of the phenomena provided by the model (Thacker, 1989). This method has been successfully implemented by Fenner et al. (2001) and Faugeras et al. (2003) to evaluate parameter identifiability of marine ecosystem models representing biogeochemical and plankton dynamics.

The elements that compose the Hessian matrix are the second derivatives of the cost function with respect to the parameters. The cost function used for identifiability comes without penalty term and is called J^S to distinguish it from the general cost function J with penalty used for optimization (see next section). J^S is composed of two terms, the negative log-likelihood of catches and the negative log-likelihood of size frequencies.

$$J^S(K) = J_C(K) + J_Q(K) \tag{8}$$

Near the global minimum, the matrix provides key indicators of the convergence and uncertainty related to optimization (Thacker, 1989). The condition number of the Hessian defined as the ratio of its largest to smallest eigenvalue, determines the rate of convergence of the minimization algorithm. For large values of the condition number the matrix is ill conditioned and nearly singular. The off-diagonal elements of the Hessian matrix correspond to the degree of correlation of pairs of parameters. Eigenvectors and eigenvalues provide important information on the uncertainties related to parameter estimation. Small eigenvalues indicate large uncertainties in the identification of parameters that make a significant contribution in the related eigenvector. Furthermore, the inverse of the Hessian matrix provides an approximation of the covariance matrix of the model parameters.

Here, we want to estimate the model parameters related to fishing activities by using the likelihood of catch and size frequency data. To ensure that the desired parameters can be identified independently and are sufficiently constrained by the data we compute the Hessian matrix of the cost function without penalty term at the global minimum. The minimum is obtained by running the model with a known set of parameters K^0 using simulated catch and size frequencies time series instead of observations in the calculation of the cost function. Running the simulation with the same parameter set we obtain a perfect match with the synthetic observations, and the cost function is equal to zero. The Hessian of the cost function can then be calculated by means of a central finite difference scheme approximation that uses the exact gradients to calculate the second derivatives:

$$\frac{\partial^2 J^S}{\partial K_i \partial K_j} \approx \frac{g_i(K + \delta K_j \varepsilon) - g_i(K - \delta K_j \varepsilon)}{4\varepsilon} + \frac{g_j(K + \delta K_i \varepsilon) - g_i(K - \delta K_i \varepsilon)}{4\varepsilon} \tag{9}$$

where ε is the step size, δK is the direction of perturbation and g is the exact gradient of the cost function provided by automatic differentiation

$$g_n(K) = \frac{\partial J^S}{\partial K_n} \tag{10}$$

2.5. Parameter estimation

Parameter estimation in the APECOSM-E model is based on the simultaneous minimization of the three terms of the cost function: the negative log-likelihood of catches, the negative log-likelihood of length frequencies and the log of the prior density function of the parameters.

$$J(K) = J_C(K) + J_Q(K) + J_P(K) \tag{11}$$

The minimization of the cost function is implemented using the n1qn3 Fortran subroutine of Gilbert and Lemaréchal (1989). This gradient-based minimization algorithm requires the calculation of the exact gradient of the cost function with respect to the parameter being estimated. For this purpose, we derive the tangent linear code by means of an automatic differentiation engine, called TAPENADE (Hascoët and Pascual, 2004), which is developed by the French National Institute for Research in Computer Science and Control (INRIA) and freely available on-line (http://tapenade.inria.fr).

Before using the tangent linear code generated by TAPENADE, it is necessary to test it by comparing the exact gradient given by the automatically differentiated code to its finite difference approximation. For this purpose, we use the Taylor test implemented by Faugeras and Maury (2005) and compute the ratio of the finite difference approximation of the gradient to the exact derivative. This test ensures that the differentiated code provides the correct derivative if for a decreasing parameter perturbation, the finite difference formulation tends to the value of the exact gradient (and their ratio tends towards 1). A previous evaluation of the model has shown that the lower limit of perturbation for which this is true is 10^{-6}, which corresponds to the precision level expected from the finite difference computation given the truncation error (Faugeras and Maury, 2005).

The parameter estimation is performed over a temporal window of 10 years, from 1984 to 1993. This window corresponds to a period of time for which a complete dataset of fleet-specific monthly catches is available for all fleets. After 1993 catches of the Maldivian bait boat are reported on a yearly basis and are therefore less suitable for parameter estimation. This furthermore allows to keep a large set of data (1994–2001) unused for parameter estimation, to assess the model predictions.

2.6. Sensitivity analysis

The sensitivity analysis measures the reaction of the model to small changes in the input parameters. By identifying the parameters that have a large effect on the output, it highlights the processes that drive the system dynamics and the parameters that should be defined accurately in order to increase the reliability of the model and its ability to forecast the evolution of the system under changing conditions.

Among the different methods, the one based on the automatic differentiation of the code belongs to the methods that are considered appropriate for sensitivity analysis of complex non-linear models with a large number of parameters (Cariboni et al., 2007;

Frey and Patil, 2002). Automatic differentiation enables to compute the first order partial derivatives of the output variables with respect to the input parameters. The value of the first derivative indicates the local sensitivity of the model outcome with respect to small changes in the parameter.

Here, we examine the sensitivity of the cost function without penalty term J^S with respect to small parameter perturbations. The relative sensitivity is defined as the derivative of the cost function with respect to parameters, multiplied by the value of the parameter and divided by the value of the cost function at the evaluation point K^1.

$$\frac{\partial J^S(K)}{\partial K}\frac{K^1}{J^S(K^1)} \tag{12}$$

We carry out the sensitivity analysis only for the parameters that are not considered for parameter estimation. In order to test the variability of the sensitivity over time, we split the simulated timeframe (1984–2001) in 3 periods of 6 years (1984–1989, 1990–1995 and 1996–2001) and calculate the relative sensitivity for each subset. This allows investigating the stability and robustness of the analysis over different periods and calculating the mean and standard deviation of the sensitivity.

Local sensitivities can be either positive or negative, depending on the sign of the derivative of Eq. (12). The absolute value of the local sensitivity informs us about the magnitude of the sensitivity. The relative sensitivity S is then obtained by dividing the absolute value of the local sensitivity by the sum of all sensitivities. Finally, we use this relative sensitivity to calculate the mean value of the sensitivity $\bar{S}$ and the standard deviation σ_S.

3. Results and discussion

3.1. Evaluation of parameter identifiability

In order to test the parameter identifiability, we compute the Hessian matrix of the cost function with respect to the parameters, at the global minimum and we calculate the condition number, the eigenvalues and eigenvectors of the Hessian. Since the optimization is based on fishing data, we assess the identifiability of 19 parameters, 18 parameters directly related to fishing activities plus one ecological parameter describing the attraction of the Maldivian Islands, which is important for the Maldivian fisheries (Dueri et al., 2012). This configuration leads to a condition number of the Hessian matrix equal to 4.2×10^6 indicating a poorly constrained inverse problem formulation. By far the two smallest eigenvalues are $\lambda_1 = 0.0248$ and $\lambda_2 = 0.19$. Looking at the corresponding eigenvectors we notice that eigenvector v_1 has significant contribution from $k_{z,ps}$ while eigenvector v_2 has significant contribution from $k_{z,bb}$ (Fig. 1). These parameters represent the steepness of the depth selectivity for purse seiners and bait boat, respectively, and the analysis suggests that they are poorly constrained by the available data and must be excluded from the optimization. The exclusion of these two parameters leads to a major improvement of the condition number which decreases to 2.8×10^4. Further exclusion of two more parameters contributing to the second eigenvector, namely the depth selectivity of bait boats $z_{s,bb}$ and the catchability of bait boats p_{bb} produces only a minimal improvement of the conditioning number (1.91×10^4), thus indicating that these two parameters can be kept in the optimization. This outcome indicates that 17 of the 19 parameters that we initially wanted to include in the parameter estimation can be reliably estimated with the available data.

3.2. Parameter optimization

The parameters estimated with the minimization algorithm include the fleet specific catchability p_k, the fleet specific increase in fishing efficiency a_k, the gear specific fishing length selectivity coefficients l_S and k_S, the gear specific depth selectivity coefficient z_s, and the parameter representing the attraction of the Maldivian Islands a_{mdv}. The parameters were estimated using the likelihoods of catch and size frequency data over a period of 10 years, from 1984 to 1993. The minimization algorithm converged after 122 iterations (Fig. 2). The parameters showing the largest relative changes in comparison to their initial values were the ones representing the catchability increase due to technological development of the three purse seine fleets (variations between 43% and 76%) and the parameter describing the steepness of the length selectivity of bait boat (43%). The parameters showing the smallest variation were the ones related to the length selectivity of purse seiners (<3%), indicating that these parameters were already well tuned before optimization. Initial and final values of estimated parameters are given in Table 2. The depth of selectivity was considerably increased by the optimization process for purse seiners from 100 m to 124 m, while the selectivity of bait boats was only slightly increased from 20 to 22 m. Comparison between the optimized fleet specific catchability increase due to technological development shows similar values for the 3 purse seine fleets with a slightly higher value for Spanish purse seiners and slightly lower values for bait boats.

3.3. Sensitivity analysis

The sensitivity analysis pointed out seven parameters having a major impact on the cost function (Fig. 3) among which we find four energetic parameters ($\{\dot{p}_{Am}\}$, $[\dot{p}_M]$, $[E_m]$ and κ) used in the DEB formulation and three ecological parameters (T_h, m_{p1} and m_{p2}). Five of them ($\{\dot{p}_{Am}\}$, $[\dot{p}_M]$, κ, T_h and m_{p2}) show a considerable variability of the local sensibility represented by the standard deviation. This indicates that the sensitivity of these parameters varies over the three periods considered for the analysis.

The results emphasize the important sensitivity of T_h which represents the upper boundary of the temperature tolerance in the functional response to temperature (Dueri et al., 2012; Kooijman, 2000). This function determines the changes in

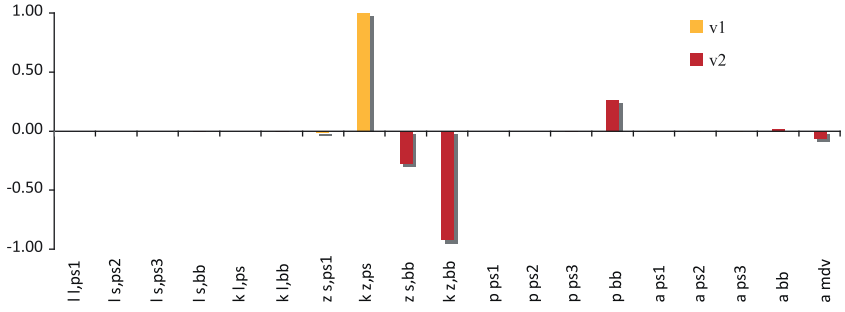

Fig. 1. Values of the elements that compose eigenvectors v_1 and v_2, corresponding to the two smallest eigenvalues of the Hessian.

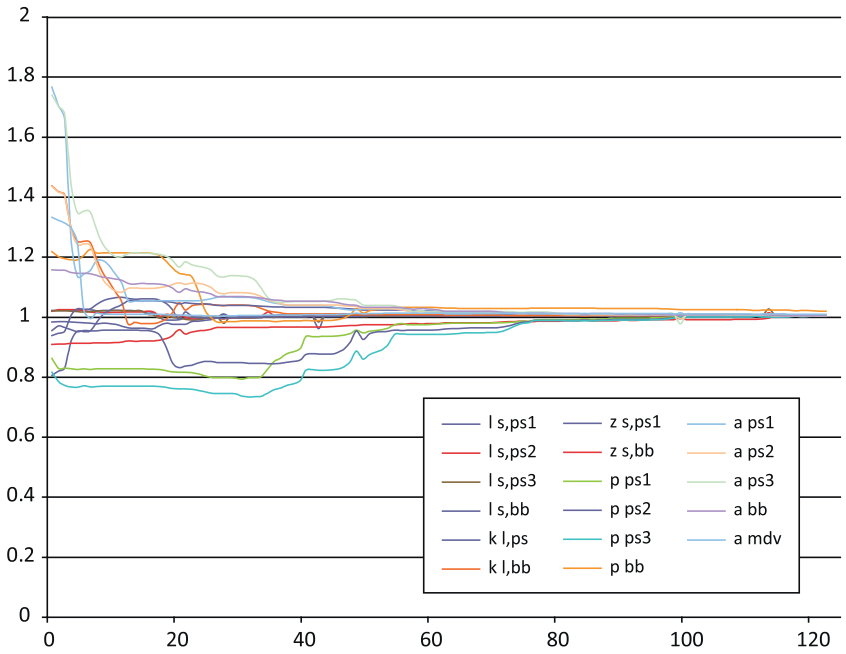

Fig. 2. Variation of the parameters as a function of the iteration number during the minimization process. Values are normalized relatively to the optimized values.

Table 2
Optimized parameters.

Parameter name	Initial value	Final value	Unit	Description
$l_{l,ps1}$	0.500	0.490	m	Length selectivity, PS France
$l_{s,ps2}$	0.500	0.486	m	Length selectivity, PS Spain
$l_{s,ps3}$	0.500	0.490	m	Length selectivity, PS Word
$l_{s,bb}$	0.450	0.482	m	Length selectivity, BB
$k_{l,ps}$	45	45.7		Steepness length selectivity, PS
$k_{l,bb}$	45	31.4		Steepness length selectivity, BB
$z_{s,ps1}$	100	124.8	m	Depth selectivity, PS
$z_{s,bb}$	20	22.1	m	Depth selectivity, BB
p_{ps1}	0.015	0.016		Fishing power PS1
p_{ps2}	0.015	0.016		Fishing power PS2
p_{ps3}	0.025	0.030		Fishing power PS3
p_{bb}	0.005	0.004		Fishing power BB
a_{ps1}	0.200	0.113		Increased efficiency, PS1
a_{ps2}	0.200	0.139		Increased efficiency, PS2
a_{ps3}	0.200	0.115		Increased efficiency, PS3
a_{bb}	0.100	0.086		Increased efficiency, BB
a_{mdv}	0.400	0.299		Maximal attraction factor for Maldives

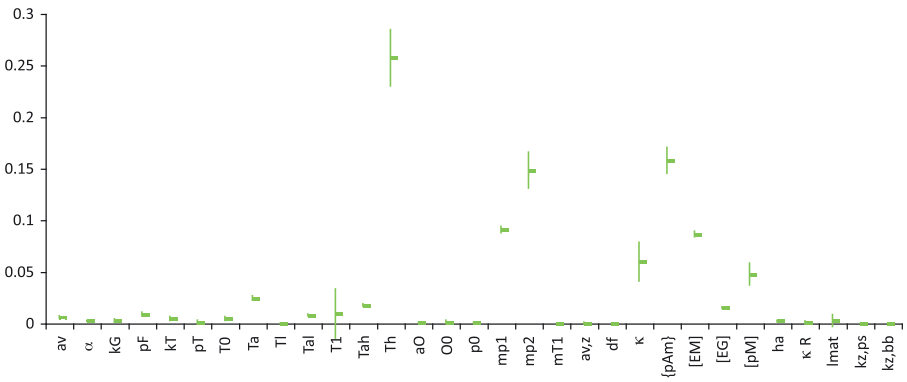

Fig. 3. Mean and standard deviation local sensitivity coefficients.

physiological rate of an organism induced by temperature variations and defines habitat preferences in relation to temperature. In the model this parameter was set according to reported habitat preferences of tropical tuna in the Indian Ocean, which are constrained between 20 and 32 °C (Stéquert and Marsac, 1989). It is not surprising that the parameter has such a major effect on the outcome since it influences temperature-related migrations as well as growth and reproduction.

The sensitivity analysis further highlights the impact of predation mortality parameters m_{p1} and m_{p2}. These parameters determine the survival of the small size fishes (larvae and juveniles) and are therefore very important for population dynamics. At the same time this process is characterised by a high level of uncertainty given the difficulties in collecting observations. In the present model the parameters were tuned. The replacement of the empirical power law function used to represent predation mortality by variable predation mortalities outputted from the APECOSM model (Maury, 2010) could help to constrain these parameters and would greatly benefit the reliability of the model outcome.

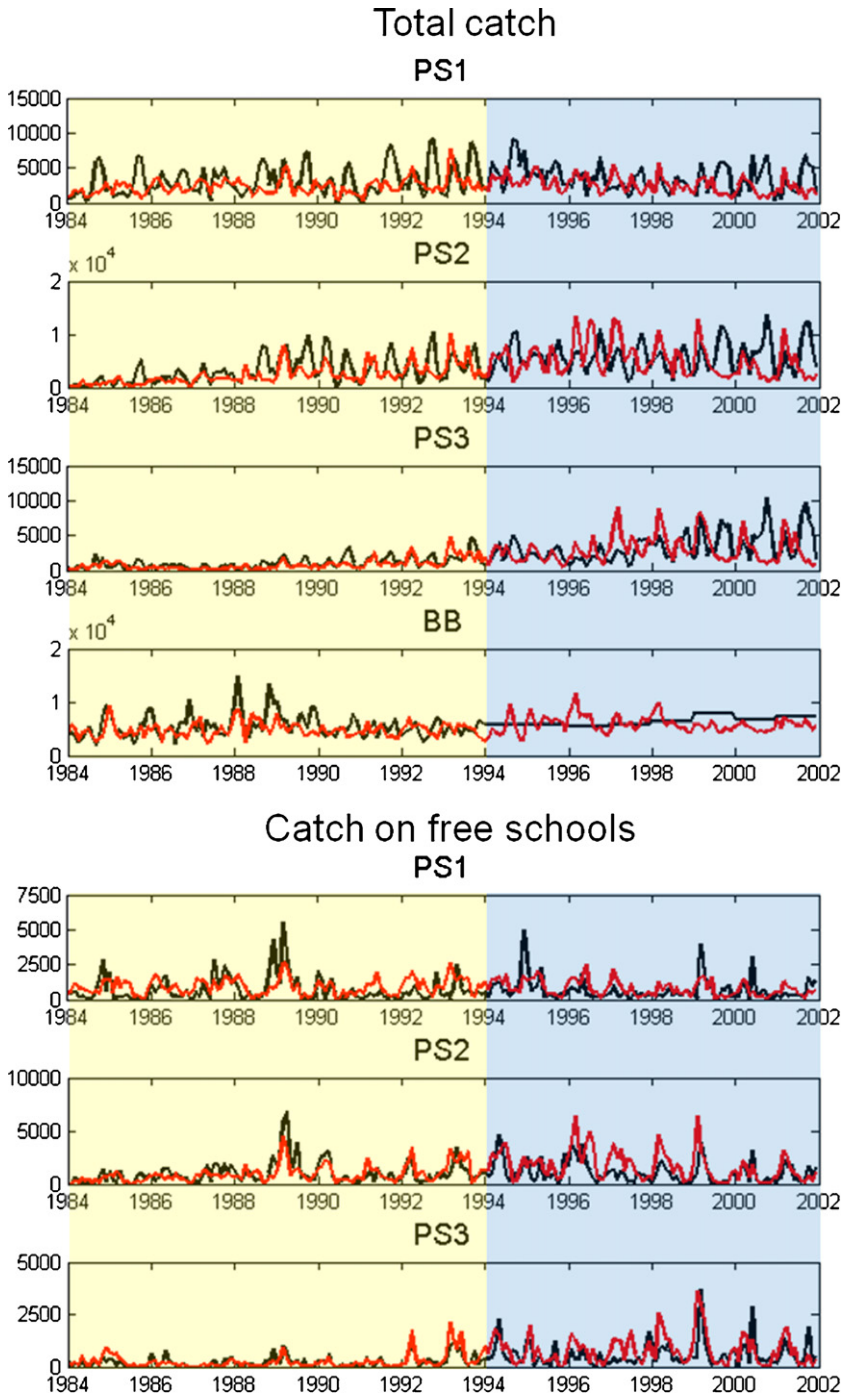

Fig. 4. Comparison between simulated (red) and observed (black) monthly aggregated catches of skipjack tuna for the different fleets: Results for total catches (top) and catches on free schools (bottom). Optimisation is performed on total catches from 1984 to 1993 and the model is running freely from 1994 to 2001. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

The results underline the sensitivity of four DEB parameters describing bio-energetic fluxes in the equations for growth, reproduction and ageing mortality of tuna. These parameters are the surface-area specific assimilation rate $\{\dot{p}_{Am}\}$, the volume-specific maintenance cost $[\dot{p}_M]$ the maximum energy density of reserves $[E_m]$, and the fixed fraction of the energy spent on growth (of structure) and somatic maintenance κ. The parameterisation of $\{\dot{p}_{Am}\}$ and $[\dot{p}_M]$ and κ were derived from Kooijman (2010) while $[E_m]$ was estimated from available growth curves (Dueri et al., 2012). As these parameters have an important effect on the model outcome it is desirable to improve the reliability of their estimation. A possible way to achieve a better confidence in the estimation is the assimilation of the data from the Regional Tuna Tagging Project of the Indian Ocean (RTTP-IO). During this program 78 326 skipjack were tagged and released from May 2005 to August 2007 in the western Indian ocean and so far more than 12 000 fish (>16%) have been recovered and recorded. Thus, this remarkable dataset is a unique source of information concerning the physiology and movements of tunas that could contribute to improve the confidence in the sensitive DEB parameters of the model by integrating them in the parameter estimation.

3.4. Comparison between simulated and observed temporal dynamics of catches and size frequencies (1984–2001)

In order to evaluate the model's ability to represent the temporal dynamics of catches and size frequencies beyond the timeframe of optimization, we compare the simulated and the observed monthly catches over the entire period of simulation 1984–2001.

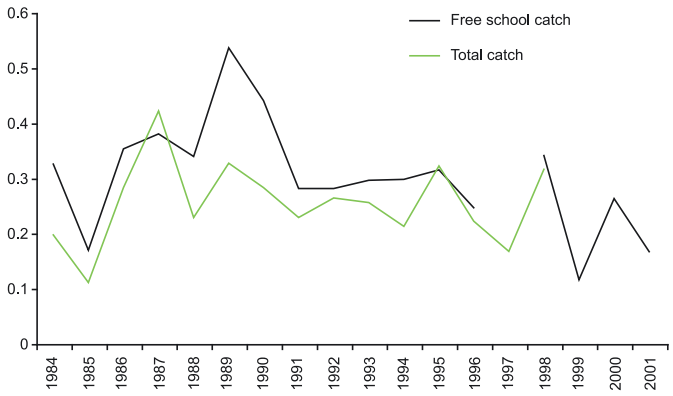

Fig. 5. Spearman's correlation coefficient for observed and computed catches (only years with significant correlations are represented).

The comparison of catches shows that while the simulation successfully represents some of the peaks, it clearly fails to capture most of the autumn peaks observed in the time series of the French and Spanish purse seiners. These missing peaks are related to a specific fishing technique that exploits fish aggregating devices (FAD). Tuna and especially skipjack are known to associate with natural and artificial floating objects. Different hypotheses have been proposed to explain this associative behaviour (Fréon and Dagorn, 2000) and concern has been raised regarding the possible "ecological trap" effect that could be caused by the increased number of FADs deployed that could eventually attract and trap tunas in areas of the ocean with low productivity

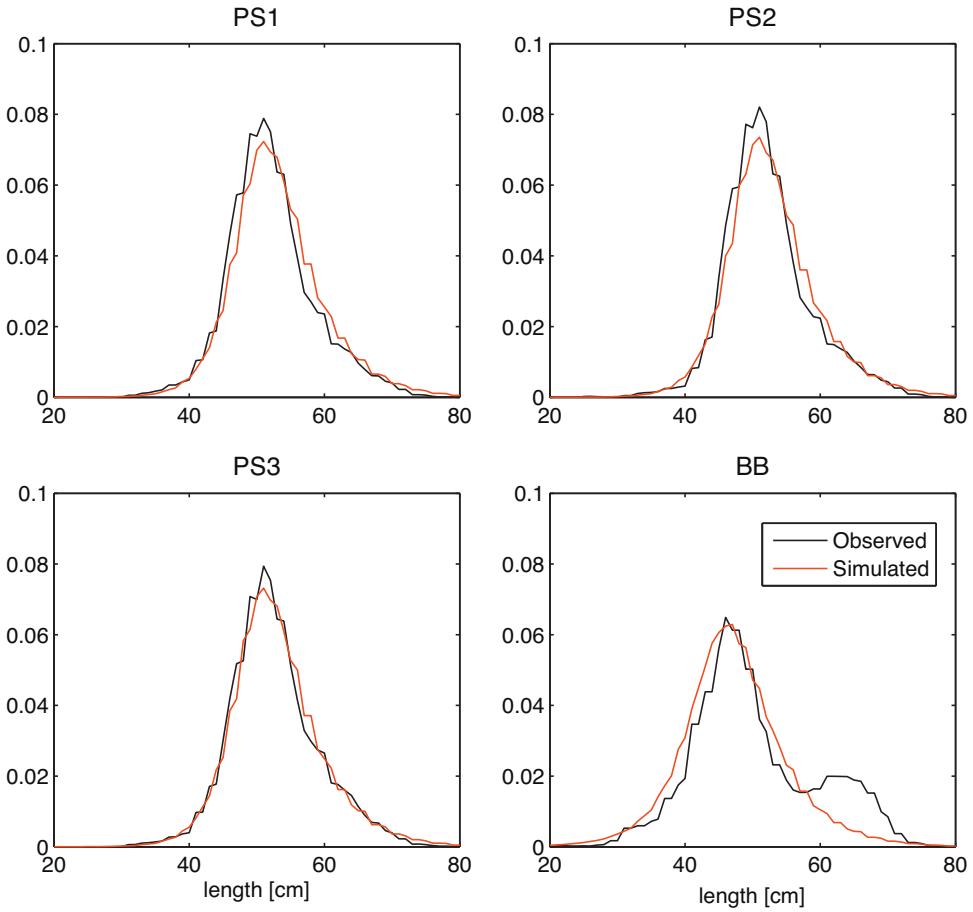

Fig. 6. Comparison between observed and simulated length frequency for the four different fleets (1984–1993).

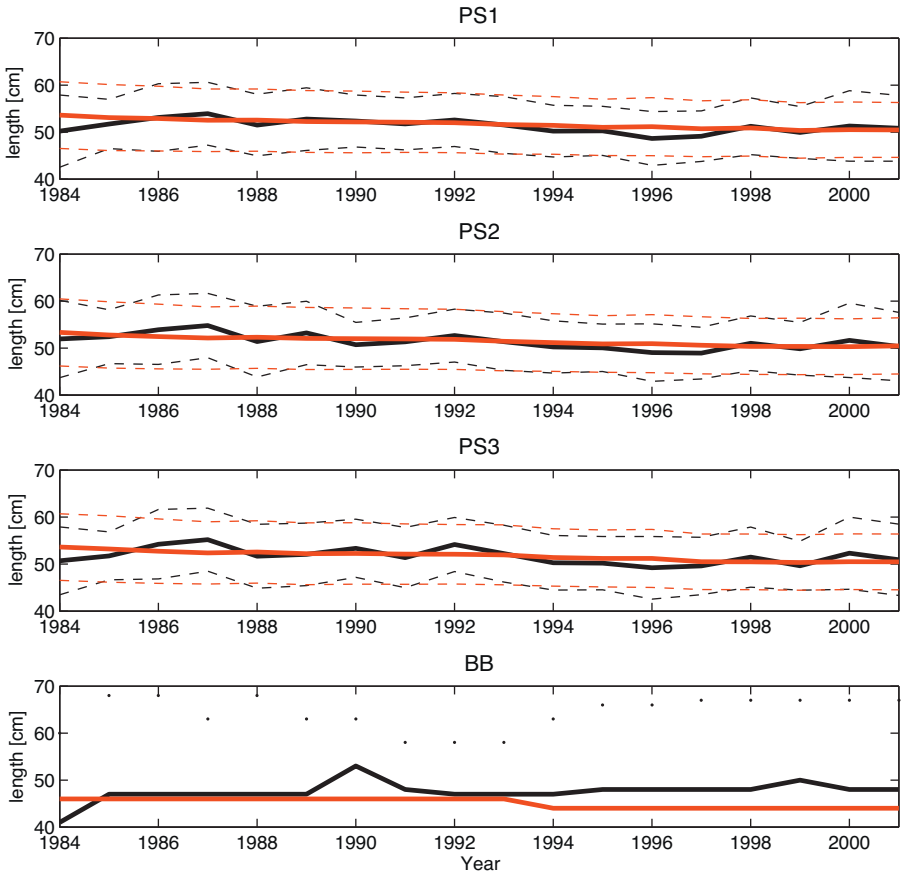

Fig. 7. Observed (blue) and computed (red) temporal evolution of size frequency for the four fleets: mean (solid line) and standard deviation (dotted line) for purse seiners or second mode (dots) for Maldivian baitboats. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

(Marsac et al., 2000; Hallier and Gaertner, 2008). In the Indian Ocean the deployment of FADs by purse seiners have considerably increased since the early 1980s and nowadays 80% of catches occurs in association with FADs (Indian Ocean Tuna Commission, 2008).

The performance of the model improves if we compare catches of free schools (not associated to FADs). This suggests that the model satisfactorily represents the spatio-temporal dynamics of tuna when it is not affected by the presence of FADs (i.e. free schools) and dominantly driven by the modelled habitat conditions, while other factors not included in the model may intervene in the association of tunas under FADs (Fig. 4). Further developments of the model that aim at distinguishing explicitly FAD fishing from free school fishing in both the model formulation and the parameter estimation are therefore to be recommended.

In order to quantify the goodness of fit, we calculate the non-parametric Spearman's rank correlation coefficient of simulated and observed catches (Fig. 5). The Spearman's r coefficient determines how tightly two variables are linked to each other. Here it is used to evaluate the temporal variability of the correlation and to quantify how well the model fits overall catches versus catches on free schools. Results from this analysis show that despite some temporal variability, the correlation is consistently higher for free school catches, with two exceptions in 1987 and 1995. Spearman's r for free schools is significant for 17 of 18 years, while for overall catches, the test is significant for 14 of 18 years.

Observed and computed length frequencies have also been compared for the four fleets (Fig. 6). The model succeeded in representing the length frequency distribution observed in purse seine catches, but it underestimates the presence of large individuals in the catches of the Maldivian bait boats. Observed bait boat length frequency around the Maldivian Islands are well known to exhibit a marked bi-modal distribution with an under-representation of the skipjack having a size between 50 and 60 cm. This pattern was not captured by the model. Adam and Anderson (1996) hypothesize that the missing size-class might migrate offshore, away from the Maldives, for some unknown reasons. Possible causes of this migration could be related to prey abundance issues. Since in the present model formulation the size distribution of preys is imposed by a power law, we are possibly missing a process, which could potentially help to explain the observed phenomena. Further improvement of the model are planned in order to implement a more realistic representation of the preys dynamics.

The temporal dynamics of size frequencies was also compared to the simulation (Fig. 7). For purse seiners, there is a generally good overlap between simulation and observation although the real data show a slight interannual variability that is not captured by the model. The simulation reveals a steady decreasing trend in the mean size frequencies, but this trend is less evident in the observations given the aforementioned interannual variability. For Maldivian baitboats, despite the lack of the bimodal distribution of the size frequency, the representation of the temporal dynamics of the mean value is satisfactory.

4. Conclusion

Optimization and sensitivity analysis were carried out on the APECOSM-E model in order to constrain the identifiable model

parameters using the available data and evaluate the sensitivity of the non-estimated parameters. Despite the generally satisfactory outcome of the model, some limitations and needs for improvement have been pointed out from the comparison of the optimized model with the observed data. To begin with, the present model formulation does not account for the aggregating effect of fish aggregation devices (FAD) and this limits the ability of the model to represent properly FAD catches which constitute an important component of the total catches. As a result, the model cannot fully account for the effect of this fishing technique on the spatial population dynamics. A proper comprehension of the attraction phenomenon is presently missing, even though a growing scientific effort is directed to improve its understanding (Taquet et al., 2007; Gaertner et al., 2008; Soria et al., 2009; Dagorn et al., 2010).

Sensitivity analysis has pointed out that several DEB parameters related to the bioenergetics of skipjack tuna should be improved in order to increase the reliability of the model results. In the present model, the DEB parameters are based on extrapolation from existing information on skipjack tuna physiology, but a better estimation of the relevant parameters based on more specific experimental testing and inclusion of other datasets (e.g. tag-recapture data), is highly recommended in order to enhance the confidence in the model results.

Acknowledgements

This work was supported by the Pelagic Fisheries Research Programme (PFRP) of the University of Hawaii, the AMPED project (www.amped.ird.fr) through a grant from the French National Research Agency (ANR), Systerra Programme, Grant Number ANR-08-STRA-03 and the MACROES project (http://www.macroes.ird.fr/) through a grant from the French National Research Agency (ANR), CEP Programme, Grant Number ANR-09-CEP-003.

The fisheries data analyzed in this publication were obtained from the IOTC (http://www.iotc.org). We wish to acknowledge the contribution of the staff of the 'Observatoire Thonier' of the Mixed Research Unit 212 'Exploited Marine Ecosystems' (IRD) for data processing and management.

References

Adam, M.S., 2010. Declining catches of skipjack in the Indian Ocean – observation from the Maldives. In: Proceedings of the 10th Meeting of the Working Party on Tropical Tuna, Indian Ocean Tuna Commission, IOTC-2010-WPTT-09.

Adam, M.S., Anderson, R.C., 1996. Skipjack tuna (Katsuwonus pelamis) in the Maldives. In: Anganuzzi, A.A., Stobberup, K.A., Webb, N.J. (Eds.), Proceedings of the Sixth Expert Consultation on Indian Ocean Tunas, Colombo, Sri Lanka, September 1995. Indo-Pacific Tuna Programme, Colombo, pp. 232–238.

Anderson, T.R., 2010. Progress in marine ecosystem modelling and the unreasonable effectiveness of mathematics. Journal of Marine Systems 81, 4–11.

Blackford, J., Allen, J.I., Anderson, T.R., Rose, K.A., 2010. Challenges for a new generation of marine ecosystem models: Overview of the Advances in Marine Ecosystem Modelling Research (AMEMR) Symposium, 23–26 June 2008, Plymouth UK Preface. Journal of Marine Systems 81, 1–3.

Brierley, A.S., Kingsford, M.J., 2009. Impacts of climate change on marine organisms and ecosystems. Current Biology 19, R602–R614.

Cariboni, J., Gatelli, D., Liska, R., Saltelli, A., 2007. The role of sensitivity analysis in ecological modelling. Ecological Modelling 203, 167–182.

Dagorn, L., Holland, K.N., Filmalter, J., 2010. Are drifting FADs essential for testing the ecological trap hypothesis? Fisheries Research 106, 6–63.

Dueri, S., Faugeras, B., Maury, O., 2012. Modelling the skipjack tuna dynamics in the Indian Ocean with APECOSM-E: Part 1. Model formulation. Ecological Modelling 245, 41–54.

Faugeras, B., Maury, O., 2005. An advection–diffusion-reaction size-structured fish population dynamics model combined with a statistical parameter estimation procedure: application to the Indian Ocean skipjack tuna fishery. Mathematical Biogeosciences and Engineering 2 (4), 719–741.

Faugeras, B., Maury, O., 2007. Modeling fish population movements: from an individual-based representation to an advection–diffusion equation. Journal of Theoretical Biology 247, 837–848.

Faugeras, B., Lévy, M., Mémery, L., Verron, J., Blum, J., Charpentier, I., 2003. Can biogeochemical fluxes be recovered from nitrate and chlorophyll data? A case study assimilating data in the Northwestern Mediterranean Sea at the JGOFS-DYFAMED station. Journal of Marine Systems 40, 99–125.

Fenner, K., Losch, M., Schröter, J., Wenzel, M., 2001. Testing a marine ecosystem model: sensitivity analysis and parameter optimization. Journal of Marine Systems 28, 45–63.

Fréon, P., Dagorn, L., 2000. Review of fish associative behaviour: toward a generalisation of the meeting point hypothesis. Reviews in Fish Biology and Fisheries 10, 183–207.

Frey, H.C., Patil, S.R., 2002. Identification and review of sensitivity analysis methods. Risk Analysis 23 (3), 553–578.

Gaertner, J.C., Taquet, M., Dagorn, L., Merigot, B., Aumeeruddy, R., Sancho, G., Itano, D., 2008. Visual censuses around drifting fish aggregating devices (FADs): a new approach for assessing the diversity of fish in open-ocean waters. Marine Ecology Progress Series 366, 175–186.

Gilbert, J.C., Lemaréchal, C., 1989. Some numerical experiments with variable-storage quasi-Newton algorithms. Mathematical Programming 45, 407–435.

Hallier, J.P., Gaertner, D., 2008. Drifting fish aggregation devices could act as an ecological trap for tropical tuna species. Marine Ecology Progress Series 353, 255–264.

Hascoët, L., Pascual, V., 2004. TAPENADE 2.1 User's Guide. Technical Report 0300, INRIA.

Indian Ocean Tuna Commission, 2008. Report of the Eleventh Session of the Scientific Committee of the IOTC. Victoria, Seychelles, 1–5 December 2008. IOTC-2008-SC-R[E], 166 pp.

Jackson, J.B.C., 2010. The future of oceans past. Philosophical Transactions of the Royal Society B 365, 3765–3778.

Kooijman, S.A.L.M., 2000. Dynamic Energy and Mass Budgets in Biological Systems, second ed. Cambridge University Press.

Kooijman, S.A.L.M., 2010. Dynamic Energy Theory for Metabolic Organisation, third ed. Cambridge University Press.

Marine Research Section, 1996. The Maldivian tuna fishery: a collection of tuna resource papers. Maldives Marine Research Bulletin 2, 176.

Marsac, F., Fonteneau, A., Ménard, F., 2000. Drifting FADs used in tuna fisheries: and ecological trap? In: Le Gall, J.Y., Cayré, P., Taquet, M. (Eds.), Pêche thonière et dispositifs de concentration de poissons. Actes Colloq. IFREMER, vol. 28, pp. 537–552.

Maury, O., 2010. An overview of APECOSM, a Spatialized Mass Balanced Apex Predators ECOSystem Model to study physiologically structured tuna population dynamics in their ecosystem. In: St John, M., Monfray, P. (Eds.), Parameterisation of Tropic Interactions in Ecosystem Modelling. Progress in Oceanography, vol. 84, pp. 113–117.

Soria, M., Dagorn, L., Potin, G., Fréon, P., 2009. First field-based experiment supporting the meeting point hypothesis for schooling in pelagic fish. Animal Behaviour 78, 1441–1446.

Stéquert, B., Marsac, F., 1989. Tropical tuna – surface fisheries in the Indian ocean. FAO fisheries technical paper no. 282. Rome, Italy, 238 pp.

Taquet, M., Sancho, G., Dagorn, L., Gaertner, J.-C., Itano, D., Aumeeruddy, R., Wendling, B., Peignon, C., 2007. Characterizing fish communities associated with drifting fish aggregating devices (FADs) in the Western Indian Ocean using underwater visual surveys. Aquatic Living Resources 20, 331–341.

Thacker, W.C., 1989. The role of the Hessian matrix in fitting models to measurements. Journal of Geophysical Research 94 (C5), 6177–6196.

Valdemarsen, J.W., 2001. Technological trend in capture fisheries. Ocean & Coastal Management 44, 635–651.

Article K : [18] O. Maury, B. Faugeras, Y-J. Shin, J.C. Poggiale, T. Ben Ari et F. Marsac. Modeling environmental effects on the size-structured energy flow through marine ecosystems. Part 1 : the model. *Progress in Oceanography* 74 (2007), p. 479–499

Available online at www.sciencedirect.com
ScienceDirect
Progress in Oceanography 74 (2007) 479–499
Progress in Oceanography
www.elsevier.com/locate/pocean

Modeling environmental effects on the size-structured energy flow through marine ecosystems. Part 1: The model

Olivier Maury [a,*], Blaise Faugeras [a], Yunne-Jai Shin [a], Jean-Christophe Poggiale [b], Tamara Ben Ari [a], Francis Marsac [a]

[a] *Institut de Recherche pour le Développement (IRD) – UR 109 Thetis, CRH, av. Jean Monnet, B.P. 171, 34203 Sète Cedex, France*
[b] *LMGEM – UMR 6117, OSU – Case 901, Campus de Luminy, 13288 Marseille Cedex 9, France*

Available online 13 May 2007

Abstract

This paper presents an original size-structured mathematical model of the energy flow through marine ecosystems, based on established ecological and physiological processes and mass conservation principles. The model is based on a nonlocal partial differential equation which represents the transfer of energy in both time and body weight (size) in marine ecosystems. The processes taken into account include size-based opportunistic trophic interactions, competition for food, allocation of energy between growth and reproduction, somatic and maturity maintenance, predatory and starvation mortality. All the physiological rates are temperature-dependent. The physiological bases of the model are derived from the dynamic energy budget theory. The model outputs the dynamic size-spectrum of marine ecosystems in term of energy content per weight class as well as many other size-dependent diagnostic variables such as growth rate, egg production or predation mortality.

In stable environmental conditions and using a reference set of parameters derived from empirical studies, the model converges toward a stationary linear log–log size-spectrum with a slope equal to −1.06, which is consistent with the values reported in empirical studies. In some cases, the distribution of the largest sizes departs from the stationary linear solution and is slightly curved downward. A sensitivity analysis to the parameters is conducted systematically. It shows that the stationary size-spectrum is not very sensitive to the parameters of the model. Numerical simulations of the effects of temperature and primary production variability on marine ecosystems size-spectra are provided in a companion paper [Maury, O., Shin, Y.-J., Faugeras, B., Ben Ari, T., Marsac, F., 2007. Modeling environmental effects on the size-structured energy flow through marine ecosystems. Part 2: simulations. Progress in Oceanography, doi:10.1016/j.pocean.2007.05.001].

Keywords: Size spectrum; Mathematical model; Predation; Bioenergetics; Dynamic energy budget (DEB) theory; Energy flow

1. Introduction

Trophic interactions between organisms are the main drivers of marine ecosystems dynamics. In particular, they allow the transfer and the dissipation of solar energy through ecosystems, along food chains, from primary

* Corresponding author.
E-mail address: maury@ird.fr (O. Maury).

0079-6611/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pocean.2007.05.002

producers to top predators. In marine systems, many species interact within complex trophic networks where bottom-up as well as top-down controls interfere continuously (e.g., Cury et al., 2003). Understanding how environmental variability such as changes in primary production or temperature impacts ecosystems and ultimately fish stocks and reciprocally how fishing upper trophic levels impacts lower trophic levels requires reliable models based on realistic representations of energy fluxes through ecosystems. However, most marine ecosystems are extremely diverse, heterogeneous and poorly known. Modelling their dynamics explicitly down to the species level is challenging. Hence, most models of marine ecosystems rely on rough species and functional groups partitioning and use fixed predation rates between groups (e.g., Polovina, 1984; Walters et al., 1997; Pauly et al., 2000). Alternatively, aggregated approaches based on size have been undertaken, taking into account allometric losses (respiration), predation and growth processes. In those approaches, phytoplankton is implicitly used as the source term of size-structured continuous mass-balance equations. The marine ecosystem is represented using a single aggregated state variable (e.g., a biomass) which experiences size-dependent growth and mortality (Platt and Denman, 1978; Silvert and Platt, 1978, 1980; Dickie et al., 1987; Cushing, 1992; Platt and Denman, 1997; Arino et al., 2004; Benoit and Rochet, 2004). Those models rest on the fundamental assumption that size is the most structuring dimension of ecological systems along which their dynamics can be projected. Many ecological traits (including population abundance, growth rate and productivity, spatial niche, trophic, competitive and facilitative relationships between species) as well as metabolic processes are indeed well correlated with body size (Sheldon et al., 1972; Blueweiss et al., 1978; Gillooly et al., 2001; Brown and Gillooly, 2003; Marquet et al., 2005; West and Brown, 2005; Woodward et al., 2005). Furthermore, because most marine organisms are highly opportunistic feeders and because prey size is limited by the allometric diameter of predator's mouth (Bone et al., 1999), predator–prey relationships are, in many marine systems, mostly determined by size (Lundvall et al., 1999; Scharf et al., 2000; Jennings et al., 2001 and Jennings et al., 2002; Shin and Cury, 2004). For instance, Jennings et al. (2001) showed that body mass explained 93% of the variation in trophic level among 15 fish communities in the North Sea. Because it captures so many aspects of ecosystem functioning, body size can therefore be used to synthesize a suite of co-varying traits into a single dimension (Cousins, 1980; Woodward et al., 2005).

As Woodward et al. (2005) state, "the challenge now is for empiricists to produce highly resolved food webs that are quantified in terms of population dynamics, energetics and chemical fluxes, and for theoreticians to develop new and more realistic size-based models, so that emerging ideas can be explored and tested more rigorously". Furthermore, "size-based models are easier and cheaper to parameterise than most food-web models" (Jennings et al., 2002). In this perspective, we model environmental influences on the dynamics of marine ecosystems with a size-spectrum approach. Primary producers are explicitly distinguished from consumer organisms and a mechanistic approach allows us to take into account various ecological and physiological processes supposed to be determining in the functioning of marine ecosystems:

- Size-structured opportunistic trophic interactions where producers are potential preys for consumers and where all consumer species are considered to be potentially prey and predator at the same time (Shin and Cury, 2004);
- Predators competition for preys;
- Allocation of energy between growth and reproduction;
- Somatic as well as maturity maintenance based on the dynamic energy budget (DEB) theory (Kooijman, 1986, 2000, 2001; Nisbet et al., 2000);
- Size-dependent nonpredatory mortality;
- Starvation mortality;
- Temperature-dependence of organism's physiological rates.

It is expected that considering explicitly the physiological bases of metabolism, the main constraints which control trophic interactions and the size-structured nature of those processes will help to better understand the various modes of energy transfer through marine ecosystems and their response to environmental forcing. Furthermore, a mass-balanced formulation is used to represent the functioning of marine ecosystems in a quantitative way, assessing the actual energy flux from primary production to apex predators as well as the top-down effects that upper trophic levels have on the overall ecosystem. To keep consistency with bioener-

getic studies and to avoid the complexity of explicit stoichiometric formulations based on chemical elements, our model is expressed in term of energy. Energy has to be understood as a currency measuring "the ability to do work" (Kooijman, 2000). It has to be noted that given homeostasis assumptions, all mass fluxes in organisms can be deduced from energy fluxes (Kooijman, 1995; Sterner and Elser, 2002). In our approach, energy is simply assumed to be proportional to biomass. This implies an assumption of strict homeostasis and constant chemical stoichiometry between organisms.

After a detailed presentation of the hypothesis and formulations of our model, a sensitivity analysis is undertaken to asses the impact of each parameter on the steady state size-spectrum. In a companion paper (Maury et al., 2007), we present numerical simulations of our model focusing on the effects of primary production and temperature variability on the size-spectrum of marine ecosystem.

2. The model

2.1. Notations and state variables

The main state variable we are dealing with is $\xi_{t,w}$, the distribution function of the energy content of the marine ecosystem (J kg^{-1} m^{-3}) at time $t \in [0, +\infty[$ and weight $w \in [0, w_{\max}]$ in 1 m^3 of seawater. $\xi_{t,w}$ is a density with respect to body weight and seawater volume. It can easily be converted into the more usual "normalized biomass size-spectrum" using the mean energetic content of one unit of biomass ψ (J kg^{-1}) which is assumed to be a constant parameter. Hence, the quantity of energy in the weight range $[w_1, w_2]$ per m^3 of seawater is given by $\int_{x=w_1}^{x=w_2} \xi_{t,x}\,dx$ and $\xi_{t,w}$ is related to $N_{t,w}$, the distribution function of the number of individuals in terms of weight (kg^{-1} m^{-3}) at (t,w) in 1 m^3 of seawater, with $\xi_{t,w} = \psi \cdot w \cdot N_{t,w}$.

The symbols u, v, w, x are continuous indices which refer all to the weight dimension. Weight is supposed to be related to length with a fixed allometric function $w = al^3$.

According to basic ecological theory, marine ecosystems can be schematically divided into three distinct components using fundamentally different means to mobilize energy: producers, consumers and decomposers (Valiela, 1995). For the sake of simplicity, the present study ignores the third component and focuses on the two first components with a particular emphasis on the consumers group (Fig. 1). Hence, our model has two main components:

- the primary producers (autotrophic organisms mostly composed of phytoplankton) which convert solar energy and mineral nutrients into biomass and whose weight belongs to $[0, w_1]$;
- the consumers (heterotrophic organisms encompassing numerous taxonomic groups of zooplankton and nekton) which gain energy solely by predation and whose weight belongs to $[w_{\text{egg}}, w_{\max}]$. Consumers do reproduce, their eggs have a weight $w_{\text{egg}} > 0$ and their maximal weight is $w_{\max} > w_1$.

The distribution function of the energy content of the producer and consumer groups are noted respectively $\xi^{\text{p}}_{t,w}$ and $\xi^{\text{c}}_{t,w}$ so that the distribution function of the energy content of the ecosystem is $\xi_{t,w} = \chi_{[0,w_1]}\xi^{\text{p}}_{t,w} + \chi_{[w_{\text{egg}},w_{\max}]}\xi^{\text{c}}_{t,w}$ with $\chi_{[x_1,x_2]}$ being the characteristic function which is equal to one in the interval $[x_1, x_2]$ and to zero elsewhere.

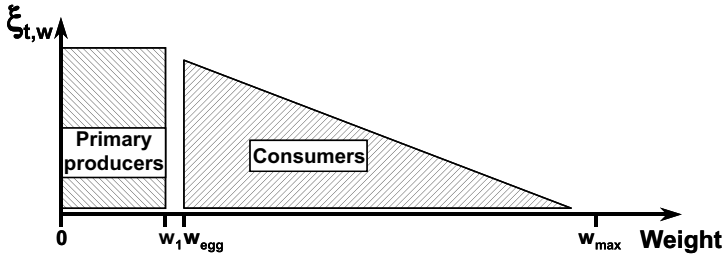

Fig. 1. Schematic representation of the weight structured ecosystem distinguishing primary phytoplanktonic producers from predatory consumers (log–log).

2.2. Dynamics

To avoid an explicit modeling of phytoplankton growth and reproduction, the energy density of producer organisms is assumed to be uniformly distributed over $[0, w_1]$. Consequently, the size-dependent predatory mortality applied by all consumer organisms (see Eq. (8)) is averaged over the producers size range $[0, w_1]$ to ensure that the producers size distribution remains constant at all time. The dynamics of phytoplanktonic organisms is then expressed as follows:

$$\frac{\mathrm{d}\xi^{\mathrm{p}}_{t,w}}{\mathrm{d}t} = \frac{1}{w_1}\left(\Pi_t - \xi^{\mathrm{p}}_{t,w}\int_{x=0}^{x=w_1}\lambda_{t,x}\,\mathrm{d}x\right) - \xi^{\mathrm{p}}_{t,w}M_I \quad \forall w \in [0; w_1] \tag{1}$$

With Π_t (J s^{-1} m^{-3}) the primary energy production which enters the system at time t, which constitutes the only external source of energy of the whole ecosystem, M_I (s^{-1}) the nonpredatory mortality rate affecting primary producers and $\lambda_{t,x}$ (s^{-1}) the mortality rate due to predation at time t and weight x.

The bio-ecological processes taken into account to model consumers are predation, mortality, assimilation and use of energy for maintenance, growth and reproduction. The basic equation used to describe the energy fluxes through the weight range of consumers combines a transport term for representing the growth process and three sink terms for predatory, nonpredatory and starvation mortality processes. It is based on the Mc Kendrick–Von Foerster equation which is usually used in population dynamics (e.g., Tuljapurkar and Caswell, 1997; Kot, 2001) and which is written as follows in the interval $]w_{\mathrm{egg}}, w_{\max}]$ assuming given initial conditions for $t = 0$:

$$\begin{cases} \dfrac{\partial\xi^{\mathrm{c}}_{t,w}}{\partial t} = -\dfrac{\partial(\gamma_{t,w}\xi^{\mathrm{c}}_{t,w})}{\partial w} - (\lambda_{t,w} + Z_w + M^{\mathrm{starv}}_{t,w})\xi^{\mathrm{c}}_{t,w} \quad \forall w \in]w_{\mathrm{egg}}; w_{\max}] \\ \xi^{\mathrm{c}}_{0,w} = \xi^{0}_{w} \end{cases} \tag{2}$$

where γ (kg s^{-1}) is the growth rate, λ (s^{-1}) is the mortality rate due to predation, Z (s^{-1}) is the loss of energy from the system due to nonpredatory mortality and M^{starv} (s^{-1}) is the starvation mortality rate. For all those coefficients, the subscripts t and w refer to time and weight.

The input of eggs R_t (J s^{-1} m^{-3}) into the system due to reproduction is taken into account assuming a Dirichlet boundary condition in $w = w_{\mathrm{egg}}$:

$$\gamma_{t,w_{\mathrm{egg}}}\xi^{\mathrm{c}}_{t,w_{\mathrm{egg}}} = R_t \tag{3}$$

The derivation of explicit expressions for all the coefficients of Eqs. (2) and (3) ($\lambda_{t,w}$, $\gamma_{t,w}$, R_t, $M^{\mathrm{starv}}_{t,w}$ and $Z_{t,w}$) are provided in the five subsections below.

2.2.1. The predation process: calculation of $\lambda_{t,w}$

Predation can be viewed as a loss of energy for preyed weight classes and a gain of energy for predating weight classes. In the model, predation is supposed to be opportunistic and only controlled by the ratio of sizes between organisms. Hence, all organisms can be potentially predators and preys at the same time, depending on their relative weight (or size) (Fig. 2).

To be able to calculate the quantity of food available to a predator, the size-based constraints on predation have to be specified. For that purpose, the selectivity $s_{u,w} \in [0, 1]$ is defined as the capability for a consumer organism of weight u to eat an encountered organism of weight w. Assuming that predation can occur if the ratio of predator length over prey length is comprised between two ρ_1 and ρ_2 extreme values (Fig. 3b), $s_{u,w}$ is a normalized function expressed as the product of two sigmoid functions which account for the limitation of ingestion when preys are either too small or too large (Fig. 3a):

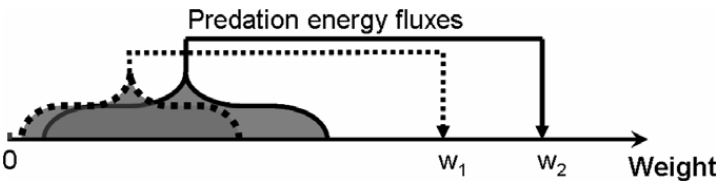

Fig. 2. Schematic representation of weight (size) structured energy flow through the ecosystem.

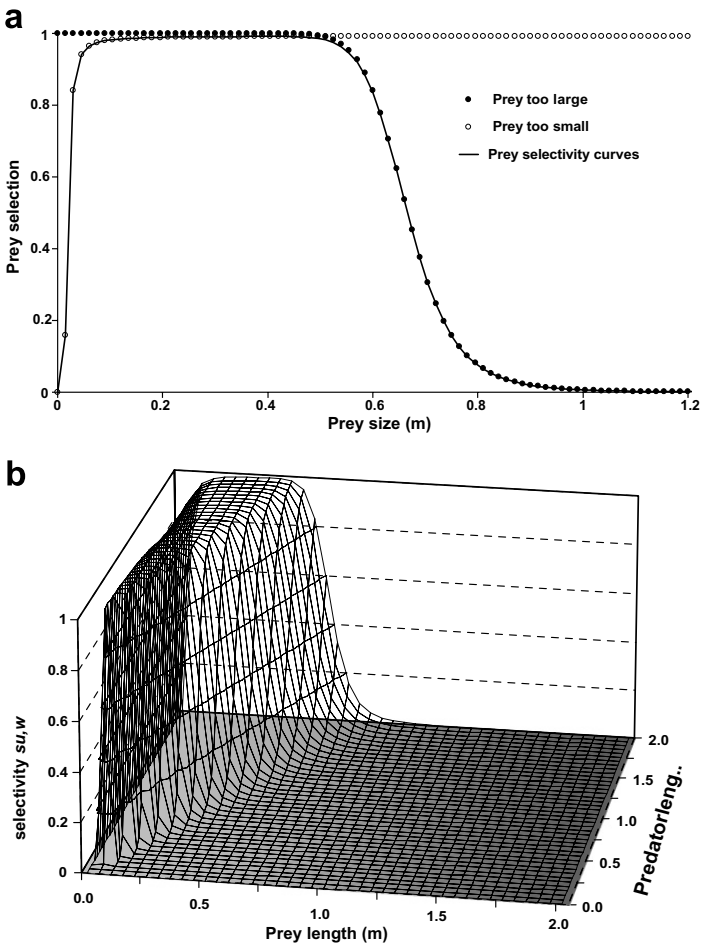

Fig. 3. (a) Limitation curves for preys too large to be ingested (black dots), preys too small to be ingested (open circles) and resulting prey selectivity function $s_{2,w}$ as a function of prey length $\left(\frac{w}{a}\right)^{1/3}$ for a 2 m long predator ($\rho_1 = 3$, $\rho_2 = 100$, $\alpha_1 = 5$ and $\alpha_2 = 0.05$). (b) Selectivity function $s_{u,w}$ versus prey length $\left(\frac{w}{a}\right)^{1/3}$ and predator length $\left(\frac{u}{a}\right)^{1/3}$ with $\rho_1 = 3$, $\rho_2 = 100$, $\alpha_1 = 5$ and $\alpha_2 = 0.05$.

$$s_{u,w} = \left(1 + e^{\alpha_1\left(\rho_1 - \left(\frac{u}{w}\right)^{1/3}\right)}\right)^{-1}\left(1 - \left(1 + e^{\alpha_2\left(\rho_2 - \left(\frac{u}{w}\right)^{1/3}\right)}\right)^{-1}\right) \quad R^{+*2} \xrightarrow{s}]0;1[\tag{4}$$

With ρ_1, ρ_2, α_1 and α_2, being constant positive parameters characterizing both the half saturation and the flatness of the sigmoid functions.

To take into account the basic physiological processes involved in the acquisition and use of energy by biological organisms, a simplified version of the dynamic energy budget (DEB) theory is used (Kooijman, 1986, 2000, 2001; Nisbet et al., 2000). In the DEB theory, the ingested energy is assimilated by organisms and stocked into reserves. A fixed fraction κ of the energy utilized from reserves is then allocated to growth of structural material and somatic maintenance, the remaining fraction $1 - \kappa$ being devoted to gonad development, maturity maintenance and egg formation. For the purpose of simplicity, neither the reserve dynamic nor the gonad development is considered explicitly in the present work. The ingested energy is supposed to be used in the same way by any organism: it is assimilated, and a fraction κ is used for somatic growth and maintenance whereas a fraction $1 - \kappa$ is allocated to reproduction and gonadic maintenance (Fig. 4). A single set of mean physiological parameters (Table 1) is used to describe the mean energy fluxes through every consumer organisms of the ecosystem: the ecosystem is modeled as a "meta-organism" characterized by a mean life history.

According to the DEB theory, the maximum amount of preyed energy that can be ingested at time t during dt by a predator is supposed to be proportional to a body surface. It follows that $E_{t,u}\,du\,dt$ (J m^{-3}), the total

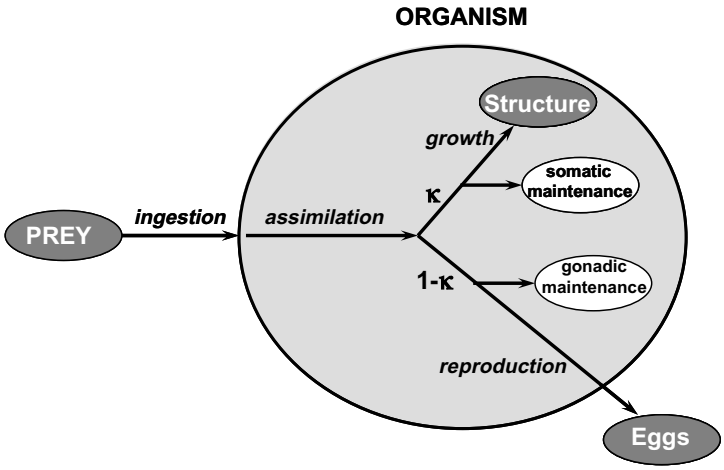

Fig. 4. Schematic representation of energy flow through organisms (simplified from Kooijman, 2000).

Table 1
Parameters used for numerical simulations (ranges are given when several studies are available)

Parameter	Designation and unit	Value	Source
A	Shape coefficient $w = al^3$ (kg m^{-3})	15	Data from Froese and Pauly (2000)
ϕ	Sex ratio (no dimension)	0.5	Arbitrary
M	Nonpredatory mortality for $l = 1$ m (s^{-1})	1.524×10^{-8}	See Appendix C
υ	Allometric coefficient of M (no dimension)	-0.2995	See Appendix C
M_{egg}	Fraction of the spawned eggs which are not fecunded	0.4	Arbitrary
ψ	Energetic content of one unit of biomass (J kg^{-1})	4×10^6	Daan (1975), Edwards et al. (1972), Krohn et al. (1996) and Kitchell et al. (1978)
ω	Maximum surface-specific ingestion rate (kg kg$^{-2/3}$ s^{-1})	5.459×10^{-7}	See Appendix B
κ	Fraction of the assimilated energy allocated to growth and somatic maintenance (no dimension)	[0.65, 0.88] 0.65	Estimations from Brill et al. (1978) and van der Veer et al. (2003)
e_A	Fraction of the ingested energy which is assimilated (no dimension)	[0.65, 0.99] 0.8	Data and estimates from Essington et al. (2001), Andersen and Riis-Vestergaard (2003), Krohn et al. (1996), Kitchell et al. (1978) and Brett and Groves (1979)
E_g	Weight specific cost of growth (Kooijman, 2000) (J kg^{-1})	7×10^6	van der Veer et al. (2003)
μ	Amount of energy required for the somatic maintenance of one unit of weight during one unit of time (J kg^{-1} s^{-1})	0.20949	See Appendix B
ρ_1	Minimum ratio of predator size over prey size	3	Floeter and Temming, 2003; Juanes, 2003 and Ménard et al. (2006);
ρ_2	Maximum ratio of predator size over prey size	100	Floeter and Temming, 2003; Juanes, 2003 and Ménard et al., 2006
α_1	Shape parameter for the selectivity curve	5	(See text)
α_2	Shape parameter for the selectivity curve	0.05	(See text)
C	Holling type II half-saturation constant (J m^3 s^{-1})	117.7	Tuned
χ	w^χ is the volume of water explored by a predator of weight w (m^3 s^{-1})	0.33	Fixed so that w^χ is proportional to length
τ_A	Arrhenius temperature-dependent correction factor (K)	$[2 \times 10^3$, $16 \times 10^3]$ 8×10^3	Brett and Groves (1979) and van der Veer et al. (2003)

The values are derived from the literature or from estimations detailed in Appendix.

amount of energy potentially preyed by all predators of weight comprised in the range $[u, u+\mathrm{d}u]$ at time t during $\mathrm{d}t$ in 1 m^3 of water, can be expressed as follows:

$$\begin{aligned} E_{t,u}\,\mathrm{d}u\,\mathrm{d}t &\propto [\text{density of predators}]_{t,u}\,\mathrm{d}u \cdot \{\text{body surface}\}_u \cdot \{\text{functional response to preys}\}_{t,u} \cdot \mathrm{d}t \\ E_{t,u}\,\mathrm{d}u\,\mathrm{d}t &= \psi\omega\frac{\xi^{\mathrm{c}}_{t,u}\,\mathrm{d}u}{u\psi}u^{2/3}f_u(p_{t,u})\mathrm{d}t \\ &= \omega\xi^{\mathrm{c}}_{t,u}u^{-1/3}f_u(p_{t,u})\mathrm{d}u\,\mathrm{d}t \\ f_u(p_{t,u}) &= \frac{p_{t,u}}{\frac{c}{u^{\chi}}+p_{t,u}}, \quad R^+ \xrightarrow{f} [0;1[\end{aligned} \tag{5}$$

where ω is the mean maximum surface-specific ingestion rate (kg $\mathrm{kg}^{-2/3}$ s^{-1}) and f_u is the functional response to the energy content of preys $p_{t,u}$ $\left(p_{t,u}=\int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v\right)$ of a weight u predator. A size-dependent Holling type II functional response without predator interference is assumed with c the half-saturation constant (J s^{-1}). u^{χ} is the volume of water explored by a predator of weight w per unit of time (m^3 s^{-1}) which is supposed to be an allometric function of predator weight (it is assumed that the volume of water explored by a predator is proportional to its swimming speed which is proportional to its body size – Froese and Pauly, 2000 – hence χ is taken equal to 0.33 cf. Table 1).

Then, according to the hypothesis of opportunistic predation (preys of a given weight are eaten in proportion to their selected biomass relatively to the biomass of all possible preys), the amount of preyed energy $E_{t,u/w}\,\mathrm{d}u\,\mathrm{d}w\,\mathrm{d}t$ (J m^{-3}) that predators in the range $[u, u+\mathrm{d}u]$ take from preys in the range $[w,\ w+\mathrm{d}w]$ at time t during $\mathrm{d}t$ is expressed as follows:

$$\begin{aligned} E_{t,u/w}\,\mathrm{d}u\,\mathrm{d}w\,\mathrm{d}t &= E_{t,u}\,\mathrm{d}u\,\mathrm{d}t\frac{s_{u,w}\xi_{t,w}\,\mathrm{d}w}{\int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v} = \omega\xi^{\mathrm{c}}_{t,u}u^{-1/3}f_u\left(\int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v\right)\frac{s_{u,w}\xi_{t,w}\,\mathrm{d}w}{\int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v}\mathrm{d}u\,\mathrm{d}t \\ &= \omega\xi^{\mathrm{c}}_{t,u}u^{-1/3}\frac{s_{u,w}\xi_{t,w}}{\frac{c}{u^{\chi}}+\int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v}\,\mathrm{d}u\,\mathrm{d}w\,\mathrm{d}t \end{aligned} \tag{6}$$

The total amount of energy preyed by all predators on preys in the range of weight $[w, w+\mathrm{d}w]$ at time t during $\mathrm{d}t$ in 1 m^3 of water is then calculated by integration over the weight range of predators:

$$E_{t,/w}\,\mathrm{d}w\,\mathrm{d}t = \int_{u=w_{\mathrm{egg}}}^{w_{\max}} E_{t,u/w}\,\mathrm{d}u\,\mathrm{d}w\,\mathrm{d}t = \omega\xi_{t,w}\int_{u=w_{\mathrm{egg}}}^{u=w_{\max}}\left[\frac{\xi^{\mathrm{c}}_{t,u}u^{-1/3}s_{u,w}}{\frac{c}{u^{\chi}}+\int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v}\right]\mathrm{d}u\,\mathrm{d}w\,\mathrm{d}t \tag{7}$$

It follows that the instantaneous mortality rate exerted by all possible predators on $\xi_{t,w}$ at time t is given by the following expression:

$$\lambda_{t,w} = \frac{E_{t,/w}}{\xi_{t,w}} = \omega\int_{u=w_{\mathrm{egg}}}^{w_{\max}}\left[\frac{\xi^{\mathrm{c}}_{t,u}u^{-1/3}s_{u,w}}{\frac{c}{u^{\chi}}+\int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v}\right]\mathrm{d}u \tag{8}$$

2.2.2. The growth process: calculation of $\gamma_{t,w}$

According to Fig. 4, growth corresponds to the use of a fraction κ of the assimilated energy diminished by a maintenance cost proportional to organism body volume and finally converted into structural material at an energy cost proportional to growth (Kooijman, 2000). Following those simple rules for energy conservation, the growth of a mean consumer organism is expressed as follows:

$$\frac{\mathrm{d}w_{t,u}}{\mathrm{d}t} = \frac{\kappa e_{\mathrm{A}}E_{t,u}}{\psi N_{t,u}} - \frac{\mu u}{\psi} - \frac{E_{\mathrm{g}}}{\psi}\frac{\mathrm{d}w_{t,u}}{\mathrm{d}t} \tag{9}$$

where $e_{\mathrm{A}} \in [0,1]$ is the mean fraction of the ingested energy which is assimilated, $\kappa \in [0,1]$ is the mean fraction of this energy which is allocated to growth and somatic maintenance, $(1-\kappa)$ being allocated to reproduction, μ is the mean amount of energy required for the somatic maintenance of one unit of weight during one unit of time (J kg^{-1} s^{-1}) and E_{g} is the mean weight specific cost of growth (Kooijman, 2000) (J kg^{-1}).

We assume that growth in length cannot be negative for most marine organisms which have an exo- or an endo-skeleton such as vertebrates, most molluscs, crustaceans, etc. Because weight is assumed to be related to

length with a fixed allometric function ($w = al^3$, $a>0$), growth in weight cannot be negative either (see the paragraph on starvation mortality for the treatment of mass conservation). It follows that the instantaneous growth rate of organisms of weight u (kg s^{-1}) can be expressed as:

$$\begin{aligned}\gamma_{t,u} &= \frac{\mathrm{d}w_{t,u}}{\mathrm{d}t} = \frac{\psi}{\psi+E_g}\left[\frac{\kappa e_A E_{t,u}}{\psi N^c_{t,u}} - \frac{\mu}{\psi}u\right]^+ = \frac{\psi}{\psi+E_g}\left[\frac{\kappa e_A E_{t,u}}{\xi^c_{t,u}}u - \frac{\mu}{\psi}u\right]^+ \\ &= \frac{\psi}{\psi+E_g}\left[\kappa e_A \omega f_u(p_{t,u})u^{2/3} - \frac{\mu}{\psi}u\right]^+ = \frac{\psi}{\psi+E_g}\left[\frac{\kappa e_A \omega \int_{v=0}^{w_{max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v}{\frac{c}{u^\chi}+\int_{v=0}^{w_{max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v}u^{2/3} - \frac{\mu}{\psi}u\right]^+\end{aligned} \tag{10}$$

Where $[x]^+$ is the function defined by

$$\begin{cases}[x]^+ = x & \text{if} \quad x \geqslant 0 \\ [x]^+ = 0 & \text{if} \quad x < 0\end{cases}.$$

At food saturation (when the functional response $f = 1$), this growth rate formulation is equivalent to a von Bertalanffy (1969) formulation of growth where anabolism is proportional to a surface (weight at a power 2/3) and catabolism is proportional to body weight.

2.2.3. The reproduction process: calculation of R_t

According to Fig. 4, reproduction corresponds to the use of a fraction $1-\kappa$ of the assimilated energy diminished by a maintenance cost proportional to $(1-\kappa)/\kappa$ times body weight (Kooijman, 2000). All sizes of both sex are supposed to reproduce permanently but only female sexual products are re-injected into the system at $w = w_{egg}$ (according to Cury and Pauly, 2000, egg size of marine fish is remarkably constant between species and approximately equals to 1 mm).

As for the expression of the growth rate and because the contribution of the weight class w to the total eggs production cannot be negative, the function $[]^+$ is used to express the egg input into the system (see the paragraph on starvation mortality for the treatment of mass conservation):

$$\begin{aligned}R_t &= (1-M_{egg})\phi\int_{w=w_{egg}}^{w_{max}}\left[e_A(1-\kappa)E_{t,w} - N^c_{t,w}\frac{1-\kappa}{\kappa}\mu w\right]^+\mathrm{d}w \\ &= (1-M_{egg})\phi\int_{w=w_{egg}}^{w_{max}}\left[\frac{(1-\kappa)e_A\omega\xi^c_{t,w}w^{-1/3}\int_{v=0}^{w_{max}} s_{w,v}\xi_{t,v}\mathrm{d}v}{\frac{c}{w^\chi}+\int_{v=0}^{w_{max}} s_{w,v}\xi_{t,v}\mathrm{d}v} - \frac{(1-\kappa)}{\kappa}\frac{\mu\xi^c_{t,w}}{\psi}\right]^+\mathrm{d}w\end{aligned} \tag{11}$$

With R (J s^{-1} m^{-3}) being the reproductive flux (input of eggs at $w = w_{egg}$), $\phi \in [0,1]$ the mean proportion of mature female in each size class, M_{egg} the fraction of the spawned eggs which are not fecunded ($M_{egg} \in [0;1]$), $(1-\kappa)$ the fraction of the assimilated energy which is allocated to reproduction and w_{egg}, the weight of eggs.

2.2.4. The starvation mortality: calculation of $M^{starv}_{t,w}$

When starvation occurs, *i.e.* when the food ration is not sufficient to meet organism's needs, growth and/or reproduction cease and structural materials of the body are lysed and used for maintaining the most important physiological functions necessary for survival (Kooijman, 2000). The starvation process leads to a quick weakening of organisms which increases mortality. At the ecosystem level, starvation is a net dissipation of energy. To conserve the mass in a consistent way when growth and/or reproduction cease due to insufficient food intake (cf. Eqs. (10) and (11)), it is considered that the quantity of energy which is needed for maintenance but which cannot be provided by food intake is removed from the ecosystem. In this perspective, starvation acts as a mortality term a the level of the ecosystem and the starvation mortality coefficient can be expressed as follows using Eqs. (10) and (11):

$$M^{starv}_{t,w} = \left[\frac{\mu}{\psi} - \frac{\kappa e_A\omega w^{-1/3}\int_{v=0}^{w_{max}} s_{w,v}\xi_{t,v}\mathrm{d}v}{\frac{c}{w^\chi}+\int_{v=0}^{w_{max}} s_{w,v}\xi_{t,v}\mathrm{d}v}\right]^+ + \left[\frac{(1-\kappa)}{\kappa}\frac{\mu}{\psi} - \frac{(1-\kappa)e_A\omega w^{-1/3}\int_{v=0}^{w_{max}} s_{w,v}\xi_{t,v}\mathrm{d}v}{\frac{c}{w^\chi}+\int_{v=0}^{w_{max}} s_{w,v}\xi_{t,v}\mathrm{d}v}\right]^+ \tag{12}$$

2.2.5. The nonpredatory mortality: calculation of $Z_{t,w}$

The mortality for other causes than predation includes diseases, parasites, ageing, etc. Since large organisms exhibit much longer life span than small organisms (e.g., Speakman, 2005), it is simply supposed to be a decreasing allometric function.

$$Z_w = Ml^{\upsilon} = M\left(\frac{w}{a}\right)^{\upsilon/3} \tag{13}$$

With M being the nonpredatory mortality rate for a 1 m long organism (s^{-1}), l being body size (m), a (kg m^{-3}) being the coefficient linking weight to cubed length ($w = al^3$) and υ a parameter.

2.3. Conservation of energy

In our model, primary production is the only supply of energy to the system. This is appropriate in open ocean ecosystems where phytoplankton is the only energy input at the basis of the food chain. Energy is injected into producer size classes which do not grow. It is only transferred to consumers through predation. The model formulation is energy conservative and losses from the system occur only through nonpredatory mortality ($M > 0$), loss of male sexual products ($\phi < 1$) and dissipation processes such as imperfect efficiency of the assimilation process ($e_A < 1$), maintenance expenditures ($\mu > 0$) and energetic cost of growth ($E_g > 0$). If $\Pi = \mu = M = E_g = 0$ and $e_A = \phi = 1$, the total quantity of energy in the system is conserved and kept constant (even if its distribution in the weight-spectrum changes through time).

2.4. Temperature effect on physiological rates

Due to its major importance in controlling chemical reactions, temperature strongly influences metabolic rates of living organisms (Clarke and Johnston, 1999; Kooijman, 2000; Pörtner, 2002; Clarke, 2004; Speakman, 2005). Despite its purely molecular basis, the description of Arrhenius (Fig. 5) based on the van't Hoff equation ($k(T) = k_\infty e^{\left(-\frac{E_a}{RT}\right)}$ with k a reaction rate, k_∞ the frequency factor, E_a the activation energy, R the gas constant and T (K) the ambient temperature) fits well temperature effects on the physiological rates of organisms at the intra-specific level (Kooijman, 2000; Clarke and Fraser, 2004). Such effects are especially important to take into account given that most marine organisms are poïkilotherms and hence their internal temperature equals ambient water temperature which is potentially variable. The Arrhenius equation does not keep a mechanistic meaning at the inter-specific level (Clarke, 2004; Clarke and Fraser, 2004). However, it still provides a good statistical description of temperature effects on metabolic rates at the ecosystem level, even if purely chemical effects are altered by complex eco-evolutionary processes acting at this scale (Clarke and Johnston, 1999; Gillooly et al., 2001, 2002; Enquist et al., 2003; Clarke, 2004; Clarke and Fraser, 2004). In our model, the Arrhenius temperature-dependent correction factor $A(T)$ is used to correct ingestion rate, maintenance rate, nonpredatory mortality rate and swimming speed.

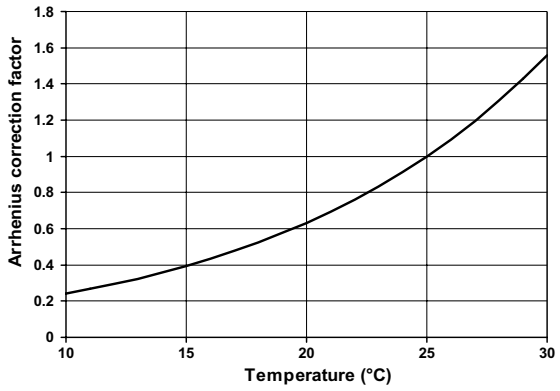

Fig. 5. Arrhenius correction factor for temperatures ranging from 10 °C to 30 °C ($T_A = 8000$ and $T_{ref} = 298.15K = 25$ °C). Each biological rate in the model is multiplied by the Arrhenius correction factor.

$$\mathrm{rate}(T) = \mathrm{rate}(T_{\mathrm{ref}}) \cdot A(T) \quad \text{with } A(T) = \mathrm{e}^{\left(\frac{\tau_A}{T_{\mathrm{ref}}} - \frac{\tau_A}{T}\right)} \tag{14}$$

With T_{ref} (K), the reference temperature and τ_A, a parameter (the "Arrhenius temperature" which equals $\frac{E_a}{R}$).

Combining Eqs. (1)–(3), (8) and (10)–(14) gives the full model which is presented in a compact form in Appendix A.

2.5. Numerical approximation

Marine ecosystems encompass very different organisms ranging from very small organisms such as phytoplankton cells (10^{-6} m, 10^{-16} kg) to very large organisms such as adult fish predators (4 m and more than 650 kg for giant bluefin tuna or swordfish for instance). To account accurately for growth and predation processes over such a large range of size would require numerically approximating the model with an extremely small resolution over an extremely high number of size intervals. Alternatively, a base α log scale can be used to ensure that processes are considered at the proper resolution whatever the size of organisms is and to keep a limited number of weight classes. Using such a length-based log scale can be done by defining $\varpi = \frac{\ln(l-\beta)}{\ln(\alpha)} - \gamma = \frac{\ln(a^{-1/3}w^{1/3}-\beta)}{\ln(\alpha)} - \gamma \iff w = a(\alpha^{\varpi+\gamma} + \beta)^3$ with α and β being fixed parameters and $\varpi = \{1, 2, 3, \ldots, n\}$. To be able to choose easily the grid characteristics, the parameters β and γ are expressed in terms of $l_{\min}$ and $l_{\max}$ which are fixed so that the grid depends only on α (Fig. 6). Because the present study focuses mostly on large consumer organisms such as fish or large meso-zooplankton ranging from 1 mm to 2 m, α is set at 1.04 which corresponds to grid cells varying from 1.5 mm for the smallest size class to 75 mm for the largest class. An irregular grid is derived calculating weight steps δw_i so that each grid point w_i is placed at the middle of its associated grid cell (Fig. 6a). The first grid point which represents producers is placed at 1.24×10^{-3} m which

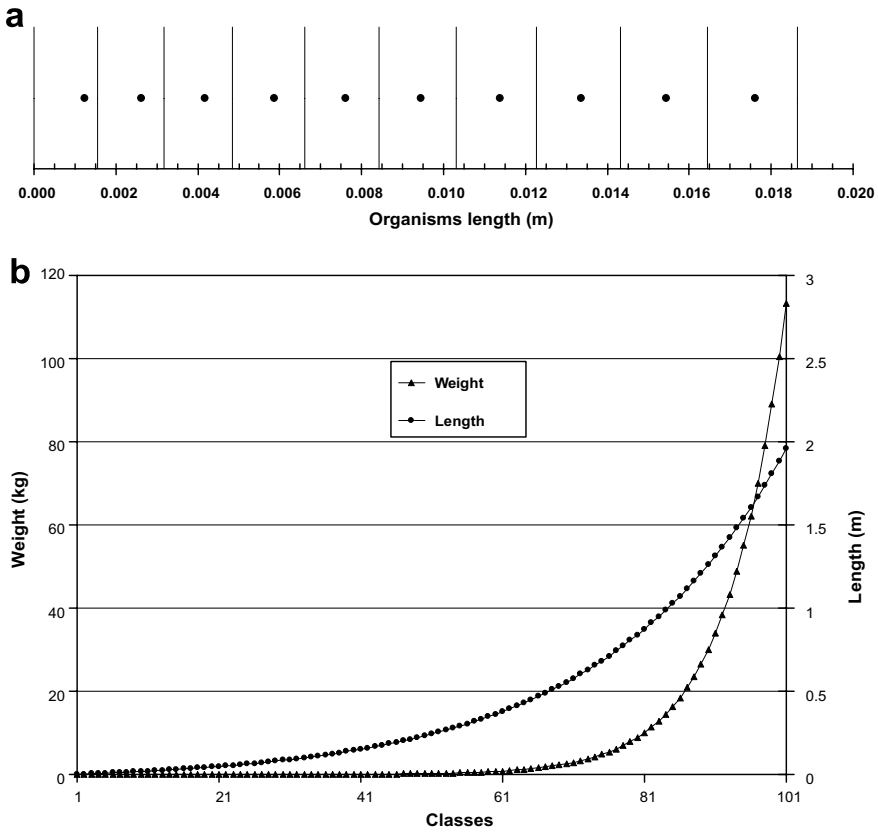

Fig. 6. (a) First 10 grid points (black dots) of the discretization used to approximate numerically the model and their associated size classes (vertical bars) used to calculate the integrals. (b) Full length/weight grid used for numerical simulations of the model. Each point represents the mean weight and the corresponding size of each of the 101 grid cells.

corresponds to the $]10^{-5}$ m, 1.56×10^{-3} m[size range. This size range obviously exceeds the phytoplankton size range (which roughly extends from 10^{-6} m to 10^{-4} m) and covers also microzooplankton and small meso-zooplankton sizes. However, it has to be kept in mind that our paper aims at representing the behaviour of a generic size-spectrum model formulated independently from the size range considered. To optimize computation time, the discretization used here focuses mostly on large organisms, as an illustration. Using our model to represent specifically small organisms such as small copepods would require refining the discretization used for small sizes. Such a grid refinement would not change the qualitative behavior of the model (but would be more costly in terms of computing time, allowing less simulations to be made). In this perspective, two coupled size-spectrum could profitably be used, one for small zooplankton and one for larger organisms such as fish and large zooplankton.

The model is integrated numerically along 101 length/weight classes from $l_{\min} = 10^{-5}$ m to $l_{\max} = 2$ m (Fig. 6b). Producers are assumed to occupy only the first length/weight class and consumers to range from the second to the 101th class (no overlap between their respective ranges).

Integrals are evaluated using first order approximations. Since the growth rate cannot be negative, a usual first order upwind finite difference scheme explicit in time is used to integrate Eq. (2). Most of the parameters used in the model have a clear physiological or ecological significance and are well documented in the literature, in both experimental and theoretical studies. The values used for simulations are given in Table 1 with the corresponding references of the literature. The maximum surface-specific ingestion rate ω as well as the maintenance rate μ are estimated given mean von Bertalanffy (1969) parameters (growth rate K and asymptotic size L_∞) of fish (cf. Appendix A). The estimation of nonpredatory mortality rate (parameters M and υ) is based on assumptions about the size-dependent mean life duration of marine organisms (Appendix B).

The value of Π_t, the primary energy production which enters the system is calculated so that the stationary concentration of phytoplankton in the reference simulations matches the value of 3144.225 J/m^3 of seawater which is approximately equivalent to 10^{-3} N mol m^{-3} and that we use as the reference concentration for producers (multiplying the redfield ratio C:N = 106:16 by the biomass free energy which is 474.6 kJ C mol^{-1} – Kooijman, 2000 – gives 3,144,225 J/mol of N). This value is then divided by the weight range of producers in the model $[1.5 \times 10^{-14}$ kg, 5.72×10^{-8} kg] to obtain the value for the distribution function of the energy content of the producers $\xi^{\mathrm{p}}_{t,w} = 549.10^8$ J kg^{-1} m^{-3}. This values is obtained in the reference simulation using $\Pi_t = 1177$ J day^{-1} m^{-3}.

2.6. Simulation experiments

In a first set of simulations, the existence of a linear steady state is tested by running the model during 50 years. The sensitivity of the steady state to the individual value of the model parameters is then explored systematically. For that purpose, the parameters ω, μ, M_{egg}, M, υ, c, κ, ρ_1, ρ_2, e_{A}, E_{g} are varied individually in a large range around their reference values (Table 1) and the influence of their variations on the stationary size-spectrum is considered.

3. Model behaviour

3.1. Steady state

The first set of numerical experiments was conducted using the reference values of the parameters (Table 1). In stable environmental conditions (constant primary production and constant temperature), the distribution of energy in the ecosystem converges from any positive initial distribution to a stationary quasi-linear size-spectrum (Fig. 7a). Only the first point (the primary producers) departs from the linear spectrum as well as the largest length classes for which the spectrum is slightly curved downward due to the slowdown of growth for large sizes close to the asymptotic length.

Fig. 7b–e provides the reader with the time evolution of the functional response function, the growth coefficient, the nonpredatory, predatory and starvation mortality coefficients and the egg production per size classes at steady state and during the transition phase. At steady state, the functional response increases with organism size from the highly food-limited small sizes to the less limited large sizes (Fig. 7b). The growth rate

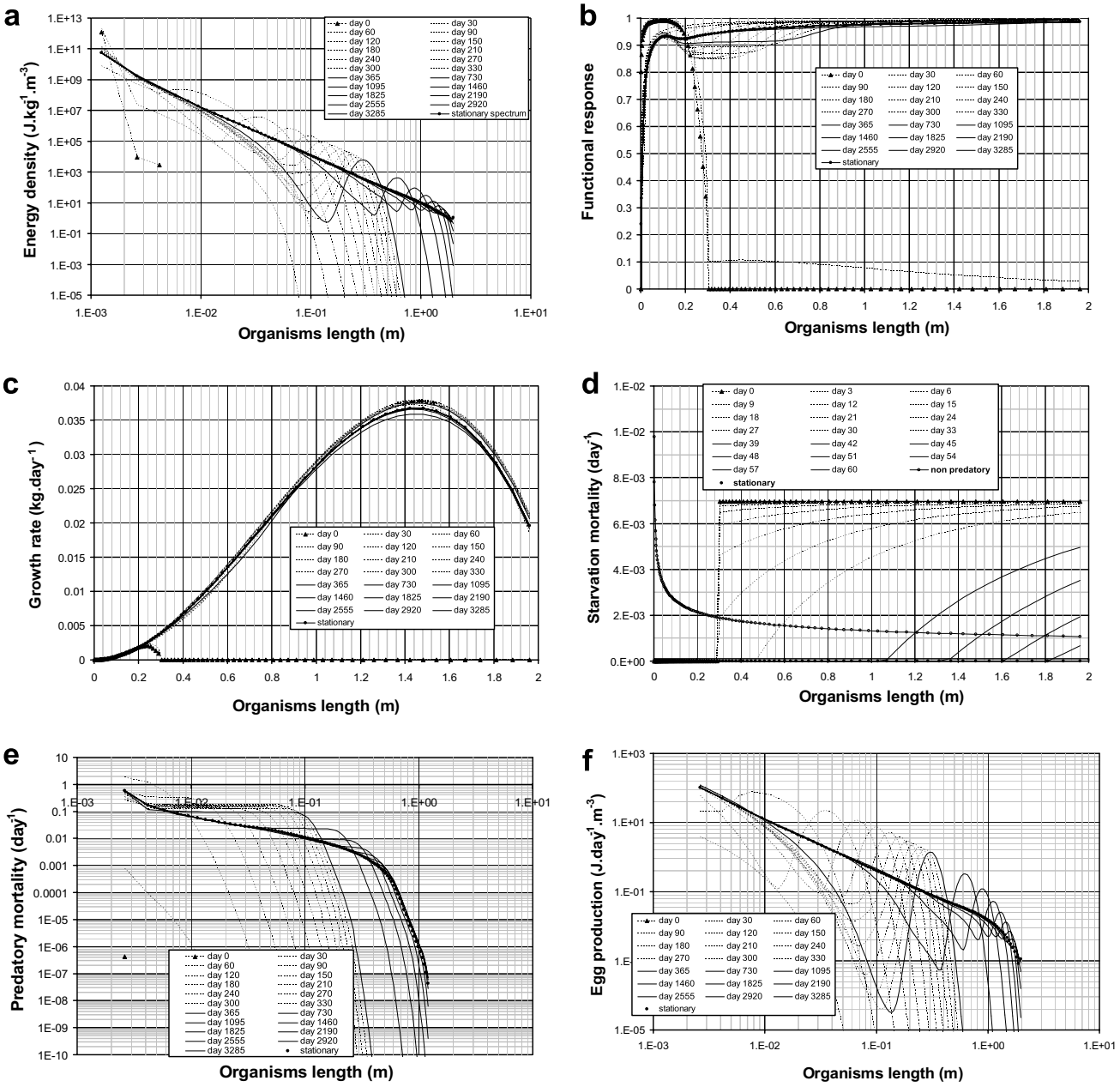

Fig. 7. Simulation of the transition toward the stationary state. (a) Size spectrum, (b) functional response, (c) growth rate, (d) starvation mortality and nonpredatory mortality (e) predatory mortality and (f) contribution of each size class to egg production. Triangles correspond to the initial energy distribution in the ecosystem (end of day 0), dotted lines correspond to the energy spectrum every 30 days except for starvation mortality where they are drawn every 3 days and continuous lines are drawn every 2 years after the first year. Black circles correspond to the steady state size-spectrum (after 50 years).

(in weight) as a function of organism size is dome-shaped, reaching a maximum for intermediate to large sizes and then decreasing down to zero for length equal to L_∞ (Fig. 7c). The log–log predatory mortality curve at steady state shows a quasi-linear decreasing trend for organisms between 2 mm and 20 cm (Fig. 7d) with higher mortality rates for producers. For larger organisms, the predation mortality decreases sharply down to zero for length above 70 cm. The log–log contribution of each size class to egg production (R_i) at steady state (Fig. 7e) exhibit a linearly decreasing trend with a downward curvature for sizes above 1.4 m, when maintenance processes are becoming to be non-negligible in Eq. (11).

When the reference values of the parameters (Table 1) are used, the slope of the stationary length-spectrum equals −3.175 which is equivalent to a slope equal to −1.058 for the weight-spectrum (Fig. 8).

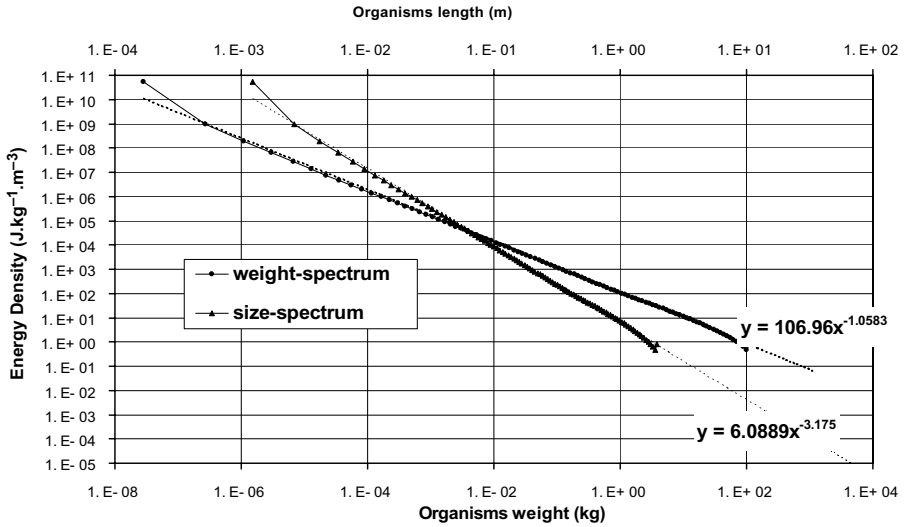

Fig. 8. Stationary size-spectrum and associated regression line as a function of weight (circles) and length (triangles).

3.2. Sensitivity to the parameters

The slope of the stationary size spectrum is not sensitive to the value of the maximum surface-specific ingestion rate (ω) but its intercept decreases when ω increases (the size spectrum is translated vertically, cf. Table 2 and Fig. 9a). The stationary size spectrum is not sensitive to the value of the maintenance rate μ (Table 2 and Fig. 9b). It has to be noted, however, that for length classes close to L_∞ (biomass is null for length greater than L_∞, cf. Appendix B), the stationary size-spectrum may departs from its linear shape and be curved downward. This is the case for low ω values or for high μ values (Table 2 and Fig. 9a and b).

Varying the value of the fraction of the spawned eggs which are not fertilized (M_{egg}) does not change the size spectrum over medium and large size classes (Table 2 and Fig. 9c). Only small size classes are sensitive to M_{egg} and depart from the linear solution when M_{egg} is smaller than 0.4. Conversely, the nonpredatory mortality coefficient M only influences the large classes of the size-spectrum, leading to a spectrum curved downward for high M values (Table 2 and Fig. 9d). Over the explored range, the exponent v of the nonpredatory mortality length-dependence has almost no effect on the size-spectrum (Table 2 and Fig. 9e).

The Holling type II half-saturation constant c has only a weak effect on the stationary size spectrum slope. However it has to be noted that decreasing its value leads to lower phytoplankton and small organism biomass which departs from the linear size spectrum. Conversely, high values of c lead to smaller L_∞ (Table 2 and Fig. 9f).

Table 2
Qualitative summary of the sensitivity analysis of the model (slope, intercept and curvature of the stationary size spectrum) to the value of its main parameters

Parameter	Designation	Slope	Intercept	Curvature
M	Nonpredatory mortality for $l = 1$ m	0	0	++
υ	Exponent of the M length-dependence	0	0	0
M_{egg}	Fraction of the spawned eggs which are not fecunded	0	0	0
ω	Maximum Surface specific ingestion rate	0	−	−
κ	Fraction of the assimilated energy allocated to growth and somatic maintenance	0	0	−
e_A	Fraction of the ingested energy which is assimilated	+	0	−
E_g	Weight specific cost of growth	−	0	+
μ	Maintenance rate	0	0	+
ρ_1 and α_1	Minimum ratio of predator size over prey size	+	0	0
ρ_2 and α_2	Maximum ratio of predator size over prey size	+	0	0
C	Holling type II half-saturation constant	0	++	+

0 = no effects, + = positive effect and − = negative effect.

The fraction of the assimilated energy which is allocated to growth and somatic maintenance (κ) slightly influences the curvature of the spectrum for small lengths (Table 2 and Fig. 9g). κ also influences positively the L_∞ value (and hence the curvature of the spectrum for large lengths). For high values of κ, the model produces unstable oscillations (waves propagating from small to large size classes cf. Fig. 9g). This unstable oscillatory phenomenon does not appear when M_{egg} is set equal to 0 (Fig. 9h).

Increasing the size of the smallest prey that can be eaten by a given predator (decreasing ρ_2 and increasing α_2) decreases substantially the slope of the stationary size-spectrum (Table 2 and Fig. 9i). Increasing the size of the largest prey that can be eaten by a given predator (increasing ρ_1 and decreasing α_1) increases the slope of the stationary size-spectrum (Table 2 and Fig. 9j).

Decreasing the fraction of the ingested energy which is assimilated (e_A) slightly decreases both the slope of the size spectrum and L_∞ (Table 2 and Fig. 9k). On the contrary, an increase of the weight specific cost of growth E_g decreases both the slope of the size spectrum and L_∞ (Table 2 and Fig. 9l).

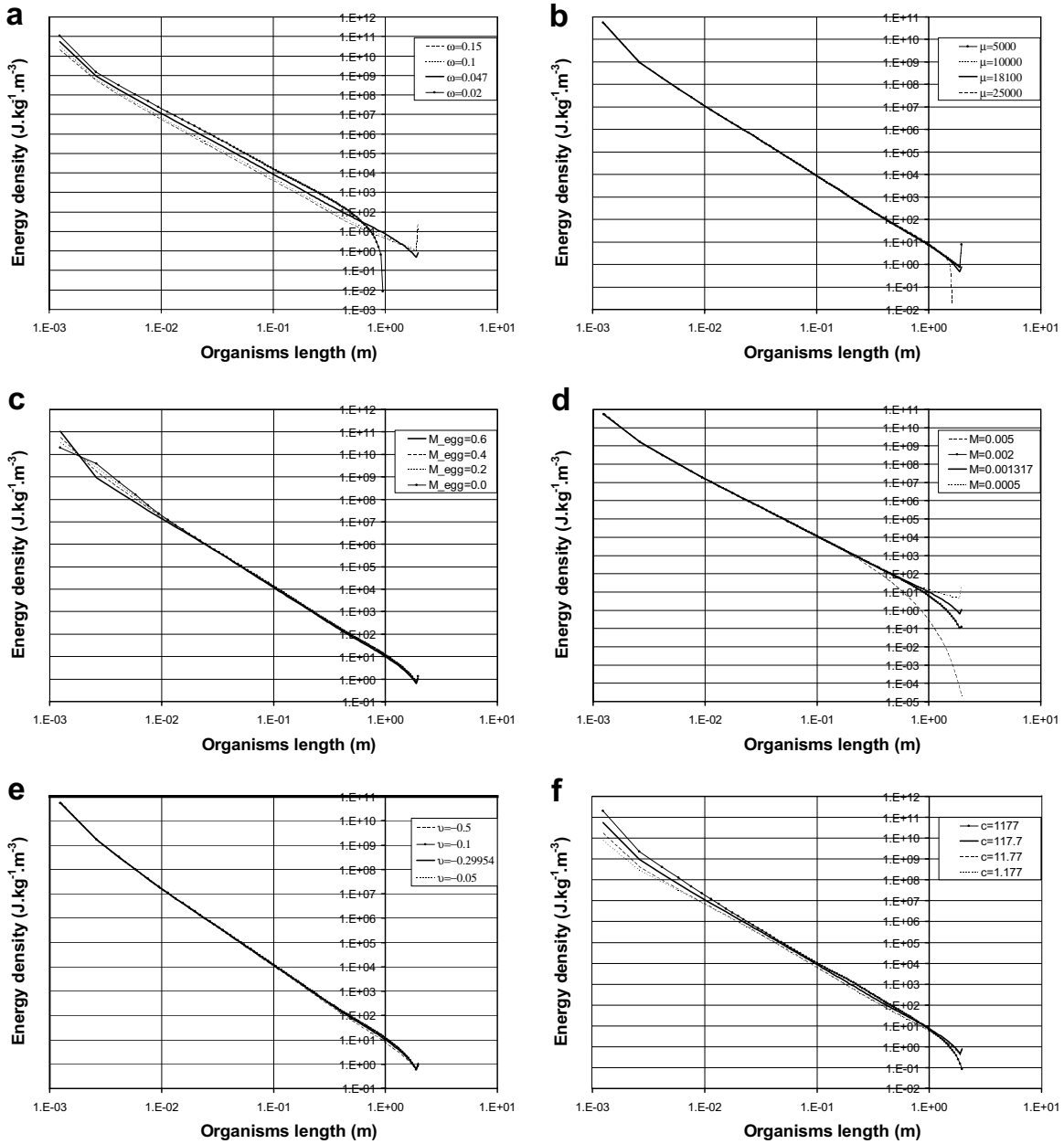

Fig. 9. Systematic sensitivity analysis of the steady state to the parameters. Different values of ω, μ, M_egg, M, v, c, κ, $\rho 2$, $\rho 1$, e_A and E_g varying in a large range around their reference values are considered respectively in (a)–(l). The sensitivity of the steady state to the parameter κ is considered in the case where $M\_egg = 0.4$ (g) and $M\_egg = 0$ (h).

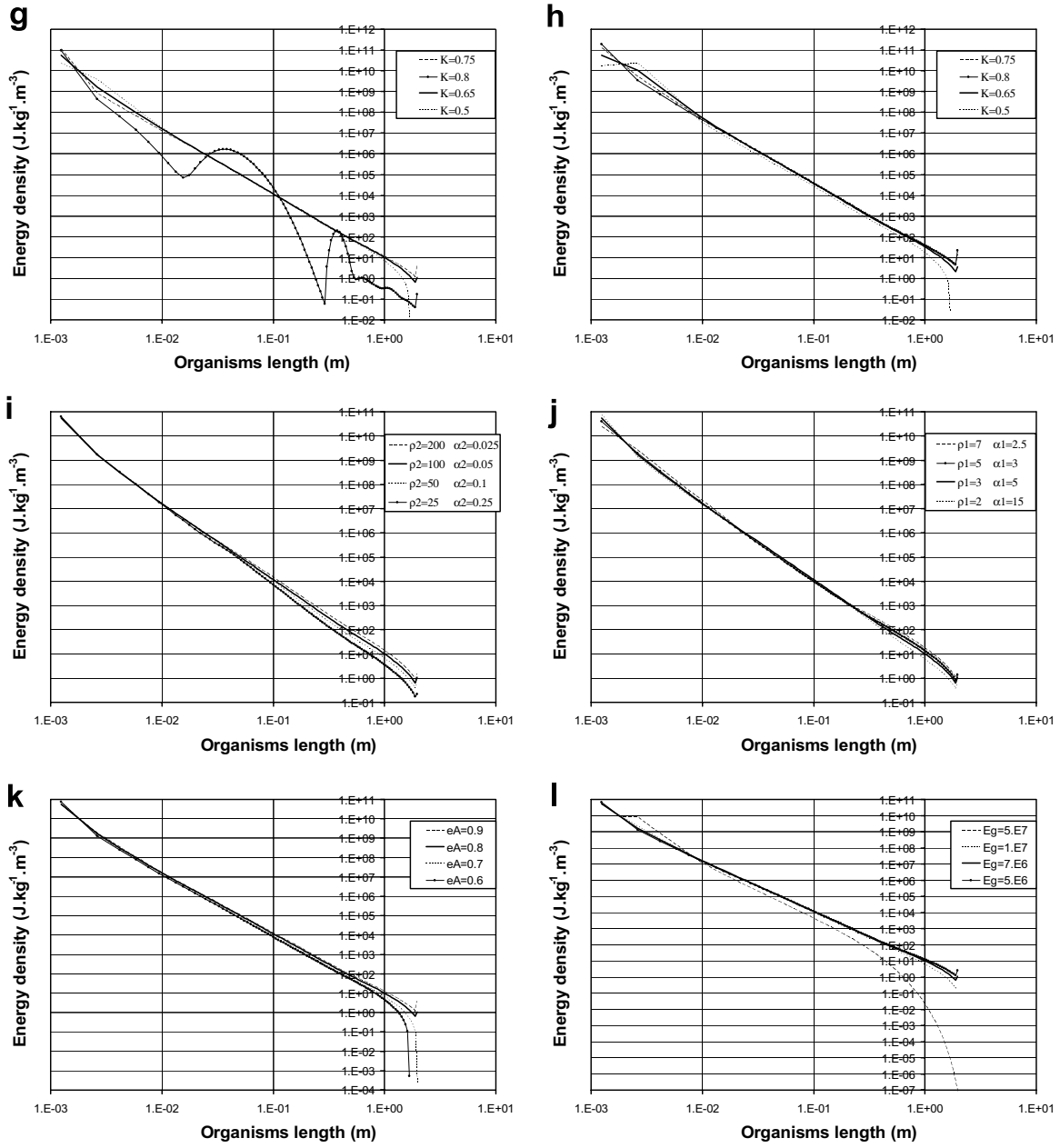

Fig. 9 (*continued*)

4. Discussion

4.1. Our model in the context of previous studies

The development of continuous size spectrum models based on allometric growth and mortality processes is a long lasting story in quantitative marine ecology (e.g., Platt and Denman, 1978; Silvert and Platt, 1978, 1980; Dickie et al., 1987; Cushing, 1992; Duplisea and Kerr, 1995; Arino et al., 2004; Benoit and Rochet, 2004). Models first dealt with constant growth rate. Later, Silvert and Platt (1980) assumed a constant size ratio between a predator and its prey. More recently, Arino et al. (2004) incorporated reproduction to the model and Benoit and Rochet (2004) linked explicitly the growth rate to the actual quantity of food being eaten and extended the predation process to any distribution of prey selectivity. In the model of Benoit and Rochet (2004), a given predator is supposed to eat all the potential preys swimming in a searched volume which increases allometrically with predator size. Like previous models, their model is built on a "supply system"

vision of the ecosystem: all the selected preys supplied in the "hunting volume" of the predator are eaten. Consequently, the growth rate of predators is not limited: if the biomass of prey tends to infinity, the growth rate of predators will also tend to infinity. Such a characteristic is not realistic and is furthermore likely to generate instability as reported by Benoit and Rochet (2004).

Conversely, our approach is based on a symmetrical "demand system" vision of the ecosystem: any organism in the ecosystem targets a maximal amount of energy proportional to its squared length to meet its growth, reproduction and maintenance needs and cannot eat more than this demand. Consequently, the growth rate of predators is limited: if the biomass of prey tends to infinity, the growth rate of predators tends to a maximum. Hence, in our model, a predator generates a mortality rate proportional to its maximal needs (and related to the biomass of prey with a Holling type II functional response) which is distributed over its prey range. Energy from prey is then shared between all their possible predators, proportionally to the mortality they exert. If predator needs for growth and/or reproduction are not satisfied, a starvation mortality coefficient is applied, which is proportional to the maintenance needs not fulfilled by assimilated energy. Our approach allows to take into account more biological and ecological processes (opportunistic size-structured predation, predators competition, allocation of energy between growth and reproduction, somatic and gonadic maintenance, starvation mortality) in a rigorous mass-balanced physiologically based formulation derived from the dynamic energy budget theory (Kooijman, 2000).

4.2. Stationary solutions

Numerical simulations show that the model produces stable solutions which do not need to be stabilized using diffusion or complex boundary conditions. In most cases with constant environmental conditions, the model converges toward a stationary log–log linear size-spectrum which is independent of initial conditions (Fig. 7). Numerically, 20 years are most of the time sufficient to approximate the stationary solution with a good precision. It is theoretically well established that size-structured predator–prey models admit a linear log–log size-spectrum as a stationary solution (Silvert and Platt, 1980; Arino et al., 2004; Benoit and Rochet, 2004) as far as the smallest sizes are put apart (Shin and Cury, 2004). Our simulations corroborate previous studies and show that this important property still holds when size-dependent opportunistic predation, predator competition, energy allocation between growth and reproduction, nonpredatory mortality and starvation mortality are explicitly taken into account as key processes governing energy flow through marine ecosystems.

From an ecological perspective, the distributed nature of predation over a large size range multiplies the weak links in ecosystems, and hence is likely to dampen oscillations between consumers and resources and enhance persistence and stability (McCann et al., 1998; McCann, 2000). In other respects, the stationary state can be considered as the "ultimate state of maturity" of an ecosystem as defined by Odum (1969). Being always submitted to perturbations, ecosystems are actually in a never-ending transient state of "maturation" toward their steady state "maturity".

Using our reference set of parameters, the slope of the simulated log–log biomass spectrum equals −1.06. This value matches fairly well with the values reported in empirical studies (e.g., Macpherson and Gordoa, 1996; Zhou and Huntley, 1997; Quiñones et al., 2003; Marquet et al., 2005). For the first size class of the spectrum however (the size class of the producers), the model departs from the linear solution. This is likely to be due to the poor representation of producers in the model, in particular to the lack of representation of phytoplankton growth and division. Furthermore, our numerical simulation grid, which focuses on consumer dynamics, has only one size-class for representing producers which likely leads to potential irregular solutions when approximating the integrals over small sizes. It has furthermore to be noted that for large sizes close to L_∞, the size spectrum is curved downward. This phenomenon corresponds in our model to the slowdown of growth around the maximum size.

4.3. Sensitivity of the simulated size-spectrum to the parameters

The slope, intercept and curvature of the stationary size-spectrum are generally not very sensitive to the parameters of the model, at least in the explored ranges (Table 2 and Fig. 9). The parameters can be classified

according to their qualitative effect on the size spectrum. Some parameters, such as the size of the smallest prey that can be eaten by a given predator (ρ_2), act only on the slope of the spectrum (cf. Fig. 9i and j) when others, such as the maximum surface specific ingestion rate (ω), act only on its intercept (Fig. 9a). Other parameters, such as the nonpredatory mortality coefficient (M), modify the curvature of the spectrum (Fig. 9d) when some others, such as the fraction of the spawned eggs which are not fertilized (M_egg), have only a local influence on the very small sizes of the spectrum (Fig. 9c). Finally, most parameters modify slightly the L_∞ value and hence influence the linearity of the spectrum for large sizes.

It has furthermore to be noted that, as suspected by Arino et al. (2004), for certain combinations of extreme values of the parameters, the stationary solution becomes unstable and oscillatory solutions appear, even in the case of stable phytoplankton production and constant temperature (Fig. 9g).

5. Conclusion

The proposed model improves previous studies by incorporating processes playing an important role in the energy fluxes through marine systems. It is furthermore based on a "demand system" approach which leads to more stable solutions than previously developed "supply system" models. Despite its simple ecological assumptions, the model seems to represent adequately the main qualitative and quantitative characteristics of marine size-spectra which have been reported in empirical studies and enables testable insights regarding the effect of environmental variability and changes on ecosystems. Those effects are explored through simulations in a companion paper (Maury et al., 2007) which focuses on temperature and primary production effects on the size spectrum.

However it has to be kept in mind that marine ecosystems encompass a large number of zoological groups which exhibit very different eco-physiological and behavioral characteristics. Each zoological group is in turn composed of a large number of species, each having various life histories (various growth rates, longevities and sizes at maturity). Hence, in real ecosystems, small organisms comprise adults of various small short-living species as well as juveniles of various large long-living species. Despite this obvious diversity, our model assumes constant physiological parameters and rules for any consumer organisms in the ecosystem. That could constitute a limitation of our approach since biodiversity plays important functional roles in ecosystems. This furthermore leads us to use simplified hypothesis about the reproduction process since all size classes are supposed to contain the same proportion of mature individuals. Formalizing and quantifying the effects of biodiversity in size-spectrum models is indeed critical and will be an important goal of our future work.

Acknowledgements

We thank Bernard Cazelles, Philippe Cury, Michel langlais, Alain Menesguen, Frédéric Ménard and Christian Mullon for their constructive criticisms of an earlier version of the manuscript and for their kind encouragements.

Appendix A. Full model equation

Combining Eqs. (1), (2), (7) and (9)–13 gives the full model equation:
$\forall w \in [0; w_1]$

$$\frac{\mathrm{d}\xi^{\mathrm{p}}_{t,w}}{\mathrm{d}t} = \frac{\Pi_t}{w_1} - A(T_t)\xi^{\mathrm{p}}_{t,w}\left(\frac{\omega}{w_1}\int_{x=0}^{w_1}\int_{u=w_{\mathrm{egg}}}^{w_{\max}}\left[\frac{\xi^{\mathrm{c}}_{t,u}u^{-1/3}s_{u,x}}{\frac{c}{u^{\chi}}+\int_{v=0}^{w_{\max}}s_{u,v}\xi_{t,v}\mathrm{d}v}\right]\mathrm{d}u\,\mathrm{d}x + M_I\right) \tag{15}$$

$w = w_{\mathrm{egg}}$

$$\gamma_{t,w_{\mathrm{egg}}}\xi^{\mathrm{c}}_{t,w_{\mathrm{egg}}} = (1-M_{\mathrm{egg}})\phi A(T_t)\int_{u=w_{\mathrm{egg}}}^{w_{\max}}\left[\frac{(1-\kappa)e_{\mathrm{A}}\omega\xi^{\mathrm{c}}_{t,u}u^{-1/3}\int_{v=0}^{w_{\max}}s_{u,v}\xi_{t,v}\,\mathrm{d}v}{\frac{c}{u^{\chi}A(T_t)}+\int_{v=w_{\min}}^{w_{\max}}s_{u,v}\xi_{t,v}\,\mathrm{d}v} - \frac{(1-\kappa)}{\kappa}\frac{\mu\xi^{\mathrm{c}}_{t,u}}{\psi}\right]^{+}\mathrm{d}u \tag{16}$$

$\forall w \in]w_{\mathrm{egg}}; w_{\max}]$

$$\begin{aligned}\frac{\partial \xi^{\mathrm{c}}_{t,w}}{\partial t} = & -\frac{\psi A(T_t)}{\psi + E_{\mathrm{g}}}\frac{\partial}{\partial w}\left(\left[\frac{\kappa e_{\mathrm{A}}\omega \int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\mathrm{d}v}{\frac{c}{w^{\chi}A(T_t)} + \int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v} w^{2/3} - \frac{\mu}{\psi} w\right]^{+} \xi^{\mathrm{c}}_{t,w}\right) \\ & - \omega A(T_t)\int_{u=w_{\mathrm{egg}}}^{w_{\max}}\left(\frac{\xi^{\mathrm{c}}_{t,u} u^{-1/3} s_{u,w}}{\frac{c}{u^{\chi}A(T_t)} + \int_{v=0}^{w_{\max}} s_{u,v}\xi_{t,v}\,\mathrm{d}v}\right)\mathrm{d}u\,\xi^{\mathrm{c}}_{t,w} \\ & - A(T_t)\left(M\left(\frac{w_1}{a}\right)^{v/3} + \left[\frac{\mu}{\psi} - \frac{\kappa e_{\mathrm{A}}\omega w^{-1/3}\int_{v=0}^{w_{\max}} s_{w,v}\xi_{t,v}\,\mathrm{d}v}{\frac{c}{w^{\chi}A(T_t)} + \int_{v=0}^{w_{\max}} s_{w,v}\xi_{t,v}\,\mathrm{d}v}\right]^{+}\right. \\ & \left. + \left[\frac{(1-\kappa)}{\kappa}\frac{\mu}{\psi} - \frac{(1-\kappa)e_{\mathrm{A}}\omega w^{-1/3}\int_{v=0}^{w_{\max}} s_{w,v}\xi_{t,v}\,\mathrm{d}v}{\frac{c}{w^{\chi}A(T_t)} + \int_{v=0}^{w_{\max}} s_{w,v}\xi_{t,v}\,\mathrm{d}v}\right]^{+}\right)\xi^{\mathrm{c}}_{t,w}\end{aligned} \tag{17}$$

where $[x]^{+}$ is the function defined by

$$\begin{cases}[x]^{+} = x & \text{if } x \geqslant 0 \\ [x]^{+} = 0 & \text{if } x < 0\end{cases}, \quad s_{u,w} = \left(1 + \mathrm{e}^{\alpha_1\left(\rho_1 - \left(\frac{u}{w}\right)^{1/3}\right)}\right)^{-1}\left(1 - \left(1 + \mathrm{e}^{\alpha_2\left(\rho_2 - \left(\frac{u}{w}\right)^{1/3}\right)}\right)^{-1}\right)$$

is the size-dependent selectivity function of preys of weight w by predators of weight u; and $A(T) = \mathrm{e}^{\left(\frac{\tau_{\mathrm{A}}}{T_{\mathrm{ref}}} - \frac{\tau_{\mathrm{A}}}{T}\right)}$ is the Arrhenius temperature-dependant correction factor.

Appendix B. Calculation of ω and μ as a function of the Von Bertalanffy growth parameters K and L_∞

At food saturation, our growth Eq. (9) is related as follows to the Von Bertalanffy growth equation:

$$\frac{\mathrm{d}w}{\mathrm{d}t} = Aw^{2/3} - Bw \quad \text{with} \begin{cases} A = \frac{\psi}{\psi + E_{\mathrm{g}}}\kappa e_{\mathrm{A}}\omega \\ B = \frac{\mu}{\psi + E_{\mathrm{g}}}\end{cases} \tag{18}$$

This well known equation can be rewritten in length and integrated between $l = 0$ and $l = l_t$ to get l_t as a function of time:

$$\frac{\mathrm{d}(al^3)}{\mathrm{d}t} = 3al^2\frac{\mathrm{d}l}{\mathrm{d}t} = A(al^3)^{2/3} - Bal^3 \quad \Longleftrightarrow \quad \frac{\mathrm{d}l}{\mathrm{d}t} = \frac{Aa^{-1/3}}{3} - \frac{B}{3}l \tag{19}$$

which after integration gives:

$$l_t = \frac{Aa^{-1/3}}{B}\left(1 - \mathrm{e}^{\frac{-B}{3}(t - t_0)}\right) \tag{20}$$

This expression is used to express ω and μ as a function of the Von Bertalanffy growth parameters K and L_∞:

$$\begin{cases} L_\infty = \frac{Aa^{-1/3}}{B} = \frac{\kappa e_{\mathrm{A}}\omega\psi a^{-1/3}}{\mu} \\ K = \frac{B}{3} = \frac{\mu}{3\left(\psi + E_{\mathrm{g}}\right)}\end{cases} \quad \Longleftrightarrow \quad \begin{cases}\omega = \frac{3KL_\infty\left(\psi + E_{\mathrm{g}}\right)}{\kappa e_{\mathrm{A}}\psi a^{-1/3}} \\ \mu = 3K(\psi + E_{\mathrm{g}})\end{cases} \tag{21}$$

For the numerical applications presented in the present paper, an asymptotic length $L_\infty = 2.2$ m is assumed with a corresponding growth rate $K = 0.2$ year^{-1} deduced from the mean statistical relationships observed between K and L_∞ by Froese and Pauly (2000).

Appendix C. Estimation of the mortality parameters M and v

To estimate the parameters M and v which determine the length-dependent nonpredatory mortality, five groups of organisms having very different mean length are considered (diatoms, copepods, and three fish of 0.1 m, 0.8 m and 1.7 m). For each group an arbitrary life span is attributed and the corresponding mortality

Table 3
Estimation of the nonpredatory mortality parameters M and υ (see text)

Species	Mean size	Estimated life span	Estimated mortality	Modeled mortality MI^{υ}
Diatoms	5×10^{-5} m	90 days	2.56×10^{-2} day^{-1}	2.56×10^{-2} day^{-1}
Copepods	5×10^{-4} m	180 days	1.28×10^{-2} day^{-1}	1.28×10^{-2} day^{-1}
Fish 0.1 m	0.1 m	730 days	3.15×10^{-3} day^{-1}	2.63×10^{-3} day^{-1}
Fish 0.8 m	0.8 m	1825 days	1.26×10^{-3} day^{-1}	1.41×10^{-3} day^{-1}
Fish 1.7 m	1.7 m	2555 days	9.01×10^{-4} day^{-1}	1.12×10^{-3} day^{-1}

is estimated assuming that the life span corresponds to the age at which only 10% of a cohort remains (Table 3). The parameters M and υ are estimated by fitting the modeled mortality curve to the estimated mortality curve (Table 3).

References

Andersen, N.G., Riis-Vestergaard, J., 2003. The effects of food consumption rate, body size and temperature on net food conversion efficiency in saithe and whiting. Journal of Fish Biology 62, 395–412.

Arino, O., Shin, Y.-J., Mullon, C., 2004. A mathematical derivation of size spectra in fish populations. Comptes Rendus de l'Académie des Sciences. Section Biologies 327, 245–254.

Benoit, E., Rochet, M.J., 2004. A continuous model of biomass size spectra governed by predation and the effects of fishing on them. Journal of Theoretical Biology 226, 9–21.

Blueweiss, L., Fox, H., Kudzma, V., Nakashima, D., Peters, R., Sams, S., 1978. Relationships between body size and some life history parameters. Oecologia 37, 257–272.

Bone, Q., Marshall, N.B., Blaxter, J.H.S., 1999. Biology of Fishes. Stanley Thornes, p. 332.

Brett, J.R., Groves, T.D.D., 1979. Physiological energetics. In: Hoar, W.S., Randall, D.J., Brett, J.R. (Eds.), Fish Physiology. Academic press.

Brill, R.W., Guernsey, D.L., Stevens, E.D., 1978. Body surface and gill heat loss rates in restrained Skipjack Tuna. In: Sharp, G.D., Dizon, A.E. (Eds.), The Physiological Ecology of Tunas. Academic Press, p. 277.

Brown, J.H., Gillooly, J.F., 2003. Ecological food webs: high-quality data facilitate theoretical unification. Proceedings of the National Academy of Sciences (USA) 100, 1467–1468.

Clarke, A., 2004. Is there a universal temperature dependence of metabolism? Functional Ecology 18, 252–256.

Clarke, A., Fraser, K.P.P., 2004. Why does metabolism scale with temperature? Functional Ecology 18, 243–251.

Clarke, A., Johnston, N.M., 1999. Scaling of metabolic rate with body mass and temperature in teleost fish. Journal of Animal Ecology 68, 893–905.

Cousins, S.H., 1980. A trophic continuum derived from plant structure, animal size and a detritus cascade. Journal of Theoretical Biology 82, 607–618.

Cury, P., Pauly, D., 2000. Patterns and propensities in reproduction and growth of marine fishes. Ecological Research 15, 101–106.

Cury, P., Shannon, L., Shin, Y.-J., 2003. The functioning of marine ecosystems a fisheries perspective. In: Sinclair, M., Valdimarsson, G. (Eds.), Responsible Fisheries in the marine Ecosystem. FAO, Cabi, 462 pp.

Cushing, J.M., 1992. A size-structured model for cannibalism. Theoretical Population Biology 42, 347–361.

Daan, N., 1975. Consumption and production in north sea cod: an assessment of the ecological status of the stock. Netherlands Journal of sea research 9, 24–55.

Dickie, L.M., Kerr, S.R., Boudreau, P.R., 1987. Size-dependent processes underlying regularities in ecosystem structure. Ecological Monograph 57 (3), 233–250.

Duplisea, D.E., Kerr, S.R., 1995. Application of a biomass size spectrum model to demersal fish data from the Scotian Shelf. Journal of Theoretical Biology 177, 263–269.

Edwards, R.R.C., Finlayson, D.M., Steele, J.H., 1972. An experimental study of the oxygen consumption, growth and metabolism of the cod. Journal of Experimental Marine Biology and Ecology 8, 299–309.

Enquist, B.J., Economo, E.P., Huxman, T.E., Allen, A.P., Ignace, D.D., Gillooly, J.F., 2003. Scaling metabolism from organisms to ecosystems. Nature 423, 639–642.

Essington, T.E., Kitchell, J.F., Walters, J.C., 2001. The von bertalanffy growth function, bioenergetics, and the consumption rates of fish. Canadian Journal of Fishery and Aquatic Science 58, 2129–2138.

Floeter, J., Temming, A., 2003. Explaining diet composition of North Sea cod (Gadus morhua): prey size preference vs. prey availability. Canadian Journal of Fishery and Aquatic Science 60, 140–150.

Froese, R., Pauly, D., (Eds.), 2000. FishBase 2000: concepts, design and data sources. ICLARM, Los Baños, Laguna, Philippines, 344 pp.

Gillooly, J.F., Brown, J.H., West, G.B., Savage, V.M., Charnov, E.L., 2001. Effects of size and temperature on metabolic rate. Science 293, 2248–2251.

Gillooly, J.F., Charnov, E.L., West, G.B., Savage, V.M., Brown, J.H., 2002. Effects of size and temperature on developmental time. Nature 417, 70–73.
Jennings, S., Pinnegar, J.K., Polunin, N.V.C., Boon, T., 2001. Weak cross-species relationships between body size and trophic level belie powerful size-based trophic structuring in fish communities. Journal of Animal Ecology 70, 934–944.
Jennings, S., Pinnegar, J.K., Polunin, N.V.C., Warr, K.J., 2002. Linking size-based and trophic analyses of benthic community structure. Marine Ecology Progress Serie 226, 77–85.
Juanes, F., 2003. The allometry of cannibalism in piscivorous fishes. Canadian Journal of Fishery and Aquatic Science 60, 594–602.
Kitchell, J.F., Neill, W.H., Dizon, A.E., Magnuson, J.J., 1978. Bioenergetic spectra of skipjack and yellowfin tunas. In: Sharp, G.D., Dizon, A.E. (Eds.), The Physiological Ecology of Tunas. Academic Press, p. 357.
Kooijman, S.A.L.M., 1986. Energy budgets can explain body size relations. Journal of Theoretical Biology 121, 269–282.
Kooijman, S.A.L.M., 1995. The stoichiometry of animal energetics. Journal of Theoretical Biology 177, 139–149.
Kooijman, S.A.L.M., 2000. Dynamic Energy and Mass Budgets in Biological Systems. Cambridge University Press.
Kooijman, S.A.L.M., 2001. Quantitative aspects of metabolic organization: a discussion of concepts. Philosophical Transactions of the Royal Society of London B356, 331–349.
Kot, M., 2001. Elements of Mathematical Ecology. Cambridge University Press, Cambridge, p.453.
Krohn, M., Reidy, S., Kerr, S., 1996. Bioenergetic analysis of the effects of temperature and prey availability on growth and condition of nothern cod. Canadian Journal of Fishery and Aquatic Science 54, 113–121.
Lundvall, D., Svanbäck, R., Persson, L., Byström, P., 1999. Size-dependent predation in piscivores: interactions between predator foraging and prey avoidance abilities. Canadian Journal of Fishery and Aquatic Science 56, 1285–1292.
Macpherson, E., Gordoa, A., 1996. Biomass spectra in benthic fish assemblages in the Benguela system. Marine Ecology Progress Serie 138, 27–32.
McCann, K.S., Hastings, A., Huxel, G.R., 1998. Weak trophic interactions and the balance of nature. Nature 395, 794–798.
McCann, K.S., 2000. The diversity–stability debate. Nature 405, 228–233.
Marquet, P.A., Quiñones, R.A., Abades, S., Labra, F., Tognelli, M., Arim, M., Rivadeneira, M., 2005. Scaling and power-laws in ecological systems. The Journal of Experimental Biology 208, 1749–1769.
Maury, O., Shin, Y.-J., Faugeras, B., Ben Ari, T., Marsac, F., 2007. Modeling environmental effects on the size-structured energy flow through marine ecosystems. Part 2: simulations. Progress in Oceanography 74 (4), 500–514.
Ménard, F., Labrune, C., Shin, Y.-J., Asine, A.-S., Bard, F.-X., 2006. Opportunistic predation in tuna: a size-based approach. Marine Ecology Progress Series 323, 223–231.
Nisbet, R.M., Muller, E.B., Lika, K., Kooijman, S.A.L.M., 2000. From molecules to ecosystems through dynamic energy budget models. Journal of Animal Ecology 69, 913–926.
Odum, E.P., 1969. The strategy of ecosystem development. Science 164, 262–270.
Pauly, D., Christensen, V., Walters, C., 2000. Ecopath, Ecosim, and Ecospace as tools for evaluating ecosystem impact of fisheries. ICES Journal of Marine Science 57, 697–706.
Platt, T., Denman, K., 1978. The structure of pelagic marine ecosystems. Rapp. Procés-Verbaux des Réunions du Conseil International d'Exploration de la Mer 173, 60–65.
Platt, T., Denman, K., 1997. Organisation in the pelagic ecosystem. Helgol. Wiss. Meeresunters 30, 575–581.
Polovina, J.J., 1984. Model of a coral reef ecosystem. I: the ECOPATH model and its application to French Frigate Schoals. Coral Reefs 3, 1–11.
Pörtner, H.O., 2002. Physiological basis of temperature-dependent biogeography: trade-offs in muscle design and performance in polar ectotherms. The Journal of Experimental Biology 205, 2217–2230.
Quiñones, R.A., Platt, T., Rodríguez, J., 2003. Patterns of biomass size-spectra from oligotrophic waters of the Northwest Atlantic. Progress in Oceanography 57, 405–427.
Scharf, F.S., Juanes, F., Rountree, R.A., 2000. Predator size–prey size relationships of marine fish predators: interspecific variation and effects of ontogeny and body size on trophic-niche breadth. Marine Ecology Progress Serie 208, 229–248.
Sheldon, R.W., Prakash, A., Sutcliffe, W.H.J., 1972. The size distribution of particles in the ocean. Limnological Oceanography 17 (3), 327–340.
Shin, Y.-J., Cury, P., 2004. Using an individual-based model of fish assemblages to study the response of size spectra to changes in fishing. Canadian Journal of Fishery and Aquatic Science 61, 414–431.
Silvert, W., Platt, T., 1978. Energy flux in the pelagic ecosystem: a time-dependent equation. Limnological Oceanography 23, 813–816.
Silvert, W., Platt, T., 1980. Dynamic energy flow model of the particle size distribution in pelagic ecosystems. In: Kerfoot, W. (Ed.), Evolution and Ecology of Zooplancton Communities. University Press of New England, Illanover, NH, pp. 754–763.
Speakman, J.R., 2005. Body size, energy metabolism and lifespan. The Journal of Experimental Biology 208, 1717–1730.
Sterner, R.W., Elser, J.J., 2002. Ecological Stoichiometry. Princeton University Press, Princeton, p. 439.
Tuljapurkar, S., Caswell, H. (Eds.), 1997. Structured-population models in marine, terrestrial, and freshwater systems. Chapman & Hall, p. 643.
Valiela, I., 1995. Marine Ecological Processes. Springer-Verlag, New York, p. 686.
van der Veer, H.W., Kooijman, S.A.L.M., van der Meer, J., 2003. Body size scaling relationships in flatfish as predicted by Dynamic Energy Budgets (DEB theory): implications for recruitment. Journal of Sea Research 50, 255–270.
von Bertalanffy, L., 1969. Basic concepts in quantitative biology of metabolism. Helgoland Wissenshaft Meeresunters 9, 5–34.
Walters, C., Christensen, V., Pauly, D., 1997. Structuring dynamic models of exploited ecosystems from trophic massbalance assessments. Review in Fish Biology and Fisheries 7 (2), 139–172.

West, G.B., Brown, J.H., 2005. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure. The Journal of Experimental Biology 208, 1575–1592.

Woodward, G., Ebenman, B., Emmerson, M., Montoya, J.M., Olesen, J.M., Valido, A., Warren, P.H., 2005. Body size in ecological networks. Trends in Ecology and Evolution 20 (7), 402–409.

Zhou, M., Huntley, M.E., 1997. Population dynamics theory of plankton based on biomass spectra. Marine Ecology Progress Serie 159, 61–73.

Article L : [21] B. Faugeras, J. Pousin et F. Fontvieille. An efficient numerical scheme for precise time integration of a diffusion - dissolution / precipitation chemical system. *Math. of Computation* 75.253 (2006), p. 209–222

MATHEMATICS OF COMPUTATION
Volume 75, Number 253, Pages 209–222
S 0025-5718(05)01782-5
Article electronically published on September 29, 2005

AN EFFICIENT NUMERICAL SCHEME FOR PRECISE TIME INTEGRATION OF A DIFFUSION-DISSOLUTION/PRECIPITATION CHEMICAL SYSTEM

BLAISE FAUGERAS, JÉRÔME POUSIN, AND FRANCK FONTVIEILLE

Abstract. A numerical scheme based on an operator splitting method and a dense output event location algorithm is proposed to integrate a diffusion-dissolution/precipitation chemical initial-boundary value problem with jumping nonlinearities. The numerical analysis of the scheme is carried out and it is proved to be of order 2 in time. This global order estimate is illustrated numerically on a test case.

1. Introduction

In this article we address the problem of the numerical integration of a complex diffusion-dissolution/precipitation chemical system of equations constituted of partial differential equations and ordinary differential equations with nonlinear discontinuous right hand side. Such systems arise in models describing the retention capacity of concrete matrices in which wastes and pollutants are embedded. The particular model we have in mind is described and studied from a mathematical point of view in [7] and [8]. It takes into account the influence of the chemical context evolution on the dissolution/precipitation rates and expresses the necessary presence of solid for dissolution by an obstacle problem. The multi-species diffusion-dissolution/precipitation model takes the form of an initial-boundary value problem in which partial differential equations (PDEs) and ordinary differential equations (ODEs) are coupled through nonlinear discontinuous terms. The system of equations for N_s species is formulated as follows. $\mathbf{C} = (C_i)_{i=1,\dots,N_s}$ is the vector of chemical species concentrations in liquid phase and $\mathbf{S} = (S_i)_{i=1,\dots,N_s}$ is the vector of chemical species concentrations in solid phase. C_i^* are nonlinear functions of C representing saturation concentrations, α_i and D_i are strictly positive constants. The following notation is also used

$$\forall z \in \mathbb{R},\ z^+ = \max(z, 0) \quad \text{and } z^- = z^+ - z \geq 0,$$

and

$$\mathrm{sgn}^+(z) = \begin{cases} 1 \text{ if } z > 0, \\ 0 \text{ otherwise.} \end{cases}$$

Received by the editor December 2, 2003 and, in revised form, September 20, 2004.
2000 *Mathematics Subject Classification.* Primary 65M12, 65G99, 35K57.
Key words and phrases. Numerical time integration, operator splitting, dense output, high order, error analysis, reaction-diffusion, jumping nonlinearities.

For $i = 1$ to N_s, we have

(1.1)
$$\begin{cases} \partial_t C_i = D_i \Delta C_i + \operatorname{sgn}^+(S_i)\alpha_i(C_i^*(C) - C_i)^+ - \alpha_i(C_i^*(C) - C_i)^- & \text{in } (0,T)\times\Omega, \\ \partial_t S_i = -\operatorname{sgn}^+(S_i)\alpha_i(C_i^*(C) - C_i)^+ + \alpha_i(C_i^*(C) - C_i)^- & \text{in } (0,T)\times\Omega, \\ C_i(0,x) = C_i^0(x) > 0,\ S_i(0,x) = S_i^0(x) > 0 & \text{in } \Omega, \\ C_i(t,x) = 0 & \text{in } (0,T)\times\partial\Omega. \end{cases}$$

The purpose of this article is to present an efficient numerical scheme of order 2 in time to integrate systems such as system (1.1). The scheme proposed in [7] is based on a simple implicit Euler method and has two main drawbacks. First a large nonlinear system has to be solved at each time step and second it is only of order 1 in time. We propose a scheme combining an operator splitting method ([13], [9]) and an event location algorithm using a dense output formula. Operator splitting methods are known to provide cheap and high order approximations to reaction-diffusion equations [2], [10], [6]. Therefore, they represent an interesting tool for dealing with large chemical systems. The event location algorithm presented in Section 3 enables us to determine the switching times at which the discontinuities occur in the reaction terms with a desired accuracy.

Throughout this article we consider a semi-discretized system of equations. Indeed a difficulty appears in the fully continuous case that we are not able to cope with easily. The switching time, t_d, is an unknown function of x, the space variable. Dealing with the continuous case then means considering reaction-diffusion equations defined on a noncylindrical domain. One can bring back the problem to a cylindrical domain by rescaling the time variable but then time and space dependent coefficients with unknown regularity appear in the equations. Thus, we consider that the chemical system is already discretized in space, using, for example, a finite difference or a finite element method. The system of ODEs we consider then reads

$$\begin{cases} \dfrac{d\mathbf{C}}{dt} = \mathbf{A}\mathbf{C} + \mathbf{F}(\mathbf{C},\mathbf{S}), \\ \dfrac{d\mathbf{S}}{dt} = -\mathbf{F}(\mathbf{C},\mathbf{S}), \\ \mathbf{C}(0) = \mathbf{C}_0,\ \mathbf{S}(0) = \mathbf{S}_0. \end{cases} \tag{1.2}$$

$\mathbf{C}$ and $\mathbf{S}$ are vectors of $\mathbb{R}^N$ and $\mathbf{A}$ is the $N \times N$ matrix resulting from the spatial discretization of the Δ operator which is symmetric negative definite. The nonlinear terms read

$$\mathbf{F}(\mathbf{C},\mathbf{S}) = (F_k(\mathbf{C},\mathbf{S}))_{k=0,\dots,N},$$

with

$$F_k(\mathbf{C},\mathbf{S}) = \begin{cases} G_k^1(\mathbf{C}), \text{ if } S_k > 0, \\ G_k^2(\mathbf{C}), \text{ if } S_k \le 0. \end{cases} \tag{1.3}$$

This paper is organized as follows. In Section 2 we present the operator splitting method and show that it can be applied to a system in which PDEs and ODEs are coupled. Estimates of the local errors are given. Section 3 deals with the numerical treatment of the discontinuities in the nonlinear reaction terms. We formalize the event location algorithm suggested in [4] and give estimates of the local errors. We then describe the scheme we propose, combining an operator splitting method and

an adaptation of the event location algorithm, and show it is of order 2. In the final Section 4 the effectiveness of the scheme is illustrated numerically.

2. Operator splitting

The main topic of this section is to present the operator splitting method which constitutes the first ingredient of the scheme we propose. We first concentrate on a classical reaction-diffusion equation and give estimates of the local errors. We then show that the method can still be applied without any order reduction ([3], [12], [5]) if an ODE is coupled to the first equation.

2.1. **Strang operator splitting.** In this section we only consider the semi-discretized diffusion-reaction equation for $\mathbf{C}$. The problem of the switching of the nonlinear discontinuous reaction terms is also left aside. We assume that locally the reaction term $\mathbf{F}(\mathbf{C}, \mathbf{S})$ is given by a smooth function $\mathbf{G}(\mathbf{C})$. As in [1], $\mathbf{G}$ is a Lipschitz function with constant L of class C^∞ such that $\mathbf{G}(0) = 0$ and the first four derivatives of $\mathbf{G}$ are bounded.

Let R^t denote the flow (also called fundamental solution operator) of the system

$$\begin{cases} \dfrac{d\mathbf{C}}{dt} = \mathbf{AC} + \mathbf{G}(\mathbf{C}),\ t > 0, \\ \mathbf{C}(0) = \mathbf{C}_0. \end{cases} \tag{2.1}$$

Let Y^t denote the flow of system (2.2),

$$\begin{cases} \dfrac{d\mathbf{C}_1}{dt} = \mathbf{G}(\mathbf{C}_1),\ t > 0, \\ \mathbf{C}_1(0) = \mathbf{C}_{0,1}, \end{cases} \tag{2.2}$$

and X^t denote the flow of system (2.3),

$$\begin{cases} \dfrac{d\mathbf{C}_2}{dt} = \mathbf{AC}_2,\ t > 0, \\ \mathbf{C}_2(0) = \mathbf{C}_{0,2}. \end{cases} \tag{2.3}$$

The idea of splitting methods is to approximate R^t by combining the two flows X^t and Y^t. Two classical approximations are given by the Strang formulas [13],

$$Z^t_{DRD} = X^{t/2} Y^t X^{t/2} \quad \text{and} \quad Z^t_{RDR} = Y^{t/2} X^t Y^{t/2}$$

(which we also denote by diffusion-reaction-diffusion or DRD-splitting and RDR-splitting in the remaining part of this paper). The following result holds.

Lemma 2.1. *Let $\mathbf{C}_0 \in \mathbb{R}^N$. For t sufficently small, the local errors for the two splitting schemes satisfy,*

$$R^t\mathbf{C}_0 - Z^t_{DRD}\mathbf{C}_0 = O(t^3)$$

$$R^t\mathbf{C}_0 - Z^t_{RDR}\mathbf{C}_0 = O(t^3).$$

Proof. This result is the particular finite dimensional case of local error estimation results obtained by Besse et al. in [1]. It can be derived using the same tools, essentially Taylor expansions and judicious estimations of the rest, with some minor adaptations due to finite dimension. For the sake of completeness we give here the

main ideas of the proof. Let us denote by $||.||$ the euclidean norm on $\mathbb{R}^N$. We may write a Duhamel formula for problem (2.1), which reads

$$R^t\mathbf{C}_0 = X^t\mathbf{C}_0 + \int_0^t X^{t-s}\mathbf{G}(R^s\mathbf{C}_0)ds,$$

and express the difference between the exact solution and the splitting solution $Z^t\mathbf{C}_0$ (DRD or RDR) as

$$R^t\mathbf{C}_0 - Z^t\mathbf{C}_0 = \int_0^t X^{t-s}(\mathbf{G}(R^s\mathbf{C}_0) - \mathbf{G}(Z^s\mathbf{C}_0))ds + \mathbf{W}(t).$$

Since $\mathbf{G}$ is Lipschitz with constant $L > 0$ such that for all $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^N$

$$||\mathbf{G}(\mathbf{C}_1) - \mathbf{G}(\mathbf{C}_2)|| \leq L||\mathbf{C}_1 - \mathbf{C}_2||.$$

The matrix $\mathbf{A}$ is negative definite. Thus for all $\mathbf{V} \in \mathbb{R}^N$ and all $t \geq 0$ the following inequality holds for the semi-group $e^{t\mathbf{A}}$,

$$||X^t\mathbf{V}|| = ||e^{t\mathbf{A}}\mathbf{V}|| \leq ||\mathbf{V}||.$$

It follows that

$$||R^t\mathbf{C}_0 - Z^t\mathbf{C}_0|| \leq L\int_0^t ||\mathbf{G}(R^s\mathbf{C}_0) - \mathbf{G}(Z^s\mathbf{C}_0)||ds + ||\mathbf{W}(t)||.$$

Then the estimates of Lemma 2.1 are obtained in the same way as in the proof of Lemma 3.1 p.13 of [1] by accounting for the following changes. In Lemma 2.2.1 p.4 from [1] the L^2, H^2, and H^4 norms are replaced by the Euclidean norm $\|\cdot\|$, the A-norm $\|\cdot\|_A$=$\|A\cdot\|$ and the A^2-norm $\|\cdot\|_{A^2}$=$\|A^2\cdot\|$, respectively . Then by using the Gronwall Lemma (p.3 from [1]), the rest $W(t)$ is estimated as $||W(t)|| = O(t^3)$ for t small and Lemma 2.1 reduces to Lemma 3.1 from [1] with the norms previously introduced. □

2.2. **Coupling reaction-diffusion equations and ODEs.** Let us now consider the coupling of equation (2.1) with the equation for $\mathbf{S}$:

$$\left\{\begin{array}{l} \dfrac{d\mathbf{C}}{dt} = \mathbf{A}\mathbf{C} + \mathbf{G}(\mathbf{C}), \\ \dfrac{d\mathbf{S}}{dt} = -\mathbf{G}(\mathbf{C}), \\ \mathbf{C}(0) = \mathbf{C}_0,\ \mathbf{S}(0) = \mathbf{S}_0. \end{array}\right. \tag{2.4}$$

The solution $(\mathbf{C}(t), \mathbf{S}(t))$ to (2.4) is denoted by $R^t(\mathbf{C}_0, \mathbf{S}_0) = (R_C^t\mathbf{C}_0, R_S^t(\mathbf{C}_0, \mathbf{S}_0))$. $\mathbf{S}(t)$ is given explicitly by

$$\mathbf{S}(t) = \mathbf{S}_0 - \int_0^t \mathbf{G}(\mathbf{C}(s))ds,$$

which can also be written $\mathbf{S} = \mathbf{H}(\mathbf{C})$. In such a situation Descombes and Massot [3] show that order reduction occurs in the DRD-splitting but not in the RDR-splitting. The problem we consider is quite similar. However, because of the particular form of function $\mathbf{H}$ which should be written as

$$\mathbf{S}(t) = \mathbf{H}(t, \mathbf{S}_0, \mathbf{C}(.)),$$

no order reduction occurs as is shown in Lemma 2.2. Let us denote by

$$(\mathbf{C}_1(t), \mathbf{S}_1(t)) = (Y_C^t\mathbf{C}_{0,1}, \mathbf{H}(t, \mathbf{S}_{0,1}, Y_C\mathbf{C}_{0,1}))$$

the solution to system

$$\left\{\begin{array}{l} \dfrac{d\mathbf{C}_1}{dt} = \mathbf{G}(\mathbf{C}_1), \\ \dfrac{d\mathbf{S}_1}{dt} = -\mathbf{G}(\mathbf{C}_1), \\ \mathbf{C}_1(0) = \mathbf{C}_{0,1},\ \mathbf{S}_1(0) = \mathbf{S}_{0,1}. \end{array}\right. \tag{2.5}$$

The DRD-splitting for system (2.4) can be written as

$$Z^t_{DRD}(\mathbf{C}_0, \mathbf{S}_0) = (Z^t_{DRD,C}\mathbf{C}_0, Z^t_{DRD,S}(\mathbf{C}_0, \mathbf{S}_0)),$$

with

$$Z^t_{DRD,C}\mathbf{C}_0 = X^{t/2}Y^t_C X^{t/2}\mathbf{C}_0, \tag{2.6}$$

$$Z^t_{DRD,S}(\mathbf{C}_0, \mathbf{S}_0) = \mathbf{H}(t, \mathbf{S}_0, Y_{\dot{C}}X^{t/2}\mathbf{C}_0), \tag{2.7}$$

and the RDR-splitting as

$$Z^t_{RDR}(\mathbf{C}_0, \mathbf{S}_0) = (Z^t_{RDR,C}\mathbf{C}_0, Z^t_{RDR,S}(\mathbf{C}_0, \mathbf{S}_0)), \tag{2.8}$$

with

$$Z^t_{RDR,C}\mathbf{C}_0 = Y^{t/2}_C X^t Y^{t/2}_C \mathbf{C}_0, \tag{2.9}$$

$$Z^t_{RDR,S}(\mathbf{C}_0, \mathbf{S}_0) = \mathbf{H}(t/2, \mathbf{H}(t/2, \mathbf{S}_0, Y_{\dot{C}}\mathbf{C}_0), Y_{\dot{C}}X^t Y^{t/2}_C \mathbf{C}_0)). \tag{2.10}$$

The following result holds.

Lemma 2.2. *Let $\mathbf{C}_0 \in \mathbb{R}^N$ and $\mathbf{S}_0 \in \mathbb{R}^N$. For t sufficiently small, the local errors for the two splitting schemes satisfy*

$$R^t_C\mathbf{C}_0 - Z^t_{DRD,C}\mathbf{C}_0 = O(t^3), \tag{2.11}$$

$$R^t_S(\mathbf{C}_0, \mathbf{S}_0) - Z^t_{DRD,S}(\mathbf{C}_0, \mathbf{S}_0) = O(t^3), \tag{2.12}$$

$$R^t_C\mathbf{C}_0 - Z^t_{RDR,C}\mathbf{C}_0 = O(t^3), \tag{2.13}$$

$$R^t_S(\mathbf{C}_0, \mathbf{S}_0) - Z^t_{RDR,S}(\mathbf{C}_0, \mathbf{S}_0) = O(t^3). \tag{2.14}$$

Proof. Equations (2.11) and (2.13) follow directly from Lemma 2.1. Let us show (2.12) and (2.14). Using the Duhamel formula, $\mathbf{C}(t)$ is given explicitly by

$$\mathbf{C}(t) = e^{t\mathbf{A}}\mathbf{C}_0 + \int_0^t e^{(t-s)\mathbf{A}}\mathbf{G}(\mathbf{C}(s))ds.$$

It follows from classical expansions that

$$\mathbf{C}(t) = \mathbf{C}_0 + t(\mathbf{A}\mathbf{C}_0 + \mathbf{G}(\mathbf{C}_0)) + O(t^2), \tag{2.15}$$

and since $\mathbf{S}(t) = \mathbf{H}(t, \mathbf{S}_0, \mathbf{C}(.))$, we obtain

$$\mathbf{S}(t) = \mathbf{S}_0 - t\mathbf{G}(\mathbf{C}_0) - \frac{t^2}{2}\mathbf{G}'(\mathbf{C}_0)(\mathbf{A}\mathbf{C}_0 + \mathbf{G}(\mathbf{C}_0)) + O(t^3). \tag{2.16}$$

From

$$Y^t_C\mathbf{C}_0 = \mathbf{C}_0 + t\mathbf{G}(\mathbf{C}_0) + O(t^2)$$

and

$$X^t\mathbf{C}_0 = e^{t\mathbf{A}}\mathbf{C}_0 = \mathbf{C}_0 + t\mathbf{A}\mathbf{C}_0 + O(t^2),$$

we deduce that

$$Y^s_C X^{t/2\mathbf{A}}\mathbf{C}_0 = \mathbf{C}_0 + \frac{t}{2}\mathbf{A}\mathbf{C}_0 + s\mathbf{G}(\mathbf{C}_0) + O(t^2) + O(s^2) + O(st)$$

and

$$\mathbf{G}(Y_C^s X^{t/2\mathbf{A}}\mathbf{C}_0) = \mathbf{G}(\mathbf{C}_0)+\frac{t}{2}\mathbf{G}'(\mathbf{C}_0)\mathbf{A}\mathbf{C}_0+s\mathbf{G}'(\mathbf{C}_0)\mathbf{G}(\mathbf{C}_0)+O(t^2)+O(s^2)+O(st).$$

Eventually since

$$Z^t_{DRD,S}(\mathbf{C}_0,\mathbf{S}_0) = \mathbf{H}(t,\mathbf{S}_0,Y_C X^{t/2}\mathbf{C}_0),$$

we obtain

$$\mathbf{S}(t) - Z^t_{DRD,S}(\mathbf{C}_0,\mathbf{S}_0) = O(t^3),$$

which proves (2.12).

The same type of arguments are used to prove (2.14). □

3. Numerical treatment of the discontinuities

The purpose of this section is twofold. First we present the event location algorihtm for a discontinuous ODE suggested in Hairer et al. [4] and prove that indeed it leads to an accurate numerical method. We then combine this algorithm to a splitting scheme and obtain a method of order 2 to integrate system (1.2).

3.1. **An event location algorithm for ODEs.**

In this section we present a numerical scheme of order $p \geq 2$ to solve a nonlinear discontinuous ODE. The main numerical tool used is an explicit Runge-Kutta method of order p with a dense output of order $p^* \geq 2$. The reader is referred to Sections II-1 to II-6 of the book by Hairer et al. [4] for a detailed description of these methods. We assume here that $p = p^*$.

Let us give some notation. An explicit Runge-Kutta method of order p to solve the ordinary differential equation

$$\left\{\begin{array}{l} y' = f(t,y), \\ y(t_0) = y_0, \end{array}\right. \tag{3.1}$$

is represented by the increment function of the method, $F(t,y,h)$. Given an initial value (t_0,y_0) and a step size h, one computes a numerical solution y_1 approximating $y(t_0+h)$ by $y_1 = y_0 + hF(t_0,y_0,h)$. The numerical solution for a point $T > t_0$ is then obtained by a step-by-step procedure

$$y_{i+1} = y_i + hF(t_i,y_i,h).$$

If the method is of order p, then the local error

$$e_{i+1} = y(t_i+h) - (y(t_i) + hF(t_i,y(t_i),h)),$$

statisfies

$$e_{i+1} = O(h^{p+1}). \tag{3.2}$$

A Runge-Kutta method with a dense output formula provides a cheap numerical approximation to $y(t_i+\theta h)$ for the whole integration interval $0 \leq \theta \leq 1$. We denote this approximation by $u_i(\theta)$, and we have

$$u_i(\theta) = y(t_i+\theta h) + O(h^{p+1}). \tag{3.3}$$

Let us now concentrate on the numerical integration on a time interval $[t_0,T]$ of the following ordinary differential equation:

$$\left\{\begin{array}{l} y' = f_1(t,y) \text{ if } g(y) > 0, \\ y' = f_2(t,y) \text{ if } g(y) \leq 0, \\ y(t_0) = y_0 \text{ and } g(y_0) > 0. \end{array}\right. \tag{3.4}$$

We assume that f_1, f_2 and g are C^∞ functions. The function g is called the switching function. We also assume that the solution $y(t)$ to (3.4) crosses the surface $\Sigma = \{y;\ g(y) = 0\}$ only once, at the point $y_d = y(t_d)$. Therefore, $y(t)$ may be written as

$$\begin{cases} y(t) = y_1(t) \text{ in } [t_0, t_d[, \\ y(t) = y_2(t) \text{ in } [t_d, T], \\ y_1(t_d) = y_2(t_d) = y_d, \end{cases}$$

where y_1 is the solution to

$$\begin{cases} y'(t) = f_1(t, y),\ t > t_0, \\ y(t_0) = y_0, \end{cases}$$

and y_2 is the solution to

$$\begin{cases} y'(t) = f_2(t, y),\ t > t_d, \\ y(t_d) = y_d. \end{cases}$$

The derivative of the solution y is, in general, discontinuous on Σ. The difficulty in the numerical integration of such a discontinuous equation is that the point (t_d, y_d) is not known in advance but has to be detected. Moreover, in order to obtain a method of order p, this point has to be detected with a precision of order p. The method proposed here relies on the event location algorithm suggested in the book by Hairer et al. [4] (Algorithm 6.4 page 195).

Algorithm 3.1.

- *Using f_1, define a Runge-Kutta method of order p with increment function F_1,*
$$y_{i+1} = y_i + hF_1(t_i, y_i, h).$$
- *Compute the solution step-by-step $y_0, y_1, \ldots$ until a sign change appears between $g(y_{n-1})$ and $g(y_n)$.*
- *Using the dense output, find θ such that $g(u_{n-1}(\theta)) = 0$.*
- *Reset $y_n = u_{n-1}(\theta)$ and $t_n = t_{n-1} + \theta h$.*
- *Using f_2, define a Runge-Kutta method of order p with increment function F_2,*
$$y_{i+1} = y_i + hF_2(t_i, y_i, h),$$
and carry on the computation from t_n to $t_N = T$.

The key point in this algorithm is that, thanks to the dense output, we are able to compute $y_n = y_d + O(h^{p+1})$ and $t_n = t_d + O(h^{p+1})$. Thus, we can show the following technical result.

Lemma 3.1. *At each time step of the scheme provided by Algorithm* 3.1*, the local error satisfies*

$$e_i = O(h^{p+1}).$$

Proof. From (3.2) it is clear that for $i = 1$ to $n - 1$,

$$e_i = y_1(t_i) - (y_1(t_{i-1}) + hF_1(t_{i-1}, y_1(t_{i-1}), h)) = O(h^{p+1}),$$

and that for $i = n + 2$ to N,

$$e_i = y_2(t_i) - (y_2(t_{i-1}) + hF_2(t_{i-1}, y_2(t_{i-1}), h)) = O(h^{p+1}).$$

It remains to show the result for e_n and e_{n+1}. Since we only know that $t_n = t_d + O(h^{p+1})$, there are two cases.

- Case $t_{n-1} < t_d \leq t_n < t_{n+1}$:
  The local error e_n reads

$$\begin{aligned} e_n &= y_2(t_n) - (y_1(t_{n-1}) + hF_1(t_{n-1}, y_1(t_{n-1}), h)) \\ &= [y_1(t_n) - (y_1(t_{n-1}) + hF_1(t_{n-1}, y_1(t_{n-1}), h))] + [y_2(t_n) - y_1(t_n)]. \end{aligned}$$

  Moreover, we have $y_1(t_n) = y_d + O(h^{p+1})$, $y_2(t_n) = y_d + O(h^{p+1})$, and we can conclude that $e_n = O(h^{p+1})$.
  Concerning the next step it is also clear that $e_{n+1} = O(h^{p+1})$.
- Case $t_{n-1} < t_n \leq t_d < t_{n+1}$:
  It is clear that $e_n = O(h^{p+1})$. The local error e_{n+1} reads

$$e_{n+1} = y_2(t_{n+1}) - (y_1(t_n) + hF_2(t_n, y_1(t_n), h)).$$

  Since $y_1(t_n) = y_d + O(h^{p+1})$, $t_n = t_d + O(h^{p+1})$, and $h = (t_{n+1} - t_d) + O(h^{p+1})$, we have that

$$hF_2(t_n, y_1(t_n), h) = (t_{n+1} - t_d)F_2(t_d, y_d, (t_{n+1} - t_d)) + O(h^{p+1})$$

  and

$$\begin{aligned} e_{n+1} &= y_2(t_{n+1}) - (y_d + (t_{n+1} - t_d)F_2(t_d, y_d, (t_{n+1} - t_d))) + O(h^{p+1}) \\ &= O((t_{n+1} - t_d)^{p+1}) + O(h^{p+1}). \end{aligned}$$

  Therefore, $e_{n+1} = O(h^{p+1})$.

□

The third step of Algorithm 3.1 is crucial. The computation of θ, such that $g(u_{n-1}(\theta)) = 0$, can be done using a dichotomy method or, for example, a second order Muller method. This latter requires that the zeros of g are separated and might require many iterations to converge depending on the "flatness" of g between t_{n-1} and t_n. However, for the applications we considered the desired accuracy on θ, y_n, and t_n can always be achieved.

3.2. **Combining the event location algorithm and the splitting scheme.** In this section we formulate the scheme proposed to integrate system (1.2). The method combines either the RDR-splitting or the DRD-splitting described in Section 2 and an event location algorithm similar to Algorithm 3.1 of Section 3.1. With those tools we construct a numerical scheme of order 2 in time for system (1.2).

Since the discontinuous nonlinear reaction terms only come up in the R-stages, it is tempting to try to detect the switching times only during these stages. However, this is not possible since the intermediate $\mathbf{C}$ or $\mathbf{S}$ values computed after the first two stages of the splitting scheme are not yet in $O(h^3)$. The computed switching time, therefore, cannot be an $O(h^3)$ approximation of the exact switching time, and we need to construct a dense output for a whole time step including the three stages of the splitting method. Hermite interpolation (Shampine [11]) provides an efficient way to construct dense output formulas. Whatever the splitting is, at each time step we have two function values $\mathbf{S}^0$, $\mathbf{S}^1$ and two derivatives $\frac{d\mathbf{S}^0}{dt}$, $\frac{d\mathbf{S}^1}{dt}$ at our disposal and can thus do cubic polynomial interpolation. The resulting formula is

$$\mathbf{u}^S(\theta) = (1-\theta)\mathbf{S}^0 + \theta\mathbf{S}^1 + \theta(\theta-1)\left((1-2\theta)(\mathbf{S}^1 - \mathbf{S}^0) + (\theta-1)h\frac{d\mathbf{S}^0}{dt} + \theta h\frac{d\mathbf{S}^1}{dt}\right).$$

A similar formula, $\mathbf{u}^C$ can be computed for $\mathbf{C}$. Since the splitting is of order 2, we have

$$\mathbf{u}^S(\theta) - \mathbf{S}(t_0 + \theta h) = O(h^3),$$

$$\mathbf{u}^C(\theta) - \mathbf{C}(t_0 + \theta h) = O(h^3).$$

These dense output formulas are used to detect the switching times. This detection is performed component by component, as illustrated in the following algorithm.

Algorithm 3.2.

- *Start from* $\mathbf{C}^0 > 0$, $\mathbf{S}^0 > 0$ *and thus with* $\mathbf{F}(\mathbf{C}, \mathbf{S}) = (G^1_k(\mathbf{C}))_{k=0,\dots,N}$.
- *Using either the RDR-splitting or the DRD-splitting, compute the solution step-by-step* $(\mathbf{C}^0, \mathbf{S}^0), (\mathbf{C}^1, \mathbf{S}^1), \dots$ *until a sign change appears, for a component* k_1, *between* $\mathbf{S}^{n-1}_{k_1}$ *and* $\mathbf{S}^n_{k_1}$.
- *Using the dense output polynomial* $\mathbf{u}^S$, *find* θ *such that* $u^S_{k_1}(\theta) = 0$.
- *Reset* $t_n = t_{n-1} + \theta h$, $\mathbf{S}^n = \mathbf{u}^S(\theta)$, *and* $\mathbf{C}^n = \mathbf{u}^C(\theta)$.
- *Change* $G^1_{k_1}$ *to* $G^2_{k_1}$ *and carry on the computation using the new reaction term* $\mathbf{F}$ *until a new sign change appears for another component* S_{k_2}.

We denote by

$$(\mathbf{C}^{n+1}, \mathbf{S}^{n+1}) = Z^h(\mathbf{C}^n, \mathbf{S}^n) = (Z^h_C \mathbf{C}^n, Z^h_S(\mathbf{C}^n, \mathbf{S}^n))$$

the numerical scheme provided by Algorithm 3.2. Let us now state a result concerning the estimation of the local errors.

Lemma 3.2. *At each time step of the scheme provided by Algorithm* 3.2*, the local error satisfies*

$$e^C_i = R^h_C \mathbf{C}(t_i) - Z^h_C \mathbf{C}(t_i) = O(h^3), \tag{3.5}$$

$$e^S_i = R^h_S(\mathbf{C}(t_i), \mathbf{S}(t_i)) - Z^h_S(\mathbf{C}(t_i), \mathbf{S}(t_i)) = O(h^3). \tag{3.6}$$

Proof. We restrict ourselves to a time interval $[0, T]$ on which only one component, S_{k_1}, switches at time t_d. Other switchings can be treated in the same way. The exact solution, $(\mathbf{C}(t), \mathbf{S}(t)) = R^t(\mathbf{C}_0, \mathbf{S}_0)$ may be written as

$$\begin{cases} (\mathbf{C}(t), \mathbf{S}(t)) = (\mathbf{C}_1(t), \mathbf{S}_1(t)) = R^t_1(\mathbf{C}_0, \mathbf{S}_0) = (R^t_{1C}\mathbf{C}_0, R^t_{1S}(\mathbf{C}_0, \mathbf{S}_0)), \text{ in } [0, t_d[, \\ (\mathbf{C}(t), \mathbf{S}(t)) = (\mathbf{C}_2(t), \mathbf{S}_2(t)) = R^t_2(\mathbf{C}_d, \mathbf{S}_d) = (R^t_{2C}\mathbf{C}_d, R^t_{2S}(\mathbf{C}_d, \mathbf{S}_d)), \text{ in } [t_d, T], \\ (\mathbf{C}_1(t_d), \mathbf{S}_1(t_d)) = (\mathbf{C}_2(t_d), \mathbf{S}_2(t_d)) = (\mathbf{C}_d, \mathbf{S}_d), \end{cases}$$

where $(\mathbf{C}_1(t), \mathbf{S}_1(t))$ is the solution to

$$\begin{cases} \dfrac{d\mathbf{C}}{dt} = \mathbf{A}\mathbf{C} + \mathbf{F}^1(\mathbf{C}), \ t > 0, \\ \dfrac{d\mathbf{S}}{dt} = -\mathbf{F}^1(\mathbf{C}), \\ \mathbf{C}(0) = \mathbf{C}_0, \ \mathbf{S}(0) = \mathbf{S}_0, \end{cases}$$

and $(\mathbf{C}_2(t), \mathbf{S}_2(t))$ is the solution to

$$\begin{cases} \dfrac{d\mathbf{C}}{dt} = \mathbf{A}\mathbf{C} + \mathbf{F}^2(\mathbf{C}), \ t > t_d, \\ \dfrac{d\mathbf{S}}{dt} = -\mathbf{F}^2(\mathbf{C}), \\ \mathbf{C}(t_d) = \mathbf{C}_d, \ \mathbf{S}(t_d) = \mathbf{S}_d, \end{cases}$$

where

$$\mathbf{F}^1(\mathbf{C}) = (G^1_k(\mathbf{C}))_{k=0,\dots,N},$$

and

$$\mathbf{F}^2(\mathbf{C}) = (G^1_1(\mathbf{C}), \dots, G^1_{k_1-1}(\mathbf{C}), G^2_{k_1}(\mathbf{C}), G^1_{k_1+1}(\mathbf{C}), \dots).$$

Before the switching, (2.13) and (2.14) (or (2.11) and (2.12)) directly show that for $i = 1$ to $n-1$,

$$e_i^C = O(h^3), \qquad e_i^S = O(h^3).$$

In the same way after the switching time we have for $i = n+2$ to N,

$$e_i^C = O(h^3), \qquad e_i^S = O(h^3).$$

It remains to show that e_n^C, e_{n+1}^C, e_n^S, and e_{n+1}^S are $O(h^3)$. Since we only know that $t_n = t_d + O(h^3)$ there are two cases.

- Case $t_{n-1} < t_d \le t_n < t_{n+1}$:

  The local error e_n^C reads

$$\begin{aligned} e_n^C &= R_C^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1}) - Z_C^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1}) \\ &= (R_{1C}^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1}) - Z_C^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1})) \\ &\quad + (R_C^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1}) - R_{1C}^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1})). \end{aligned}$$

  Again, from (2.13) we know that

$$(R_{1C}^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1}) - Z_C^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1})) = O(h^3).$$

  Moreover, since

$$\begin{aligned} &(R_C^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1}) - R_{1C}^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1})) \\ &\quad = (R_{2C}^{(t_n - t_d)} - R_{1C}^{(t_n - t_d)})R_{1C}^{(t_d - t_{n-1})}\mathbf{C}(t_{n-1}) \\ &\quad = (R_{2C}^{(t_n - t_d)} - R_{1C}^{(t_n - t_d)})\mathbf{C}_d, \end{aligned}$$

  we obtain using the same expansion as in (2.15),

$$(R_C^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1}) - R_{1C}^{(t_n - t_{n-1})}\mathbf{C}(t_{n-1})) = O(t_n - t_d) = O(h^3),$$

  and this proves that

$$e_n^C = O(h^3).$$

  The same type of manipulations enable us to show that

$$e_n^S = O(h^3).$$

  It is also clear that $e_{n+1}^C = O(h^3)$, and that $e_{n+1}^S = O(h^3)$.

- Case $t_{n-1} < t_n \le t_d < t_{n+1}$:

  The arguments of the proof are similar to those of the previous case.

□

4. Global error estimate and numerical illustration

Theorem 4.1. *For all $\mathbf{C}_0$, $\mathbf{S}_0 \in \mathbb{R}^N$ and all $T > 0$, there exists h_0 such that for all $h \in]0, h_0]$, for all n such that $nh \leq T$*

$$R_C^{nh}\mathbf{C}_0 - (Z_C^h)^n\mathbf{C}_0 = O(h^2), \tag{4.1}$$

$$R_S^{nh}(\mathbf{C}_0, \mathbf{S}_0) - (Z_S^h)^n(\mathbf{C}_0, \mathbf{S}_0) = O(h^2). \tag{4.2}$$

Proof. We only prove (4.1). As noticed in [2] the triangle inequality yields

$$||(Z_C^h)^n\mathbf{C}_0 - R_C^{nh}\mathbf{C}_0|| \leq \sum_{j=0}^{n-1} ||(Z_C^h)^{n-1-j}(Z_C^h)R_C^{jh}\mathbf{C}_0 - (Z_C^h)^{n-1-j}R_C^{(j+1)h}\mathbf{C}_0||.$$

By using the fact that X^t is unitary with respect to the Euclidean norm and that the functions G^i are Lipchitzian with constant L, we refer to [1, p.8] where, for deriving, there exists a constant K such that for $\mathbf{C}_1$ and $\mathbf{C}_2 \in \mathbb{R}^N$ and all $t \in [0, 1]$

$$||Z_C^t\mathbf{C}_1 - Z_C^t\mathbf{C}_2|| \leq (1 + Kt)||\mathbf{C}_1 - \mathbf{C}_2||.$$

Therefore,

$$||(Z_C^h)^n\mathbf{C}_0 - R_C^{nh}\mathbf{C}_0|| \leq \sum_{j=0}^{n-1}(1 + Kh)^{n-1-j}||(Z_C^h - R_C^h)R_C^{jh}\mathbf{C}_0||.$$

Now from Lemma 3.2 we deduce that there exists a constant $\tilde{K}$ such that for all j such that $jh \leq T$

$$||(Z_C^h - R_C^h)R_C^{jh}\mathbf{C}_0|| \leq \tilde{K}h^3,$$

and eventually

$$\begin{aligned} ||(Z_C^h)^n\mathbf{C}_0 - R_C^{nh}\mathbf{C}_0|| &\leq \tilde{K}\sum_{j=0}^{n-1} e^{(n-1-j)Kh}h^3, \\ &\leq \tilde{K}e^{nhK}(nh)h^2, \\ &\leq \tilde{K}(T)h^2. \end{aligned}$$

□

Let us illustrate this result by a numerical experiment with a simple test case. We consider the following system of equations set on the one dimensional domain $(0, 1)$,

$$\left\{\begin{array}{lll} \partial_t C = \Delta C + \alpha C(1-C) & \text{if } S > S_d, \\ \quad\;\; = \Delta C + \beta C & \text{if } S \leq S_d, \\ \partial_t S = -\alpha C(1-C) & \text{if } S > S_d, \\ \quad\;\; = -\beta C & \text{if } S \leq S_d, \end{array}\right. \tag{4.3}$$

where $\alpha = 0.5, \beta = 0.25$, and $S_d = 1$ are constants. Initial and boundary conditions for C are determined by the exact solution

$$C(t, x) = \left(\frac{1}{1 + \exp(\sqrt{\frac{\alpha}{6}}x - \frac{5}{6}\alpha t)}\right)^2$$

to Fisher's equation

$$\partial_t C = \Delta C + \alpha C(1 - C).$$

Hence,

$$C(0,x) = (\frac{1}{1+\exp(\sqrt{\frac{\alpha}{6}}x)})^2,$$

$$C(t,0) = (\frac{1}{1+\exp(-\frac{5}{6}\alpha t)})^2,$$

and

$$C(t,1) = (\frac{1}{1+\exp(\sqrt{\frac{\alpha}{6}} - \frac{5}{6}\alpha t)})^2.$$

Initial conditions for S are given by

$$S(0,x) = 1 + \exp(-(x-1/2)^2).$$

The diffusion operator is discretized using second order finite differences with a step size $\Delta x = 10^{-2}$, and its time integration is performed using the unconditionally stable second order Crank-Nicolson scheme. Reaction terms are integrated with a second order explicit Runge-Kutta scheme. A reference solution is computed for the classical splitting method and for the method proposed in this paper with a time step $h_{\text{ref}} = \frac{0.1}{2^{14}}$.

Figure 1 shows a zoom in on $S(t,x)$ where the discontinuity of the derivative $\partial_t S$ clearly appears when S crosses the surface $S = S_d = 1$.

Solutions are computed using five different time steps, $h = \frac{0.1}{2^9}, \frac{0.1}{2^{10}}, \frac{0.1}{2^{11}}, \frac{0.1}{2^{12}}$ and $h = \frac{0.1}{2^{13}}$. For each solution the global errors

$$E_C = ||\mathbf{C}_h(T) - \mathbf{C}_{h_{\text{ref}}}(T)||,$$

$$E_S = ||\mathbf{S}_h(T) - \mathbf{S}_{h_{\text{ref}}}(T)||,$$

are computed at $T = 0.1$ ($||.||$ denoted the euclidian norm). Figure 2 shows $-\log(E_C)$ and $-\log(E_S)$ versus $-\log(h)$ when the classical splitting method is used to compute the solution to problem (4.3). The convergence curve is very

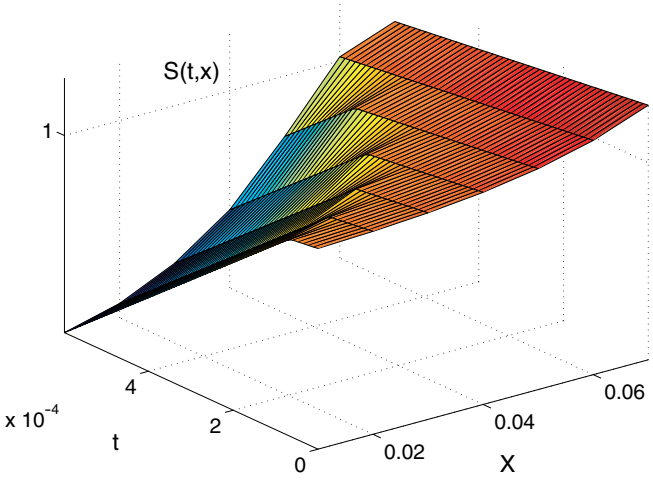

FIGURE 1. Zoom in on $S(t,x)$ crossing the surface $S = 1$, computed with the proposed scheme.

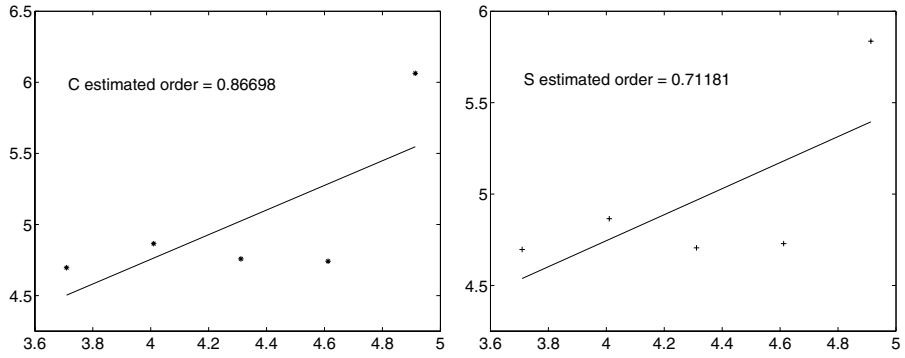

FIGURE 2. $-\log(E)$ versus $-\log(h)$. Convergence curve for the classical splitting (left C and right S).

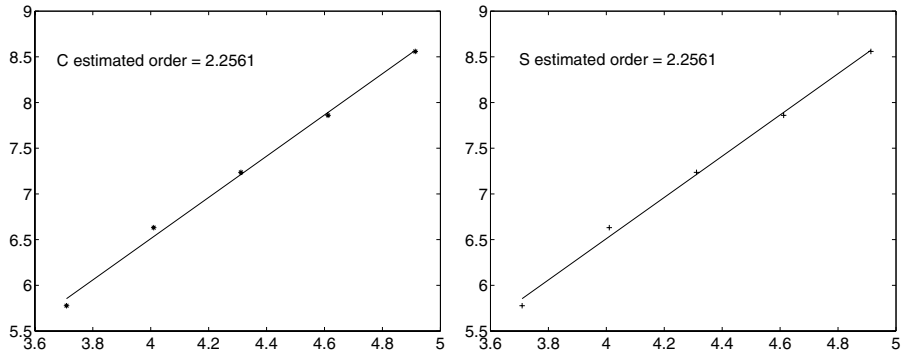

FIGURE 3. $-\log(E)$ versus $-\log(h)$. Convergence curve for the proposed scheme (left C and right S).

perturbed and the estimated order of the scheme is less then 1. This is not surprising since the method is not able to deal with the discontinuities correctly. On the other hand Figure 3 shows $-\log(E_C)$ and $-\log(E_S)$ versus $-\log(h)$ when the method proposed in this paper is used. The estimated order is about 2, which is in agreement with the theoretical result.

REFERENCES

1. C. Besse, B. Bidegaray, and S. Descombes, *Order estimates in time of splitting methods for the nonlinear schrödinger equation*, SIAM J. Numer. Anal. **40** (2002), no. 5, 26–40. MR1921908 (2003k:65099)
2. S. Descombes, *Convergence of a splitting method of high order for reaction-diffusion systems*, Math. Comp. **70** (2001), no. 236, 1484–1501. MR1836914 (2002c:65152)
3. S. Descombes and M. Massot, *Operator splitting for nonlinear reaction-diffusion systems with an entropic structure : singular perturbation and order reduction*, Numerische Mathematik (2004) Vol. 97, No. 4, 667–698.
4. E. Hairer, S.P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer Series in Computational Mathematics, Springer Verlag, 1993. MR1227985 (94c:65005)

5. E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, Springer Series in Computational Mathematics, Springer Verlag, 1993. MR1111480 (92a:65016)
6. C. Lubich and T. Jahnke, *Error bounds for exponential operator splitting*, Technical report, Universitat Tubingen, Germany, 1999; BIT **40** (2000), 735–744. MR1799313 (2001k:65143)
7. E. Maisse, *Analyse et simulations numériques de phénomènes de diffusion-dissolution/précipitation en milieux poreux, appliquées au stockage de déchets*, Ph.D. thesis, Université Claude Bernard Lyon I, 1998.
8. E. Maisse and J. Pousin, *Diffusion and dissolution/precipitation in an open porous reactive medium*, J. Comp. Appl. Math. **82** (1997), 279–280. MR1473546 (98g:35170)
9. G.I. Marchuk, *Splitting and alternating direction methods*, Handbook of numerical analysis, vol. I, North-Holland, Amsterdam, 1990, pp. 197–462. MR1039325
10. M. Schatzman, *Toward non commutative numerical analysis : high order integration in time*, Journal of Scientific Computing **17** (2002), no. 1-3, 107–125. MR1910554
11. L.F. Shampine, *Interpolation for Runge-Kutta methods*, SIAM J. Numer. Anal. **22** (1985), 1014–1027. MR0799125 (86j:65014)
12. B. Sportisse, *An analysis of operator splitting techniques in the stiff case*, J. Comput. Phys. **161** (2000), no. 1, 140–168. MR1762076 (2001f:65107)
13. G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal. **5** (1968), 506–517. MR0235754 (38:4057)

CNRS I35, Les Algorithmes, 2000 Route des Lucioles, BP 121, 06903 Sophia Antipolis cedex France
E-mail address: Blaise.Faugeras@unice.fr

MAPLY, Centre de Mathématique INSA de Lyon, Bat. Léonard de Vinci, 21, Av. Jean Capelle, 69100 Villeurbanne Cedex, France
E-mail address: Jerome.Pousin@insa-lyon.fr

MAPLY, Centre de Mathématique INSA de Lyon, Bat. Léonard de Vinci, 21, Av. Jean Capelle, 69100 Villeurbanne Cedex, France
E-mail address: Franck.Fontvieille@insa-lyon.fr

Article M : [20] B. FAUGERAS et J. POUSIN. Variational asymptotic derivation of an elastic model arising from the problem of 3D automatic segmentation of cardiac images. *Analysis and Applications (AA)* 2.4 (2004), p. 275–307

Analysis and Applications, Vol. 2, No. 4 (2004) 275–307

World Scientific
www.worldscientific.com

VARIATIONAL ASYMPTOTIC DERIVATION OF AN ELASTIC MODEL ARISING FROM THE PROBLEM OF 3D AUTOMATIC SEGMENTATION OF CARDIAC IMAGES

BLAISE FAUGERAS*

Institut de Recherche pour le Développement
Centre de Recherche IRD-IFREMER
Av. Jean Monnet, BP 171
34200 Sète, France
Blaise.Faugeras@ifremer.fr

JÉRÔME POUSIN

MAPLY, Centre de Mathématique INSA de Lyon
Bat. Léonard de Vinci, 21, Av. Jean Capelle
69100 Villeurbanne Cedex, France

Received 3 April 2003
Revised 18 November 2003

Segmentation of 3D cardiac images using a deformable elastic model of the heart proved to be significantly improved by applying special boundary conditions on the elastic model [15]. The purpose of this paper is to derive those boundary conditions by means of a rigourous convergence result. We consider a simplified two-layer elastic shell model and show that when the thickness ε of the thin external fibrous layer tends to 0 it can be replaced by the above mentioned boundary conditions on the internal layer. A mixed variational formulation of the problem in curvilinear coordinates is introduced. This formulation is then scaled in order to be defined over an ε-independent domain. Finally, several *a priori* estimations on the solution are obtained which enable us to pass to the limit and prove our result.

Keywords: Segmentation; cardiac; elasticity; shell; mixed variational formulation; asymptotic analysis.

Mathematics Subject Classification 2000: 74B05, 74K25, 35B40, 35B45

1. Introduction

By means of Magnetic Resonance, one can get a clinical M.R. volume dataset. Such a volume dataset is denoted by a matrix V with X rows, Y columns and Z slices which represents a discrete grid of volume elements (or voxels)

*Corresponding author.

$v \in \{1,\dots,X\} \times \{1,\dots,Y\} \times \{1,\dots,Z\}$. For each voxel v, we denote by $I\colon \mathbb{N}^3 \to \mathbb{Z}$ the grey level function $v \mapsto I(v)$. Data are anisotropic with equal sampling in the x and y directions but a coarser density in the z direction. By image segmentation we refer to processes identifying all voxels which belong together according to a homogeneity criterion (most often a grey level criterion). Segmentation is required for the identification of the object (that is, the heart) in the M.R. volume data. Here, we deal with edge-based algorithms which try to detect the borderline of a structure (that is, the discontinuity surfaces of the "gradient" of the grey level function I). A force field is computed from the "gradient" of the function I by using a Gradient Vector Flow technique.

In order to address the problem of 3D automatic segmentation of cardiac M.R. multi-slices image sequences, a strategy based on an elastic simulation of the human heart has been proposed by Vincent *et al.* [16]. It can be summarized as follows: an *a priori* template (object) representing the heart is immersed into the image data and submitted to a force field which pulls the boundary of the object towards the image edges. This method has several advantages but one drawback concerns the regularity of the displacement field and the smoothness of the final object boundary. As an alternative to classical geometrical curvature-based boundary regularization techniques, Pham *et al.* [15] propose to add boundary constraints modeling crudely some biomechanical properties of the heart. They consider a simplified three-layer elastic model of the heart composed of a middle homogeneous isotropic layer and two surrounding thin layers of myocardial fibers with a directional structure. The aim of this model is to mimic the elastic properties of the heart resulting from the fiber structure of the muscle oriented in the longitudinal direction. It is an efficient tool for image segmentation but not a complete myocardium model. For a more realistic elastic model of the heart we refer to Caillerie *et al.* [6]. It is announced but not proved in [15] that the fibrous layers can be replaced by boundary conditions on the middle layer when the thickness of the external layers tends to 0. These conditions increase the stiffness of the boundary and smooth the displacement field at the interface of the elastic object by imposing preferential directions of deformation in the tangent space (see Fig. 1). We are not going here to get into the details of the numerical method used and refer the readers to Pham [14] and Pebay *et al.* [13]. However, it is worth noticing that the use of a 3D complex geometric template is necessary for the efficiency of the method. Therefore, we describe the thin layers with a shell-kind model using curvilinear coordinates. The purpose of this article is to obtain the above mentionned boundary conditions by means of a rigourous convergence result.

In order to simplify the mathematical analysis, we only consider two layers: an internal layer of fixed thickness ε_l and an external layer of thickness ε. These two layers have a common side, which is a surface $\hat{S}$ of $\mathbb{R}^3$. Therefore, the heart is represented by an elastic shell occupying a domain $\hat{\Omega}_\varepsilon = \hat{\Omega}^- \cup \hat{\Omega}^+_\varepsilon$, where $\hat{\Omega}^-$ is the internal layer and $\hat{\Omega}^+_\varepsilon$ is the external layer. The border of $\hat{\Omega}_\varepsilon$ is $\partial\hat{\Omega}_\varepsilon = \hat{\Gamma}^+_\varepsilon \cup \hat{\Gamma}^- \cup \hat{\Gamma}_{l,\varepsilon}$, where $\hat{\Gamma}^+_\varepsilon$ is the external border, $\hat{\Gamma}^-$ is the internal border and $\hat{\Gamma}_{l,\varepsilon}$ is the lateral border (see Fig. 2).

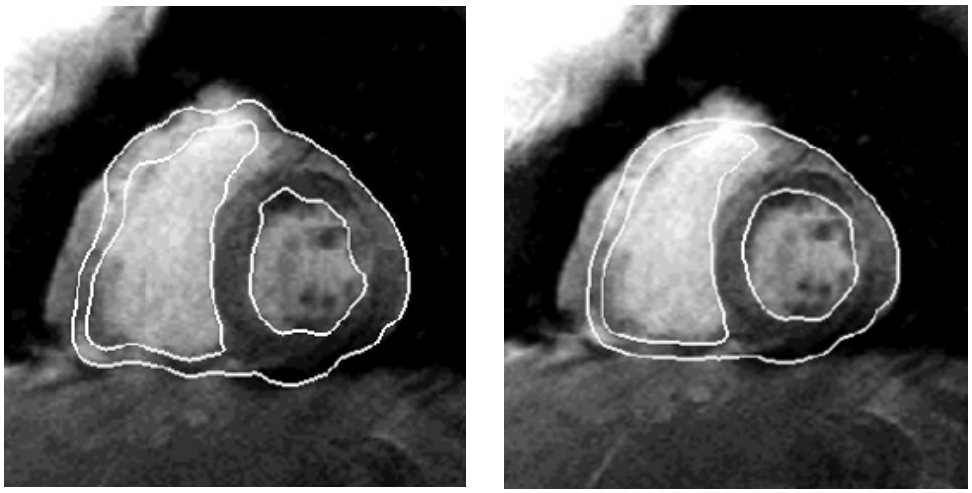

Fig. 1. Impact of the regularization on segmentation results for a mid-ventricular slice: without (left) and with (right) applied boundary conditions (from Pham [14]).

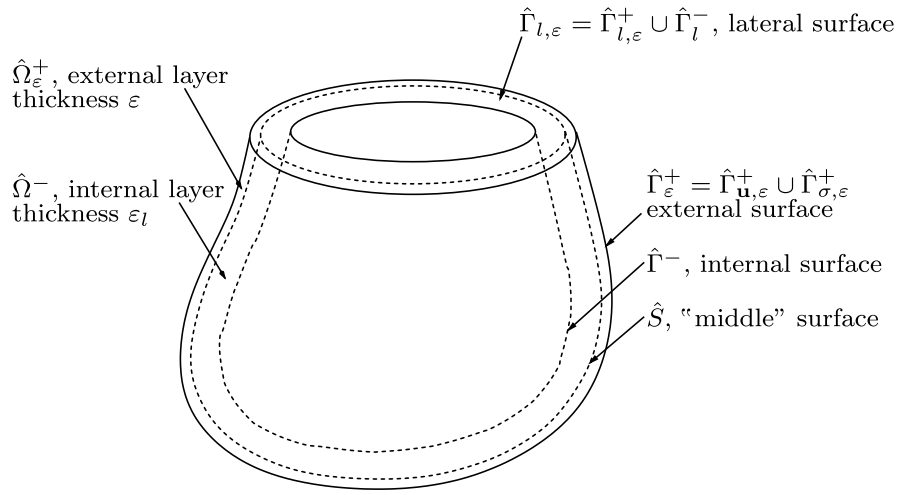

Fig. 2. The domain $\hat{\Omega}_\varepsilon = \hat{\Omega}^+_\varepsilon \cup \hat{\Omega}^-$.

We use the following classical conventions and notations throughout this work. Greek indices and exponents (except ε) belong to the set $\{1, 2\}$, whereas Latin indices belong to the set $\{1, 2, 3\}$. The summation convention with respect to repeated indices and exponents is systematically used. The Euclidean scalar product, the vector product and tensorial product of $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$ are denoted $\mathbf{a} \cdot \mathbf{b}$, $\mathbf{a} \times \mathbf{b}$ and $\mathbf{a} \otimes \mathbf{b}$, respectively; the Euclidean norm is denoted $|| \cdot ||$.

Let $(\mathbf{e}_i)$ be the canonical orthonormal basis of the Euclidean space $\mathbb{R}^3$. In cartesian coordinates the displacement field for any material point is represented by $\hat{\mathbf{u}} = \hat{u}_i \mathbf{e}_i$. The deformation is described by the Green–Lagrange strain tensor, which is linearized under the small deformation assumption:

$$\hat{e}_{ij}(\hat{\mathbf{u}}) = \frac{1}{2}\left(\frac{\partial \hat{u}_i}{\partial \hat{x}_j} + \frac{\partial \hat{u}_j}{\partial \hat{x}_i}\right).$$

If $\hat{\sigma}$ denotes the stress tensor, the constitutive law or stress-strain relation for the homogeneous isotropic internal layer takes the form:

$$\hat{\sigma}(\hat{\mathbf{u}}) = \lambda \ \mathrm{trace}(\hat{e}(\hat{\mathbf{u}}))I + 2\mu\hat{e}(\hat{\mathbf{u}}), \tag{1.1}$$

where λ and μ are the Lamé constants, and I is the identity tensor. Equivalently, we have,

$$\hat{e}(\hat{\mathbf{u}}) = \frac{1+\nu}{E}\hat{\sigma}(\hat{\mathbf{u}}) - \frac{\nu}{E}\ \mathrm{trace}(\hat{\sigma}(\hat{\mathbf{u}}))I, \tag{1.2}$$

where E is the Young modulus and ν the Poisson ratio. The following classical relations hold

$$\frac{\nu}{E} = \frac{\lambda}{4\mu(\lambda+\mu)}, \quad \frac{1+\nu}{E} = \frac{1}{2\mu}. \tag{1.3}$$

If $\hat{\mathbf{d}}$ is the 3D orientation vector of fibers belonging to the tangent space and μ_e is the second Lamé coefficient for the external layer, the constitutive law for this layer reads as follows [15]:

$$\hat{\sigma}(\hat{\mathbf{u}}) = (\hat{\mathbf{d}} \cdot \hat{e}(\hat{\mathbf{u}})\hat{\mathbf{d}})\hat{\mathbf{d}} \otimes \hat{\mathbf{d}} + 2\mu_e\varepsilon\hat{e}(\hat{\mathbf{u}}). \tag{1.4}$$

We will show in Sec. 3.1 that the inverse relation is well defined for all $\varepsilon > 0$. In the context of bonded joint with soft material, similar constitutive law models have been proposed in [12] or in [3].

We assume that the elastic body is submitted to a volumic force field $\hat{\mathbf{f}}$ such that $\hat{\mathbf{f}} = 0$ in $\hat{\Omega}_\varepsilon^+$. The equilibrium state is expressed by:

$$\begin{cases} \mathrm{div}(\hat{\sigma}(\hat{\mathbf{u}})) + \hat{\mathbf{f}} = 0 & \text{in } \hat{\Omega}_\varepsilon, \\ \hat{\sigma}(\hat{\mathbf{u}}) = \lambda \ \mathrm{trace}(\hat{e}(\hat{\mathbf{u}}))I + 2\mu\hat{e}(\hat{\mathbf{u}}) & \text{in } \hat{\Omega}^-, \\ \hat{\sigma}(\hat{\mathbf{u}}) = (\hat{\mathbf{d}} \cdot \hat{e}(\hat{\mathbf{u}})\hat{\mathbf{d}})\hat{\mathbf{d}} \otimes \hat{\mathbf{d}} + 2\mu_e\varepsilon\hat{e}(\hat{\mathbf{u}}) & \text{in } \hat{\Omega}_\varepsilon^+, \\ \hat{\mathbf{u}} = 0 & \text{on } \hat{\Gamma}^- \cup \hat{\Gamma}_{l,\varepsilon} \cup \hat{\Gamma}^+_{\mathbf{u},\varepsilon}, \\ \hat{\sigma}\mathbf{n} = 0 & \text{on } \hat{\Gamma}^+_{\sigma,\varepsilon}, \\ \hat{\mathbf{u}}^- = \hat{\mathbf{u}}^+ \ \text{ and } \ \hat{\sigma}^-\mathbf{n} = \hat{\sigma}^+\mathbf{n} & \text{on } \hat{S}, \end{cases} \tag{1.5}$$

where $\hat{\Gamma}_\varepsilon^+ = \hat{\Gamma}^+_{\mathbf{u},\varepsilon} \cup \hat{\Gamma}^+_{\sigma,\varepsilon}$ and $\mathrm{meas}(\hat{\Gamma}^+_{\mathbf{u},\varepsilon}) \neq 0$. $\hat{\mathbf{u}}^+$ (respectively $\hat{\mathbf{u}}^-$) is the restriction of $\hat{\mathbf{u}}$ to $\hat{\Omega}_\varepsilon^+$ (respectively $\hat{\Omega}^-$). The same notation applies to $\hat{\sigma}$. The vector $\mathbf{n}$ denotes the normal unit vector pointing outwards of $\hat{\Omega}_\varepsilon$ on $\hat{\Gamma}^+_{\sigma,\varepsilon}$ and outwards of $\hat{\Omega}^-$ on $\hat{S}$.

The goal of this work is to prove that when the thickness of the external layer, ε, tends to 0, the asymptotic model is given by:

$$\begin{cases} \mathrm{div}(\hat{\sigma}(\hat{\mathbf{u}})) + \hat{\mathbf{f}} = 0 & \text{in } \hat{\Omega}^-, \\ \hat{\sigma}(\hat{\mathbf{u}}) = \lambda \ \mathrm{trace}(\hat{e}(\hat{\mathbf{u}}))I + 2\mu\hat{e}(\hat{\mathbf{u}}) & \text{in } \hat{\Omega}^-, \\ \hat{\mathbf{u}} = 0 & \text{on } \hat{\Gamma}^- \cup \hat{\Gamma}_l^-, \\ \hat{\sigma}\mathbf{n} = -2\mu_e\hat{u}_n\mathbf{n} - \mu_e\hat{\mathbf{u}}_T & \text{on } \hat{S}, \end{cases} \tag{1.6}$$

where $\hat{\Gamma}_l^-$ is such that $\hat{\Gamma}_{l,\varepsilon} = \hat{\Gamma}^+_{l,\varepsilon} \cup \hat{\Gamma}_l^-$, $\hat{u}_n\mathbf{n}$ is the component of $\hat{\mathbf{u}}$ normal to the surface $\hat{S}$ and $\hat{\mathbf{u}}_T$ is the tangential component.

It is worth noticing here that $\hat{\Gamma}^+_{\mathbf{u},\varepsilon}$ and $\hat{\Gamma}^+_{\sigma,\varepsilon}$ disappear in the limit process. However, if, on the one hand, one can choose $\text{meas}(\hat{\Gamma}^+_{\sigma,\varepsilon}) = 0$, it is on the other hand necessary to have $\text{meas}(\hat{\Gamma}^+_{\mathbf{u},\varepsilon}) \neq 0$. The Dirichlet boundary condition on $\hat{\Gamma}^+_{\mathbf{u},\varepsilon}$ plays an important role in the proof of Theorem 6.7 at the end of the paper. It should also be noticed that the new boundary condition on $\hat{S}$ does not depend on the fibers direction $\hat{\mathbf{d}}$. If $\hat{\mathbf{d}}$ has a non zero component in the normal direction $\mathbf{n}$, the asymptotic model will be dramatically different.

An overview of the article is as follows. In the next section we collect most of the notation to be used in the remainder of the paper recalling basic notions on curvilinear coordinates. Using this notation in Sec. 3, we derive some estimations concerning the stress-strain relations in the internal layer, Ω^-, and in the external layer Ω^+_ε. In Sec. 4, we introduce the mixed variational formulation of the elasticity problem (1.5) and show its well-posedness. The problem is then reformulated in Sec. 5 over an ε-independent domain Ω. The main result of this paper is obtained in Sec. 6, in which we first prove several *a priori* estimations on the solution to the scaled problem before passing to the limit as ε tends to 0.

Let Ω be an open subset in $\mathbb{R}^3$. $L^2(\Omega)$, $||\cdot||_{0,\Omega}$ and $H^1(\Omega)$, $||\cdot||_{1,\Omega}$ denote the usual Sobolev spaces of real-valued functions. Boldface lowercase letters denote vector-valued functions and boldface uppercase letters denote matrix valued functions. The norms are denoted in the same way as for real-valued functions. For instance, if $\mathbf{v} \in (L^2(\Omega))^3$, we note $||\mathbf{v}||^2_{0,\Omega} = \sum_i ||v_i||^2_{0,\Omega}$.

2. Preliminaries

2.1. *Curvilinear coordinates*

All needed notions of differential geometry may be found, e.g., in [8]. The presentation given in this section is very close to the one given in [9]. We consider a shell described by a surface $\hat{S}$, the thickness of which is $\varepsilon_l + \varepsilon$. We assume that the surface $\hat{S}$ is a bounded, two-dimensional submanifold of $\mathbb{R}^3$, which, for simplicity, admits an atlas consisting of one chart only. Let $\boldsymbol{\psi}$ be this chart. We are thus given once and for all a domain $w \subset \mathbb{R}^2$ and an injective mapping $\boldsymbol{\psi} \in \mathcal{C}^3(\bar{w}, \mathbb{R}^3)$, such that

$$\hat{S} = \boldsymbol{\psi}(\bar{w}).$$

We assume that w has a Lipschitz-continuous boundary, γ. Let $y = (y_\alpha)$ denote a generic point in the set $\bar{w}$ and let $\partial_\alpha = \partial/\partial y_\alpha$. Let $\boldsymbol{\psi}$ be such that the two vectors

$$\mathbf{a}_\alpha(y) = \partial_\alpha \boldsymbol{\psi}(y),$$

are linearly independent at all points $y \in \bar{w}$. They form the covariant basis of the tangent plane, $T(\hat{S})$, to the surface $\hat{S}$ at the point $\boldsymbol{\psi}(y)$. The two vectors $\mathbf{a}^\alpha(y)$ of the same tangent plane defined by the relations

$$\mathbf{a}^\alpha(y) \cdot \mathbf{a}_\beta(y) = \delta^\alpha_\beta,$$

constitute its contravariant basis. Let us also define

$$\mathbf{a}_3(y) = \mathbf{a}^3(y) = \frac{\mathbf{a}_1(y) \times \mathbf{a}_2(y)}{||\mathbf{a}_1(y) \times \mathbf{a}_2(y)||},$$

which is a chart-independent (modulo multiplication by -1) unit normal vector to the tangent plane. One then defines the metric tensor, $(a_{\alpha\beta})$ or $(a^{\alpha\beta})$ (in covariant or contravariant components), the curvature tensor, $(b_{\alpha\beta})$ or (b_α^β) (in covariant or mixed components), and the Christoffel symbols $\Gamma_{\alpha\beta}^\rho$, of the surface $\hat{S}$ by letting

$$a_{\alpha\beta} = \mathbf{a}_\alpha \cdot \mathbf{a}_\beta, \quad a^{\alpha\beta} = \mathbf{a}^\alpha \cdot \mathbf{a}^\beta, \tag{2.1}$$

$$b_{\alpha\beta} = \mathbf{a}_3 \cdot \partial_\beta \mathbf{a}_\alpha, \quad b_\alpha^\beta = a^{\beta\alpha} b_{\sigma\alpha}, \tag{2.2}$$

$$\Gamma_{\alpha\beta}^\rho = \mathbf{a}^\rho \cdot \partial_\beta \mathbf{a}_\alpha. \tag{2.3}$$

Note the symmetries:

$$a_{\alpha\beta} = a_{\beta\alpha}, \quad a^{\alpha\beta} = a^{\beta\alpha}, \quad b_{\alpha\beta} = b_{\beta\alpha}, \quad \Gamma_{\alpha\beta}^\rho = \Gamma_{\beta\alpha}^\rho.$$

The area element along $\hat{S}$ is $\sqrt{a}dy$, where

$$a = \det(a_{\alpha\beta}). \tag{2.4}$$

The function a is continuous on the set $\bar{w}$ and there exits a constant a_0, such that

$$0 < a_0 \leq a(y), \quad \forall y \in \bar{w}. \tag{2.5}$$

For each $\varepsilon > 0$ we define the sets:

$$\begin{aligned}
\Omega_\varepsilon &= w \times]-\varepsilon_l, \varepsilon[, \\
\Omega_\varepsilon^+ &= w \times]0, \varepsilon[, \\
\Omega^- &= w \times]-\varepsilon_l, 0[, \\
\Gamma_{l,\varepsilon}^+ &= \gamma \times [0, \varepsilon], \\
\Gamma_l^- &= \gamma \times [-\varepsilon_l, 0], \\
\Gamma^- &= w \times \{-\varepsilon_l\}, \\
\Gamma_{\mathbf{u},\varepsilon}^+ &= w_\mathbf{u} \times \{\varepsilon\}, \\
\Gamma_{\sigma,\varepsilon}^+ &= w_\sigma \times \{\varepsilon\}, \\
S &= w \times \{0\},
\end{aligned}$$

with $w = w_\mathbf{u} \cup w_\sigma$ and $\text{meas}(w_\mathbf{u}) \neq 0$. Note that $\Gamma_{l,\varepsilon}^+ \cup \Gamma_l^- \cup \Gamma^- \cup \Gamma_{\mathbf{u},\varepsilon}^+ \cup \Gamma_{\sigma,\varepsilon}^+ = \partial\Omega_\varepsilon$ constitutes a partition of the boundary of the set Ω_ε (see Fig. 3).

Let $x^\varepsilon = (x_i^\varepsilon)$ denote a generic point in $\bar{\Omega}_\varepsilon$, and let $\partial_i^\varepsilon = \partial/\partial x_i^\varepsilon$; hence $x_\alpha^\varepsilon = y_\alpha$ and $\partial_\alpha^\varepsilon = \partial_\alpha$. The initial configuration of the shell is the image of Ω_ε by the mapping $\mathbf{\Psi}$: $\bar{\Omega}_\varepsilon \to \mathbb{R}^3$ defined by

$$\mathbf{\Psi}(x^\varepsilon) = \boldsymbol{\psi}(y) + x_3^\varepsilon \mathbf{a}_3(y), \quad \forall x^\varepsilon = (y, x_3^\varepsilon) \in \bar{\Omega}_\varepsilon.$$

It can then be shown (cf. [8]) that there exists $\varepsilon_0 > 0$, such that the mapping $\mathbf{\Psi}$ is a $\mathcal{C}^2$-diffeomorphism, and the three vectors,

$$\mathbf{g}_i^\varepsilon(x^\varepsilon) = \partial_i^\varepsilon \mathbf{\Psi}(x^\varepsilon),$$

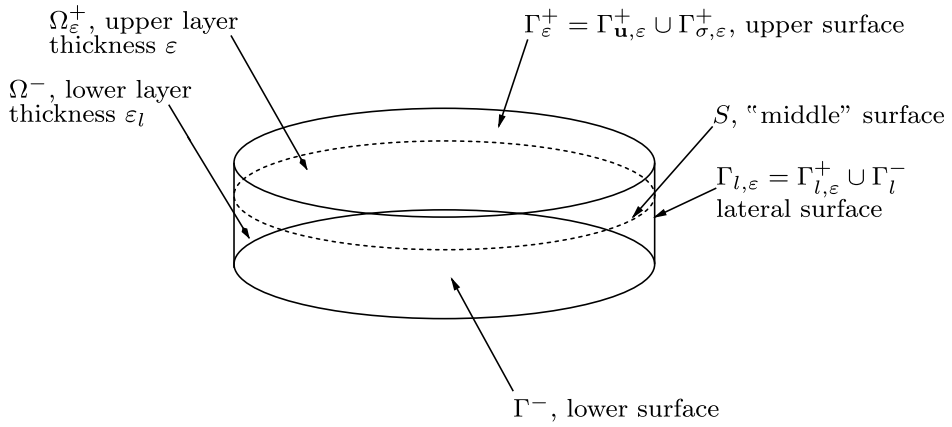

Fig. 3. The domain $\Omega_\varepsilon = \Omega_\varepsilon^+ \cup \Omega^-$.

are linearly independent at all points $x^\varepsilon \in \bar{\Omega}_\varepsilon$ for all $0 < \varepsilon < \varepsilon_0$. Therefore, we make a geometrical assumption on the thicknesses of the two layers of the shell heart model:

$$0 < \varepsilon \leq \varepsilon_l \leq \varepsilon_0.$$

The three vectors $\mathbf{g}_i^\varepsilon(x^\varepsilon)$ define the covariant basis at the point $\boldsymbol{\Psi}(x^\varepsilon)$. It is clear that $\mathbf{g}_3^\varepsilon = \mathbf{a}_3$ is the unit vector normal to $\hat{S}$. We choose it to be pointing outwards of Ω^- and for the remainder of this work, we use indifferently the notations $\mathbf{n}$ or $\mathbf{g}_3^\varepsilon$. The three vectors $\mathbf{g}^{i,\varepsilon}(x^\varepsilon)$ defined by

$$\mathbf{g}^{j,\varepsilon}(x^\varepsilon) \cdot \mathbf{g}_i^\varepsilon(x^\varepsilon) = \delta_i^j,$$

form the contravariant basis. One then defines the metric tensor (g_{ij}^ε) or $(g^{ij,\varepsilon})$ (in covariant or contravariant components) and the Christoffel symbols of the manifold $\boldsymbol{\Psi}(\bar{\Omega}_\varepsilon)$ by letting (we omit the explicit dependence on x^ε)

$$g_{ij}^\varepsilon = \mathbf{g}_i^\varepsilon \cdot \mathbf{g}_j^\varepsilon, \quad g^{ij,\varepsilon} = \mathbf{g}^{i,\varepsilon} \cdot \mathbf{g}^{j,\varepsilon},$$

$$\Gamma_{ij}^{p,\varepsilon} = \mathbf{g}^{p,\varepsilon} \cdot \partial_i^\varepsilon \mathbf{g}_j^\varepsilon.$$

Note the symmetries

$$g_{ij}^\varepsilon = g_{ji}^\varepsilon, \quad g^{ij,\varepsilon} = g^{ji,\varepsilon}, \quad \Gamma_{ij}^{p,\varepsilon} = \Gamma_{ji}^{p,\varepsilon},$$

and the relations

$$\Gamma_{\alpha 3}^{3,\varepsilon} = \Gamma_{33}^{p,\varepsilon} = 0 \quad \text{in } \bar{\Omega}_\varepsilon.$$

The volume element in the set $\boldsymbol{\Psi}(\Omega_\varepsilon)$ is $\sqrt{g^\varepsilon}dx^\varepsilon$, where

$$g^\varepsilon = \det(g_{ij}^\varepsilon). \tag{2.6}$$

2.2. *Vectors and tensors in curvilinear coordinates*

With all the notations defined in the preceding section, a vector field or a second order symmetric tensor field defined on the shell may be represented in the curvilinear system by its covariant or contravariant components:

$$v_i^\varepsilon \mathbf{g}^{i,\varepsilon} = v^{i,\varepsilon}\mathbf{g}_i^\varepsilon,$$
$$\tau_{ij}^\varepsilon \mathbf{g}^{i,\varepsilon} \otimes \mathbf{g}^{j,\varepsilon} = \tau^{ij,\varepsilon}\mathbf{g}_i^\varepsilon \otimes \mathbf{g}_j^\varepsilon.$$

One can relate covariant and contravariant components, thanks to the relations

$$\begin{aligned} v^{i,\varepsilon} &= g^{ik,\varepsilon}v_k^\varepsilon, & v_i^\varepsilon &= g_{ik}^\varepsilon v^{k,\varepsilon}, \\ \tau^{ij,\varepsilon} &= \frac{1}{2}(g^{ik,\varepsilon}g^{jl,\varepsilon} + g^{jk,\varepsilon}g^{il,\varepsilon})\tau_{kl}^\varepsilon, & \tau_{ij}^\varepsilon &= \frac{1}{2}(g_{ik}^\varepsilon g_{jl}^\varepsilon + g_{jk}^\varepsilon g_{il}^\varepsilon)\tau^{kl,\varepsilon}, \\ &= G^{ijkl,\varepsilon}\tau_{kl}^\varepsilon, & &= H_{ijkl}^\varepsilon \tau^{kl,\varepsilon}. \end{aligned}$$

Concerning the fourth-order tensors $(G^{ijkl,\varepsilon})$ and (H_{ijkl}^ε), the following relations hold for each $\varepsilon > 0$

$$G^{\alpha\beta k3,\varepsilon} = G^{333\alpha,\varepsilon} = 0,$$
$$H_{\alpha\beta k3}^\varepsilon = H_{333\alpha}^\varepsilon = 0,$$

and

$$G^{ijkl,\varepsilon} = G^{jikl,\varepsilon} = G^{klij,\varepsilon},$$
$$H_{ijkl}^\varepsilon = H_{jikl}^\varepsilon = H_{klij}^\varepsilon.$$

Both tensors are symmetric, positive definite, and uniform with respect to $x^\varepsilon \in \bar{\Omega}_\varepsilon$. The scalar product between two vectors, $u^{i,\varepsilon}\mathbf{g}_i^\varepsilon$ and $v_i^\varepsilon \mathbf{g}^{i,\varepsilon}$ can be written as

$$(u^{i,\varepsilon}\mathbf{g}_i^\varepsilon)\cdot(v_i^\varepsilon \mathbf{g}^{i,\varepsilon}) = u^{i,\varepsilon}v_i^\varepsilon.$$

The second-order inner product between two tensors can be written as

$$(\tau^{ij,\varepsilon}\mathbf{g}_i^\varepsilon \otimes \mathbf{g}_j^\varepsilon)\colon(\sigma_{ij}^\varepsilon \mathbf{g}^{i,\varepsilon} \otimes \mathbf{g}^{j,\varepsilon}) = \tau^{ij,\varepsilon}\sigma_{ij}^\varepsilon.$$

Using the fourth-order tensor $G^{ijkl,\varepsilon}$, this expression can be transformed to

$$\tau^{ij,\varepsilon}\sigma_{ij}^\varepsilon = G^{ijkl,\varepsilon}\tau_{kl}^\varepsilon\sigma_{ij}^\varepsilon.$$

Let us now introduce the vectorial notation which we will use. $\mathbf{S}$ denotes the set of all symmetric matrices of order 3. Any $(\tau_{ij}) \in \mathbf{S}$ can be represented by a vector $\boldsymbol{\tau} \in \mathbb{R}^6$:

$$\boldsymbol{\tau} = (\tau_{11}, \sqrt{2}\tau_{12}, \tau_{22}, \sqrt{2}\tau_{13}, \sqrt{2}\tau_{23}, \tau_{33})^T.$$

We also note

$$\boldsymbol{\tau}_T = (\tau_{11}, \sqrt{2}\tau_{12}, \tau_{22})^T, \quad \boldsymbol{\tau}_N = (\sqrt{2}\tau_{13}, \sqrt{2}\tau_{23}, \tau_{33})^T.$$

The fourth-order tensor $G^{ijkl,\varepsilon}$ can be represented by the 6×6 symmetric matrix $\mathbf{G}^\varepsilon$:

$$\mathbf{G}^\varepsilon = \left(\begin{array}{c|c} \mathbf{G}_T^\varepsilon & 0 \\ \hline 0 & \mathbf{G}_N^\varepsilon \end{array}\right),$$

with

$$\mathbf{G}_T^\varepsilon = \begin{pmatrix} g^{11,\varepsilon}g^{11,\varepsilon} & \sqrt{2}g^{11,\varepsilon}g^{12,\varepsilon} & g^{12,\varepsilon}g^{12,\varepsilon} \\ \sqrt{2}g^{11,\varepsilon}g^{12,\varepsilon} & g^{11,\varepsilon}g^{22,\varepsilon}+g^{12,\varepsilon}g^{12,\varepsilon} & \sqrt{2}g^{12,\varepsilon}g^{22,\varepsilon} \\ g^{12,\varepsilon}g^{12,\varepsilon} & \sqrt{2}g^{12,\varepsilon}g^{22,\varepsilon} & g^{22,\varepsilon}g^{22,\varepsilon} \end{pmatrix},$$

and

$$\mathbf{G}_N^\varepsilon = \begin{pmatrix} g^{11,\varepsilon}g^{33,\varepsilon} & g^{12,\varepsilon}g^{33,\varepsilon} & 0 \\ g^{12,\varepsilon}g^{33,\varepsilon} & g^{22,\varepsilon}g^{33,\varepsilon} & 0 \\ 0 & 0 & g^{33,\varepsilon}g^{33,\varepsilon} \end{pmatrix}.$$

Recalling that the (g_{ij}^ε) matrix is the inverse of the $(g^{ij,\varepsilon})$ matrix, we note that

$$(\mathbf{G}^\varepsilon)^{-1} = \mathbf{H}^\varepsilon = \left(\begin{array}{c|c} \mathbf{H}_T^\varepsilon & 0 \\ \hline 0 & \mathbf{H}_N^\varepsilon \end{array}\right),$$

with

$$\mathbf{H}_T^\varepsilon = \begin{pmatrix} g_{11}^\varepsilon g_{11}^\varepsilon & \sqrt{2}g_{11}^\varepsilon g_{12}^\varepsilon & g_{12}^\varepsilon g_{12}^\varepsilon \\ \sqrt{2}g_{11}^\varepsilon g_{12}^\varepsilon & g_{11}^\varepsilon g_{22}^\varepsilon + g_{12}^\varepsilon g_{12}^\varepsilon & \sqrt{2}g_{12}^\varepsilon g_{22}^\varepsilon \\ g_{12}^\varepsilon g_{12}^\varepsilon & \sqrt{2}g_{12}^\varepsilon g_{22}^\varepsilon & g_{22}^\varepsilon g_{22}^\varepsilon \end{pmatrix},$$

and

$$\mathbf{H}_N^\varepsilon = \begin{pmatrix} g_{11}^\varepsilon g_{33}^\varepsilon & g_{12}^\varepsilon g_{33}^\varepsilon & 0 \\ g_{12}^\varepsilon g_{33}^\varepsilon & g_{22}^\varepsilon g_{33}^\varepsilon & 0 \\ 0 & 0 & g_{33}^\varepsilon g_{33}^\varepsilon \end{pmatrix}.$$

In vectorial notation the second-order inner product between two symmetric tensors is written

$$\begin{aligned} (\tau^{ij,\varepsilon}\mathbf{g}_i^\varepsilon \otimes \mathbf{g}_j^\varepsilon)\colon(\sigma_{ij}^\varepsilon \mathbf{g}^{i,\varepsilon}\otimes\mathbf{g}^{j,\varepsilon}) &= G^{ijkl,\varepsilon}\tau_{kl}^\varepsilon\sigma_{ij}^\varepsilon, \\ &= \boldsymbol{\tau}^\varepsilon\cdot\mathbf{G}^\varepsilon\boldsymbol{\sigma}^\varepsilon, \\ &= \boldsymbol{\tau}_T^\varepsilon\cdot\mathbf{G}_T^\varepsilon\boldsymbol{\sigma}_T^\varepsilon + \boldsymbol{\tau}_N^\varepsilon\cdot\mathbf{G}_N^\varepsilon\boldsymbol{\sigma}_N^\varepsilon. \end{aligned}$$

The fact that $(G^{ijkl,\varepsilon})$ is symmetric, positive definite, and uniform with respect to $x^\varepsilon \in \bar{\Omega}_\varepsilon$ implies that there exists a constant $c_G^\varepsilon > 0$ depending on Ω_ε only (thus on the small parameter ε), such that

$$\boldsymbol{\tau}\cdot\mathbf{G}^\varepsilon\boldsymbol{\tau} \geq c_G^\varepsilon\boldsymbol{\tau}\cdot\boldsymbol{\tau} = c_G^\varepsilon\tau_{ij}\tau_{ij}, \tag{2.7}$$

for all $x^\varepsilon \in \bar{\Omega}_\varepsilon$ and all $(\tau_{ij}) \in \mathbf{S}$.

From the continuity of $x^\varepsilon \to \mathbf{G}^\varepsilon(x^\varepsilon)$ on $\bar{\Omega}^\varepsilon$ we also deduce that there exists a constant C_G^ε, such that

$$\boldsymbol{\tau}\cdot\mathbf{G}^\varepsilon\boldsymbol{\sigma} \leq C_G^\varepsilon||\boldsymbol{\tau}||||\boldsymbol{\sigma}|| \tag{2.8}$$

for all $x^\varepsilon \in \bar{\Omega}_\varepsilon$ and all $(\tau_{ij}), (\sigma_{ij}) \in \mathbf{S}$.

It is clear that if we consider the restrictions of functions $G^{ijkl,\varepsilon}$ to Ω^-, inequality (2.7) still holds with an ε-independent constant $C_G > 0$. To emphasize the fact that the restriction to Ω^- of geometrical quantities such as $\mathbf{g}_i^\varepsilon$, $G^{ijkl,\varepsilon}$, ... are ε-independent, we omit the exponents ε in what follows. For example, $\mathbf{g}_i$ denotes the restriction of $\mathbf{g}_i^\varepsilon$ to Ω^-. We then have

$$\boldsymbol{\tau} \cdot \mathbf{G}\boldsymbol{\tau} \geq c_G \boldsymbol{\tau} \cdot \boldsymbol{\tau} = c_G \tau_{ij} \tau_{ij}, \tag{2.9}$$

for all $x^\varepsilon \in \bar{\Omega}^-$ and all $(\tau_{ij}) \in \mathbf{S}$.

To conclude this section, let us recall that given the covariant components $(u_i^\varepsilon) = \mathbf{u}^\varepsilon$ of an arbitrary displacement field $u_i^\varepsilon \mathbf{g}^{i,\varepsilon}$ of the points of the shell, the covariant components of the linearized strain tensor read

$$\begin{aligned} e_{\alpha\beta}^\varepsilon(\mathbf{u}^\varepsilon) &= \frac{1}{2}(\partial_\alpha^\varepsilon u_\beta^\varepsilon + \partial_\beta^\varepsilon u_\alpha^\varepsilon) - \Gamma_{\alpha\beta}^{p,\varepsilon} u_p^\varepsilon, \\ e_{\alpha 3}^\varepsilon(\mathbf{u}^\varepsilon) &= \frac{1}{2}(\partial_\alpha^\varepsilon u_3^\varepsilon + \partial_3^\varepsilon u_\alpha^\varepsilon) - \Gamma_{\alpha 3}^{\rho,\varepsilon} u_\rho^\varepsilon, \\ e_{33}^\varepsilon(\mathbf{u}^\varepsilon) &= \partial_3^\varepsilon u_3^\varepsilon. \end{aligned} \tag{2.10}$$

Using our vectorial notation, the associated vector of $\mathbb{R}^6$ is denoted by

$$\mathbf{e}^\varepsilon(\mathbf{u}^\varepsilon) = (\mathbf{e}_T^\varepsilon(\mathbf{u}^\varepsilon),\ \mathbf{e}_N^\varepsilon(\mathbf{u}^\varepsilon)).$$

3. Strain-Stress Relation

In this section the strain-stress relations in the internal and external layers are expressed using the vectorial notation. We introduce a new basis of $\mathbb{R}^3$ in order to derive estimations (3.3), (3.4) and (3.7), (3.8) which are needed in the remaining part of the paper.

3.1. *Strain-stress relation in the external layer*

Assume that the linearized strain tensor is described by its *contravariant* components, $e^{kl,\varepsilon}(\mathbf{u}^\varepsilon)$ and that the stress tensor is described by its *covariant* components $\sigma_{ij}^\varepsilon(\mathbf{u}^\varepsilon)$. Assume that the orientation vector of fibers is tangent to the surface $\hat{S}$ and that it is defined by its covariant components, d_α. These components are assumed to be x_3-independent, that is to say, $d_\alpha = d_\alpha(x_1, x_2)$. We also assume that $d_\alpha \in \mathcal{C}^0(\bar{w}, \mathbb{R})$ and that for all $(x_1, x_2) \in \bar{w}$, $\mathbf{d} \neq 0$.

Omiting the explicit dependence on $\mathbf{u}^\varepsilon$, the constitutive law (1.4) for the external fibrous layer then reads

$$\begin{aligned} \sigma_{ij}^\varepsilon \mathbf{g}^{i,\varepsilon} \otimes \mathbf{g}^{j,\varepsilon} &= (e^{kl,\varepsilon} d_k d_l) d_i d_j \mathbf{g}^{i,\varepsilon} \otimes \mathbf{g}^{j,\varepsilon} \\ &\quad + 2\mu_e \varepsilon \frac{1}{2}(g_{ik}^\varepsilon g_{jl}^\varepsilon + g_{jk}^\varepsilon g_{il}^\varepsilon) e^{kl,\varepsilon} \mathbf{g}^{i,\varepsilon} \otimes \mathbf{g}^{j,\varepsilon}. \end{aligned} \tag{3.1}$$

This relation can be written as

$$\sigma_{ij}^\varepsilon = B_{ijkl}^\varepsilon e^{kl,\varepsilon}, \tag{3.2}$$

where

$$B^{\varepsilon}_{ijkl} = d_i d_j d_k d_l + 2\mu_e \varepsilon \frac{1}{2}(g^{\varepsilon}_{ik} g^{\varepsilon}_{jl} + g^{\varepsilon}_{jk} g^{\varepsilon}_{il}).$$

Note the symmetries

$$B^{\varepsilon}_{ijkl} = B^{\varepsilon}_{jikl} = B^{\varepsilon}_{klij}.$$

Since $d_3 = 0$ the following relation holds

$$B^{\varepsilon}_{\alpha\beta k3} = B^{\varepsilon}_{\alpha 333} = 0.$$

The fourth-order symmetric tensor (B^{ε}_{ijkl}) defined by its covariant components is known as the stiffness tensor. In order to establish the mixed variational formulation of the problem, we need to use the inverse relation and the associated compliance tensor $(C^{ijkl,\varepsilon})$ defined by its contravariant components. Let us show that $(C^{ijkl,\varepsilon})$, the inverse of (B^{ε}_{ijkl}), exists for all $\varepsilon > 0$.

The contravariant components $C^{ijkl,\varepsilon}\colon \bar{\Omega}^{+}_{\varepsilon} \to \mathbb{R}$ of the compliance tensor $(C^{ijkl,\varepsilon})$ are obtained by inverting the matrix of covariant components of the stiffness tensor, $B^{\varepsilon}_{ijkl}\colon \bar{\Omega}^{+}_{\varepsilon} \to \mathbb{R}$. In vectorial notation, relation (3.2) reads

$$\boldsymbol{\sigma}^{\varepsilon} = \mathbf{B}^{\varepsilon} \mathbf{e}^{\varepsilon},$$

where

$$\boldsymbol{\sigma}^{\varepsilon} = (\sigma^{\varepsilon}_{11}, \sqrt{2}\sigma^{\varepsilon}_{12}, \sigma^{\varepsilon}_{22}, \sqrt{2}\sigma^{\varepsilon}_{13}, \sqrt{2}\sigma^{\varepsilon}_{23}, \sigma^{\varepsilon}_{33})^T,$$
$$\mathbf{e}^{\varepsilon} = (e^{11,\varepsilon}, \sqrt{2}e^{12,\varepsilon}, e^{22,\varepsilon}, \sqrt{2}e^{13,\varepsilon}, \sqrt{2}e^{23,\varepsilon}, e^{33,\varepsilon})^T.$$

$\mathbf{B}^{\varepsilon}$ is the 6×6 matrix defined by

$$\mathbf{B}^{\varepsilon} = \mathbf{D} + 2\mu_e \varepsilon \mathbf{H}^{\varepsilon},$$

with

$$\mathbf{B}^{\varepsilon} = \left(\begin{array}{c|c} \mathbf{B}^{\varepsilon}_T & 0 \\ \hline 0 & \mathbf{B}^{\varepsilon}_N \end{array}\right), \quad \mathbf{H}^{\varepsilon} = \left(\begin{array}{c|c} \mathbf{H}^{\varepsilon}_T & 0 \\ \hline 0 & \mathbf{H}^{\varepsilon}_N \end{array}\right), \quad \mathbf{D} = \left(\begin{array}{c|c} \mathbf{D}_T & 0 \\ \hline 0 & 0 \end{array}\right),$$

and

$$\mathbf{D}_T = \begin{pmatrix} (d_1)^4 & \sqrt{2}(d_1)^3 d_2 & (d_1)^2 (d_2)^2 \\ \sqrt{2}(d_1)^3 d_2 & 2(d_1)^2 (d_2)^2 & \sqrt{2} d_1 (d_2)^3 \\ (d_1)^2 (d_2)^2 & \sqrt{2} d_1 (d_2)^3 & (d_2)^4 \end{pmatrix}.$$

$\mathbf{H}^{\varepsilon}$ is symmetric, positive definite and uniform with respect to $x^{\varepsilon} \in \bar{\Omega}_{\varepsilon}$ similar to the fourth-order tensor $(H^{ijkl,\varepsilon})$ is. $\mathbf{D}$ is symmetric and non-negative as its rank is one and its only non-zero eigenvalue is $\mathrm{trace}(\mathbf{D}_T) > 0$. Consequently, for all $\varepsilon > 0$, $\mathbf{B}^{\varepsilon}$ is symmetric, positive definite and therefore invertible. Moreover, as $(\mathbf{H}^{\varepsilon}_N)^{-1} = \mathbf{G}^{\varepsilon}_N$, we have

$$(\mathbf{B}^{\varepsilon})^{-1} = \mathbf{C}^{\varepsilon} = \left(\begin{array}{c|c} (\mathbf{B}^{\varepsilon}_T)^{-1} & 0 \\ \hline 0 & \dfrac{1}{2\mu_e \varepsilon} \mathbf{G}^{\varepsilon}_N \end{array}\right).$$

In order to obtain a simple expression for $(\mathbf{B}_T^\varepsilon)^{-1}$ one has to notice that $\mathbf{H}_T^\varepsilon$ is symmetric, positive definite and $\mathbf{D}_T$ is symmetric, positive definite and uniform with respect to $x^\varepsilon \in \bar{\Omega}_\varepsilon$. Therefore, it follows from a classical result (see Appendix A) on the simultaneous reduction of two quadratic forms that there exists a 3×3 invertible matrix $\mathbf{P}_T^\varepsilon$, such that

$$(\mathbf{P}_T^\varepsilon)^T \mathbf{H}_T^\varepsilon \mathbf{P}_T^\varepsilon = \mathbf{I},$$
$$(\mathbf{P}_T^\varepsilon)^T \mathbf{D}_T \mathbf{P}_T^\varepsilon = \mathbf{S}^\varepsilon,$$

with

$$\mathbf{S}^\varepsilon = \begin{pmatrix} s_\varepsilon & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Note that for all $\varepsilon > 0$ and for all $x^\varepsilon \in \bar{\Omega}_\varepsilon$, $s_\varepsilon > 0$.

We obtain

$$(\mathbf{P}_T^\varepsilon)^T \mathbf{B}_T^\varepsilon \mathbf{P}_T^\varepsilon = \begin{pmatrix} s_\varepsilon + 2\mu_e\varepsilon & 0 & 0 \\ 0 & 2\mu_e\varepsilon & 0 \\ 0 & 0 & 2\mu_e\varepsilon \end{pmatrix},$$

and therefore

$$(\mathbf{Q}_T^\varepsilon)^T (\mathbf{B}_T^\varepsilon)^{-1} \mathbf{Q}_T^\varepsilon = \begin{pmatrix} \dfrac{1}{s_\varepsilon + 2\mu_e\varepsilon} & 0 & 0 \\ 0 & \dfrac{1}{2\mu_e\varepsilon} & 0 \\ 0 & 0 & \dfrac{1}{2\mu_e\varepsilon} \end{pmatrix},$$

where

$$\mathbf{Q}_T^\varepsilon = ((\mathbf{P}_T^\varepsilon)^T)^{-1}.$$

The columns of the matrix

$$\mathbf{Q}^\varepsilon = \left(\begin{array}{c|c} \mathbf{Q}_T^\varepsilon & 0 \\ \hline 0 & \mathbf{I}_3 \end{array}\right)$$

define a new basis of $\mathbb{R}^6$. Any $(\tau_{ij}) \in \mathbf{S}$ represented by a vector

$$\boldsymbol{\tau} = (\tau_{11}, \sqrt{2}\tau_{12}, \tau_{22}, \sqrt{2}\tau_{13}, \sqrt{2}\tau_{23}, \tau_{33})^T,$$

in the canonical basis of $\mathbb{R}^6$ is represented by a vector $\tilde{\boldsymbol{\tau}}$ in this new basis. We have $\boldsymbol{\tau}_T = \mathbf{Q}_T^\varepsilon \tilde{\boldsymbol{\tau}}_T$ and $\boldsymbol{\tau}_N = \tilde{\boldsymbol{\tau}}_N$.

Since $\mathbf{G}_N^\varepsilon$ is symmetric, positive definite and uniform with respect to $x^\varepsilon \in \bar{\Omega}_\varepsilon$, there exist two constants $C_G^\varepsilon > 0$ and $c_G^\varepsilon > 0$ depending on Ω_ε, such that

$$\begin{aligned}
\boldsymbol{\tau}\cdot\mathbf{C}^\varepsilon\boldsymbol{\tau} &= \boldsymbol{\tau}\cdot(\mathbf{B}^\varepsilon)^{-1}\boldsymbol{\tau} = \tilde{\boldsymbol{\tau}}\cdot(\mathbf{Q}^\varepsilon)^T(\mathbf{B}^\varepsilon)^{-1}\mathbf{Q}^\varepsilon\tilde{\boldsymbol{\tau}},\\
&= \tilde{\boldsymbol{\tau}}_T\cdot(\mathbf{Q}_T^\varepsilon)^T(\mathbf{B}_T^\varepsilon)^{-1}\mathbf{Q}_T^\varepsilon\tilde{\boldsymbol{\tau}}_T + \frac{1}{2\mu_e\varepsilon}\boldsymbol{\tau}_N\cdot\mathbf{G}_N^\varepsilon\boldsymbol{\tau}_N,\\
&= \frac{1}{s_\varepsilon+2\mu_e\varepsilon}\tilde{\tau}_{11}^2 + \frac{1}{\mu_e\varepsilon}\tilde{\tau}_{12}^2 + \frac{1}{2\mu_e\varepsilon}\tilde{\tau}_{22}^2 + \frac{1}{2\mu_e\varepsilon}\boldsymbol{\tau}_N\cdot\mathbf{G}_N^\varepsilon\boldsymbol{\tau}_N,\\
&\geq \frac{1}{s_\varepsilon+2\mu_e\varepsilon}\tilde{\tau}_{11}^2 + \frac{1}{\mu_e\varepsilon}\tilde{\tau}_{12}^2 + \frac{1}{2\mu_e\varepsilon}\tilde{\tau}_{22}^2 + \frac{c_G^\varepsilon}{2\mu_e\varepsilon}(2\tau_{13}^2+2\tau_{23}^2+\tau_{33}^2), \qquad (3.3)
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\sigma}\cdot\mathbf{C}^\varepsilon\boldsymbol{\tau} &= \boldsymbol{\sigma}\cdot(\mathbf{B}^\varepsilon)^{-1}\boldsymbol{\tau} = \tilde{\boldsymbol{\sigma}}\cdot(\mathbf{Q}^\varepsilon)^T(\mathbf{B}^\varepsilon)^{-1}\mathbf{Q}^\varepsilon\tilde{\boldsymbol{\tau}},\\
&= \tilde{\boldsymbol{\sigma}}_T\cdot(\mathbf{Q}_T^\varepsilon)^T(\mathbf{B}_T^\varepsilon)^{-1}\mathbf{Q}_T^\varepsilon\tilde{\boldsymbol{\sigma}}_T + \frac{1}{2\mu_e\varepsilon}\boldsymbol{\sigma}_N\cdot\mathbf{G}_N^\varepsilon\boldsymbol{\tau}_N,\\
&= \frac{1}{s_\varepsilon+2\mu_e\varepsilon}\tilde{\sigma}_{11}\tilde{\tau}_{11} + \frac{1}{\mu_e\varepsilon}\tilde{\sigma}_{12}\tilde{\tau}_{12} + \frac{1}{2\mu_e\varepsilon}\tilde{\sigma}_{22}\tilde{\tau}_{22} + \frac{1}{2\mu_e\varepsilon}\boldsymbol{\sigma}_N\cdot\mathbf{G}_N^\varepsilon\boldsymbol{\tau}_N,\\
&\leq \frac{1}{s_\varepsilon+2\mu_e\varepsilon}\tilde{\sigma}_{11}\tilde{\tau}_{11} + \frac{1}{\mu_e\varepsilon}\tilde{\sigma}_{12}\tilde{\tau}_{12} + \frac{1}{2\mu_e\varepsilon}\tilde{\sigma}_{22}\tilde{\tau}_{22}\\
&\quad + \frac{C_G^\varepsilon}{2\mu_e\varepsilon}(2\sigma_{13}^2+2\sigma_{23}^2+\sigma_{33}^2)^{1/2}(2\tau_{13}^2+2\tau_{23}^2+\tau_{33}^2)^{1/2}, \qquad (3.4)
\end{aligned}$$

for all $(\tau_{ij}), (\sigma_{ij}) \in \mathbf{S}$ and all $x^\varepsilon \in \bar{\Omega}_\varepsilon^+$.

3.2. *Strain-stress relation in the internal layer*

In the curvilinear coordinate system, the stress-strain relation (1.2) for the homogeneous isotropic internal layer can be written as

$$e^{ij} = A^{ijkl}\sigma_{kl}, \qquad (3.5)$$

where the fourth-order symmetric tensor A is represented by its contravariant components $A^{ijkl}\colon \bar{\Omega}^- \to \mathbb{R}$

$$A^{ijkl} = \frac{1+\nu}{2E}\left(g^{ik}g^{jl}+g^{jk}g^{il}\right) - \frac{\nu}{E}g^{ij}g^{kl}. \qquad (3.6)$$

Note the symmetries

$$A^{ijkl} = A^{jikl} = A^{klij},$$

and the relations

$$A^{\alpha\beta k3} = A^{\alpha 333} = 0.$$

It is classical that A is positive definite and uniform with respect to $x^\varepsilon \in \bar{\Omega}^-$. With our vectorial notation and the change of basis defined by the matrix $\mathbf{Q}$, the

following relations hold: There exist two constants $C_A > 0$ and $c_A > 0$ depending on Ω^-, such that

$$\begin{aligned}\boldsymbol{\tau} \cdot \mathbf{A}\boldsymbol{\tau} &= \tilde{\boldsymbol{\tau}} \cdot (\mathbf{Q})^T \mathbf{A}\mathbf{Q}\tilde{\boldsymbol{\tau}}, \\ &= \tilde{\boldsymbol{\tau}}_T \cdot (\mathbf{Q}_T)^T \mathbf{A}_T \mathbf{Q}_T \tilde{\boldsymbol{\tau}}_T + \boldsymbol{\tau}_N \cdot \mathbf{A}_N \boldsymbol{\tau}_N, \\ &\geq c_A(2\tilde{\tau}_{13}^2 + 2\tilde{\tau}_{23}^2 + \tilde{\tau}_{33}^2 + 2\tau_{13}^2 + 2\tau_{23}^2 + \tau_{33}^2), \end{aligned} \tag{3.7}$$

and

$$\begin{aligned}\boldsymbol{\sigma} \cdot \mathbf{A}\boldsymbol{\tau} &= \tilde{\boldsymbol{\sigma}} \cdot \mathbf{Q}^T \mathbf{A}\mathbf{Q}\tilde{\boldsymbol{\tau}}, \\ &= \tilde{\boldsymbol{\sigma}}_T \cdot (\mathbf{Q}_T)^T \mathbf{A}_T \mathbf{Q}_T \tilde{\boldsymbol{\sigma}}_T + \boldsymbol{\sigma}_N \cdot \mathbf{A}_N \boldsymbol{\tau}_N, \\ &\leq C_A(\tilde{\sigma}_{11}^2 + 2\tilde{\sigma}_{12}^2 + \tilde{\sigma}_{22}^2 + 2\sigma_{13}^2 + 2\sigma_{23}^2 + \sigma_{33}^2)^{1/2} \\ &\quad \times (\tilde{\tau}_{11}^2 + 2\tilde{\tau}_{12}^2 + \tilde{\tau}_{22}^2 + 2\tau_{13}^2 + 2\tau_{23}^2 + \tau_{33}^2)^{1/2}. \end{aligned} \tag{3.8}$$

for all $(\tau_{ij}), (\sigma_{ij}) \in \mathbf{S}$ and all $x^\varepsilon \in \bar{\Omega}^-$.

4. Mixed Variational Formulation in Curvilinear Coordinates

This section aims to give the mixed variational formulation of the elasticity problem (1.5) using the notation introduced in the preceding sections. Well-posedness is then proved thanks to Brezzi's theorem.

The unknowns of the mixed variational formulation of the problem expressed in curvilinear coordinates are:

- the vector field

$$\mathbf{u}^\varepsilon = (u_i^\varepsilon)\colon \bar{\Omega}_\varepsilon \to \mathbb{R}^3,$$

  where the three functions $u_i^\varepsilon\colon \bar{\Omega}_\varepsilon \to \mathbb{R}$ are the covariant components of the displacement field of the points of the shell;
- the symmetric tensor field

$$\sigma^\varepsilon = (\sigma_{ij}^\varepsilon)\colon \bar{\Omega}_\varepsilon \to \mathbb{R}^9,$$

  where the nine functions $\sigma_{ij}^\varepsilon\colon \bar{\Omega}_\varepsilon \to \mathbb{R}$ are the covariant components of the stress tensor.

In what follows $\mathbf{v}^+$ (respectively $\mathbf{v}^-$) denotes the restriction of $\mathbf{v}$ to Ω^+ (respectively Ω^-). Let us introduce some functional spaces, namely

$$\begin{aligned}\mathbf{V}^\varepsilon = \{&\mathbf{v},\ \mathbf{v}^- \in (H^1(\Omega^-))^3,\ \mathbf{v}^+ \in H^1(\Omega_\varepsilon^+)^3, \\ &\mathbf{v} = 0 \text{ on } \Gamma^- \cup \Gamma_l \cup \Gamma_{\mathbf{u}}^+,\ \mathbf{v}^- = \mathbf{v}^+ \text{ on } S\}.\end{aligned}$$

$\mathbf{V}^\varepsilon$ is the Hilbert space of admissible displacement fields compatible with the transition condition on S. It is equipped with the norm

$$||\mathbf{v}||_{1,\Omega_\varepsilon} = [||\mathbf{v}||_{1,\Omega_\varepsilon^+}^2 + ||\mathbf{v}||_{1,\Omega^-}^2]^{1/2}.$$

Also, $\Sigma^\varepsilon = \{\tau = (\tau_{ij}) \in (L^2(\Omega_\varepsilon))^9,\ \tau_{ij} = \tau_{ji}\}$ is the Hilbert space of stress tensors. It is equipped with the norm

$$||\tau||_{0,\Omega_\varepsilon} = \left[\sum_{i,j} ||\tau_{ij}||^2_{0,\Omega_\varepsilon}\right]^{1/2}.$$

We assume that the applied volumic force field is defined by its contravariant components, $f^i \mathbf{g}_i^\varepsilon$, and make the following assumption

$$\begin{aligned} &\mathbf{f} = (f^i) = 0, \quad \text{in } \Omega_\varepsilon^+, \\ &\mathbf{f} \in (L^2(\Omega^-))^3. \end{aligned}$$

From the equations of the strong formulation of the elasticity problem (1.5), one classically deduces the mixed variational formulation expressed in terms of the curvilinear coordinates x_i^ε of the reference configuration $\mathbf{\Psi}(\bar{\Omega}_\varepsilon)$. The unknowns $\mathbf{u}^\varepsilon$ and σ^ε satisfy:

$$\begin{cases} \mathbf{u}^\varepsilon \in \mathbf{V}^\varepsilon, \quad \sigma^\varepsilon \in \Sigma^\varepsilon, \\ \displaystyle\int_{\Omega^-} A^{ijkl}\sigma_{kl}^\varepsilon \tau_{ij}\sqrt{g}dx^\varepsilon + \int_{\Omega_\varepsilon^+} C^{ijkl,\varepsilon}\sigma_{kl}^\varepsilon \tau_{ij}\sqrt{g^\varepsilon}dx^\varepsilon \\ \quad = \displaystyle\int_{\Omega^-} G^{ijkl} e_{kl}(\mathbf{u}^\varepsilon)\tau_{ij}\sqrt{g}dx^\varepsilon + \int_{\Omega_\varepsilon^+} G^{ijkl,\varepsilon} e_{kl}^\varepsilon(\mathbf{u}^\varepsilon)\tau_{ij}\sqrt{g^\varepsilon}dx^\varepsilon, \quad \forall \tau \in \Sigma^\varepsilon, \\ \displaystyle\int_{\Omega^-} G^{ijkl} e_{kl}(\mathbf{v})\sigma_{ij}^\varepsilon \sqrt{g}dx^\varepsilon + \int_{\Omega_\varepsilon^+} G^{ijkl,\varepsilon} e_{kl}^\varepsilon(\mathbf{v})\sigma_{ij}^\varepsilon \sqrt{g^\varepsilon}dx^\varepsilon \\ \quad = \displaystyle\int_{\Omega^-} f^i v_i \sqrt{g}dx^\varepsilon, \quad \forall \mathbf{v} \in \mathbf{V}^\varepsilon. \end{cases}$$

Using vectorial notation, we define:

$$\begin{aligned} A^\varepsilon(\sigma,\tau) &= \int_{\Omega^-} [\tilde{\boldsymbol{\sigma}}_T \cdot (\mathbf{Q}_T)^T \mathbf{A}_T \mathbf{Q}_T \tilde{\boldsymbol{\tau}}_T + \boldsymbol{\sigma}_N \cdot \mathbf{A}_N \boldsymbol{\tau}_N]\sqrt{g}dx^\varepsilon \\ &\quad + \int_{\Omega_\varepsilon^+} [\tilde{\boldsymbol{\sigma}}_T \cdot (\mathbf{Q}_T^\varepsilon)^T \mathbf{C}_T^\varepsilon \mathbf{Q}_T^\varepsilon \tilde{\boldsymbol{\tau}}_T + \boldsymbol{\sigma}_N \cdot \mathbf{C}_N^\varepsilon \boldsymbol{\tau}_N]\sqrt{g^\varepsilon}dx^\varepsilon, \\ &= \int_{\Omega^-} [\tilde{\boldsymbol{\sigma}}_T \cdot (\mathbf{Q}_T)^T \mathbf{A}_T \mathbf{Q}_T \tilde{\boldsymbol{\tau}}_T + \boldsymbol{\sigma}_N \cdot \mathbf{A}_N \boldsymbol{\tau}_N]\sqrt{g}dx^\varepsilon \\ &\quad + \int_{\Omega_\varepsilon^+} \left[\frac{1}{s_\varepsilon + 2\mu_e\varepsilon}\tilde{\sigma}_{11}\tilde{\tau}_{11} + \frac{1}{\mu_e\varepsilon}\tilde{\sigma}_{12}\tilde{\tau}_{12} + \frac{1}{2\mu_e\varepsilon}\tilde{\sigma}_{22}\tilde{\tau}_{22}\right. \\ &\quad \left. + \frac{1}{2\mu_e\varepsilon}\boldsymbol{\sigma}_N \cdot \mathbf{G}_N^\varepsilon \boldsymbol{\tau}_N\right]\sqrt{g^\varepsilon}dx^\varepsilon, \\ B^\varepsilon(\mathbf{v},\tau) &= \int_{\Omega^-} [\tilde{\mathbf{e}}_T(\mathbf{v}) \cdot (\mathbf{Q}_T)^T \mathbf{G}_T \mathbf{Q}_T \tilde{\boldsymbol{\tau}}_T + \mathbf{e}_N(\mathbf{v}) \cdot \mathbf{G}_N \boldsymbol{\tau}_N]\sqrt{g}dx^\varepsilon \end{aligned}$$

$$
\begin{aligned}
&+\int_{\Omega_\varepsilon^+}[\tilde{\mathbf{e}}_T^\varepsilon(\mathbf{v})\cdot(\mathbf{Q}_T^\varepsilon)^T\mathbf{G}_T^\varepsilon\mathbf{Q}_T^\varepsilon\tilde{\boldsymbol{\tau}}_T+\mathbf{e}_N^\varepsilon(\mathbf{v})\cdot\mathbf{G}_N^\varepsilon\boldsymbol{\tau}_N]\sqrt{g^\varepsilon}dx^\varepsilon,\\
=&\int_{\Omega^-}[\tilde{\mathbf{e}}_T(\mathbf{v})\cdot\tilde{\boldsymbol{\tau}}_T+\mathbf{e}_N(\mathbf{v})\cdot\mathbf{G}_N\boldsymbol{\tau}_N]\sqrt{g}dx^\varepsilon\\
&+\int_{\Omega_\varepsilon^+}[\tilde{\mathbf{e}}_T^\varepsilon(\mathbf{v})\cdot\tilde{\boldsymbol{\tau}}_T+\mathbf{e}_N^\varepsilon(\mathbf{v})\cdot\mathbf{G}_N^\varepsilon\boldsymbol{\tau}_N]\sqrt{g^\varepsilon}dx^\varepsilon,\\
L(\mathbf{v})=&\int_{\Omega^-}f^iv_i\sqrt{g}dx^\varepsilon.
\end{aligned}
\tag{4.1}
$$

With this notation the variational mixed formulation reads

$$\mathbf{u}^\varepsilon\in\mathbf{V}^\varepsilon,\quad \sigma^\varepsilon\in\Sigma^\varepsilon, \tag{4.2}$$

$$A^\varepsilon(\sigma^\varepsilon,\tau)=B^\varepsilon(\mathbf{u}^\varepsilon,\tau),\quad\forall\tau\in\Sigma^\varepsilon, \tag{4.3}$$

$$B^\varepsilon(\mathbf{v},\sigma^\varepsilon)=L(\mathbf{v}),\quad\forall\mathbf{v}\in\mathbf{V}^\varepsilon. \tag{4.4}$$

and the following result holds.

Theorem 4.1. *There exists a unique solution* $(\mathbf{u}^\varepsilon,\sigma^\varepsilon)$ *to problem* (4.2)–(4.4). *Moreover, there exist two positive constants,* C_σ^ε *and* $C_{\mathbf{u}}^\varepsilon$ *depending on* ε *only such that*

$$
\begin{aligned}
&||\sigma^\varepsilon||_{0,\Omega^\varepsilon}\le C_\sigma^\varepsilon||\mathbf{f}||_{0,\Omega^-},\\
&||\mathbf{u}^\varepsilon||_{1,\Omega^\varepsilon}\le C_{\mathbf{u}}^\varepsilon||\mathbf{f}||_{0,\Omega^-}.
\end{aligned}
$$

Proof. It is a direct consequence of Brezzi's theorem [4] (also see Babuška and Aziz [1]). Let us first note that since $\boldsymbol{\sigma}=\mathbf{Q}^\varepsilon\tilde{\boldsymbol{\sigma}}$, since $x^\varepsilon\to\mathbf{Q}^\varepsilon(x^\varepsilon)$ is continuous on $\bar\Omega^\varepsilon$ and since $\mathbf{Q}^\varepsilon$ is invertible for all $x^\varepsilon\in\bar\Omega^\varepsilon$, there exist two constants $c^\varepsilon, C^\varepsilon>0$, such that

$$c^\varepsilon||\tilde\sigma||_{0,\Omega^\varepsilon}\le||\sigma||_{0,\Omega^\varepsilon}\le C^\varepsilon||\tilde\sigma||_{0,\Omega^\varepsilon},\quad\forall\sigma\in\Sigma^\varepsilon. \tag{4.5}$$

Since $x^\varepsilon\to g^\varepsilon(x^\varepsilon)$ is continuous on $\bar\Omega^\varepsilon$ and strictly positive, there exist two constants $g_0^\varepsilon, g_1^\varepsilon>0$, such that

$$g_0^\varepsilon\le g^\varepsilon\le g_1^\varepsilon,\quad\forall x^\varepsilon\in\bar\Omega^\varepsilon. \tag{4.6}$$

From (4.5), (4.6), (3.4) and (3.8), we deduce that the bilinear form $A^\varepsilon(\sigma,\tau)$ is continuous on $\Sigma^\varepsilon\times\Sigma^\varepsilon$. There exists a positive constant M_A^ε, such that

$$|A^\varepsilon(\sigma,\tau)|\le M_A^\varepsilon||\sigma||_{0,\Omega^\varepsilon}||\tau||_{0,\Omega^\varepsilon},\quad\forall\sigma,\ \tau\in\Sigma^\varepsilon.$$

Since $x^\varepsilon\to\Gamma_{ij}^{k,\varepsilon}$ is continuous on $\bar\Omega^\varepsilon$, it follows from (2.10) that there exists a constant $C^\varepsilon>0$, such that

$$||e^\varepsilon(\mathbf{v})||_{0,\Omega_\varepsilon}\le C^\varepsilon||\mathbf{v}||_{1,\Omega_\varepsilon},\quad\forall\mathbf{v}\in(H^1(\Omega_\varepsilon))^3. \tag{4.7}$$

We deduce from (4.5) to (4.7) and (2.8) that the bilinear form $B^\varepsilon(\mathbf{v},\tau)$ is continuous on $\mathbf{V}^\varepsilon\times\Sigma^\varepsilon$.

We deduce from (4.5), (4.6), (3.3) and (3.7) that there exists a constant $m_A^\varepsilon > 0$, such that

$$A^\varepsilon(\sigma,\sigma) \geq m_A^\varepsilon ||\sigma||^2_{0,\Omega^\varepsilon}, \quad \forall \sigma \in \Sigma^\varepsilon.$$

Eventually, the inf-sup condition

$$\inf_{\substack{\mathbf{v}\in\mathbf{V}^\varepsilon \\ ||\mathbf{v}||_{1,\Omega_\varepsilon}=1}} \sup_{\substack{\tau\in\Sigma^\varepsilon \\ ||\tau||_{0,\Omega_\varepsilon}=1}} B(\mathbf{v},\tau) > 0$$

follows essentially from Korn's inequality in curvilinear coordinates (see, for example, [8]). There exists a constant $C^\varepsilon = C^\varepsilon(\Omega_\varepsilon, \mathbf{\Psi}, \Gamma_l \cup \Gamma^- \cup \Gamma^+_{\mathbf{u}})$, such that

$$||\mathbf{v}||_{1,\Omega_\varepsilon} \leq C^\varepsilon ||e^\varepsilon(\mathbf{v})||_{0,\Omega_\varepsilon}, \quad \forall \mathbf{v} \in \mathbf{V}^\varepsilon.$$

This condition can be written as: there exists a constant $\beta^\varepsilon > 0$, such that

$$\sup_{\tau\in\Sigma^\varepsilon} \frac{B^\varepsilon(\mathbf{v},\tau)}{||\tau||_{0,\Omega^\varepsilon}} \geq \beta^\varepsilon ||\mathbf{v}||_{1,\Omega^\varepsilon}, \quad \forall \mathbf{v} \in V^\varepsilon,$$

and one then has the classical bounds:

$$||\sigma^\varepsilon||_{0,\Omega^\varepsilon} \leq \frac{1}{\beta^\varepsilon}\left(1 + \frac{M_A^\varepsilon}{m_A^\varepsilon}\right)||\mathbf{f}||_{0,\Omega^-},$$
$$||\mathbf{u}^\varepsilon||_{1,\Omega^\varepsilon} \leq \frac{M_A^\varepsilon}{(\beta^\varepsilon)^2}\left(1 + \frac{M_A^\varepsilon}{m_A^\varepsilon}\right)||\mathbf{f}||_{0,\Omega^-}. \qquad \square$$

5. Formulation over a Domain Independent of ε

Let us define the sets

$$\begin{aligned}
\Omega &= w \times]-\varepsilon_l, 1[,\\
\Omega^+ &= w \times]0, 1[,\\
\Omega^- &= w \times]-\varepsilon_l, 0[,\\
\Gamma_l^+ &= \gamma \times [0,1],\\
\Gamma_l^- &= \gamma \times [-\varepsilon_l, 0],\\
\Gamma^- &= w \times \{-\varepsilon_l\},\\
\Gamma^+_{\mathbf{u}} &= w_{\mathbf{u}} \times \{1\},\\
\Gamma^+_\sigma &= w_\sigma \times \{1\}.
\end{aligned}$$

Let $x = (x_i)$ denote a generic point in the set $\bar{\Omega}$, and let $\partial_i = \partial/\partial x_i$. With $x^\varepsilon \in \bar{\Omega}_\varepsilon$, we associate the point $x = (x_i) \in \bar{\Omega}$, defined by

$$\begin{aligned}
x_\alpha &= x^\varepsilon_\alpha \ (=y_\alpha),\\
x_3 &= x^\varepsilon_3 \quad \text{if } x^\varepsilon \in \Omega^-,\\
x_3 &= (x^\varepsilon_3/\varepsilon) \quad \text{if } x^\varepsilon \in \Omega^+_\varepsilon.
\end{aligned}$$

We thus have

$$\begin{aligned}
&\partial^\varepsilon_\alpha = \partial_\alpha, \\
&\partial^\varepsilon_3 = \partial_3 \qquad \text{if } x^\varepsilon \in \Omega^-, \\
&\partial^\varepsilon_3 = (\partial_3/\varepsilon) \quad \text{if } x^\varepsilon \in \Omega^+_\varepsilon.
\end{aligned}$$

The functions

$$g_{ij},\ g^{ij},\ g,\ \Gamma^p_{ij}\colon\ \bar\Omega^- \to \mathbb{R},$$

are not affected by the scaling. On the other hand, with these same functions defined on $\bar\Omega^+_\varepsilon$,

$$g^\varepsilon_{ij},\ g^{ij,\varepsilon},\ g^\varepsilon,\ \Gamma^{p,\varepsilon}_{ij}\colon\ \bar\Omega^+_\varepsilon \to \mathbb{R},$$

we associate the functions

$$g_{ij}(\varepsilon),\ g^{ij}(\varepsilon),\ g(\varepsilon),\ \Gamma^p_{ij}(\varepsilon)\colon\ \bar\Omega^+ \to \mathbb{R},$$

defined for all $x^\varepsilon \in \Omega^+_\varepsilon$ by

$$\begin{aligned}
g_{ij}(\varepsilon)(x) = g^\varepsilon_{ij}(x^\varepsilon), \quad g^{ij}(\varepsilon)(x) = g^{ij,\varepsilon}(x^\varepsilon), \\
g(\varepsilon)(x) = g^\varepsilon(x^\varepsilon), \quad \Gamma^p_{ij}(\varepsilon)(x) = \Gamma^{p,\varepsilon}_{ij}(x^\varepsilon).
\end{aligned} \tag{5.1}$$

With the unknowns $\mathbf{u}^\varepsilon\colon \bar\Omega_\varepsilon \to \mathbb{R}^3$ and $\sigma^\varepsilon\colon \bar\Omega_\varepsilon \to \mathbb{R}^9$ of problem (4.2)–(4.4), we associate the scaled unknowns $\mathbf{u}(\varepsilon)\colon \bar\Omega \to \mathbb{R}^3$ and $\sigma(\varepsilon)\colon \bar\Omega \to \mathbb{R}^9$, defined by

$$\begin{aligned}
&\mathbf{u}(\varepsilon)(x) = \mathbf{u}^\varepsilon(x^\varepsilon) \quad \forall x^\varepsilon \in \bar\Omega_\varepsilon, \\
&\sigma(\varepsilon)(x) = \sigma^\varepsilon(x^\varepsilon) \quad \forall x^\varepsilon \in \bar\Omega_\varepsilon.
\end{aligned}$$

With any vector field $\mathbf{v} = (v_i) \in H^1(\Omega^+)^3$, we associate the symmetric tensor $(e_{ij}(\varepsilon)(\mathbf{v})) \in (L^2(\Omega^+)^9)$, defined by

$$\begin{aligned}
e_{\alpha\beta}(\varepsilon)(\mathbf{v}) &= \frac{1}{2}(\partial_\alpha v_\beta + \partial_\beta v_\alpha) - \Gamma^p_{\alpha\beta}(\varepsilon) v_p, \\
e_{\alpha 3}(\varepsilon)(\mathbf{v}) &= \frac{1}{2}\left(\partial_\alpha v_3 + \frac{1}{\varepsilon}\partial_3 v_\alpha\right) - \Gamma^\rho_{\alpha 3}(\varepsilon) v_\rho, \\
e_{33}(\varepsilon)(\mathbf{v}) &= \frac{1}{\varepsilon}\partial_3 v_3.
\end{aligned}$$

Let us now introduce the functional spaces $\mathbf{V}$ and Σ:

$$\begin{aligned}
\mathbf{V} = \{&\mathbf{v},\ \mathbf{v}^- \in (H^1(\Omega^-))^3,\ \mathbf{v}^+ \in (H^1(\Omega^+))^3, \\
&\mathbf{v} = 0 \text{ on } \Gamma^- \cup \Gamma_l \cup \Gamma^+_{\mathbf{u}},\ \mathbf{v}^- = \mathbf{v}^+ \text{ on } S\}.
\end{aligned}$$

$\mathbf{V}$ is the Hilbert space of admissible displacement fields compatible with the transition condition on S. Also

$$\Sigma = \{\tau = (\tau_{ij}) \in L^2(\Omega)^9,\ \tau_{ij} = \tau_{ji}\}$$

is the Hilbert space of stress tensors.

Eventually, the following notations are used in the scaled variational mixed formulation.

$$
\begin{aligned}
A(\varepsilon)(\sigma,\tau) &= A^-(\sigma,\tau) + A^+(\varepsilon)(\sigma,\tau),\\
A^-(\sigma,\tau) &= \int_{\Omega^-} [\tilde{\boldsymbol{\sigma}}_T \cdot (\mathbf{Q}_T)^T \mathbf{A}_T \mathbf{Q}_T \tilde{\boldsymbol{\tau}}_T + \boldsymbol{\sigma}_N \cdot \mathbf{A}_N \boldsymbol{\tau}_N] \sqrt{g} dx,\\
A^+(\varepsilon)(\sigma,\tau) &= \int_{\Omega^+} \left[\frac{\varepsilon}{s(\varepsilon) + 2\mu_e \varepsilon} \tilde{\sigma}_{11} \tilde{\tau}_{11} + \frac{1}{\mu_e} \tilde{\sigma}_{12} \tilde{\tau}_{12} + \frac{1}{2\mu_e} \tilde{\sigma}_{22} \tilde{\tau}_{22} \right.\\
&\qquad \left. + \frac{1}{2\mu_e} \boldsymbol{\sigma}_N \cdot \mathbf{G}_N(\varepsilon) \boldsymbol{\tau}_N \right] \sqrt{g(\varepsilon)} dx,\\
B(\varepsilon)(\mathbf{v},\tau) &= B^-(\mathbf{v},\tau) + B^+(\varepsilon)(\mathbf{v},\tau),\\
B^-(\mathbf{v},\tau) &= \int_{\Omega^-} [\tilde{\mathbf{e}}_T(\mathbf{v}) \cdot \tilde{\boldsymbol{\tau}}_T + \mathbf{e}_N(\mathbf{v}) \cdot \mathbf{G}_N \boldsymbol{\tau}_N] \sqrt{g}\, dx,\\
B^+(\varepsilon)(\mathbf{v},\tau) &= \int_{\Omega^+} [\varepsilon \tilde{\mathbf{e}}_T(\varepsilon)(\mathbf{v}) \cdot \tilde{\boldsymbol{\tau}}_T + \varepsilon \mathbf{e}_N(\varepsilon)(\mathbf{v}) \cdot \mathbf{G}_N(\varepsilon) \boldsymbol{\tau}_N] \sqrt{g(\varepsilon)}\, dx,\\
L(\mathbf{v}) &= \int_{\Omega^-} f^i v_i \sqrt{g}\, dx.
\end{aligned}
$$

The scaled unknowns $\mathbf{u}(\varepsilon)$ and $\sigma(\varepsilon)$ solve the scaled variational mixed formulation, (5.2)–(5.4), now posed over the set Ω, and thus over a domain which is independent of ε,

$$\mathbf{u}(\varepsilon) \in \mathbf{V}, \quad \sigma(\varepsilon) \in \Sigma, \tag{5.2}$$

$$A(\varepsilon)(\sigma(\varepsilon),\tau) = B(\varepsilon)(\mathbf{u}(\varepsilon),\tau) \quad \forall \tau \in \Sigma, \tag{5.3}$$

$$B(\varepsilon)(\mathbf{v},\sigma(\varepsilon)) = L(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{V}. \tag{5.4}$$

In the following lemmas, we gather properties needed in the sequel concerning the behavior of different functions as $\varepsilon \to 0$. $||\cdot||_{0,\infty,\bar{\Omega}^+}$ denotes the usual norm of the space $\mathcal{C}^0(\bar{\Omega}^+)$. The constant ε_0 is defined in Sec. 2.1.

Lemma 5.1. *The functions $g_{ij}(\varepsilon)$, $g^{ij}(\varepsilon)$, $g(\varepsilon)$, $\Gamma^p_{ij}(\varepsilon)$ are defined as in* (5.1) *and the functions a_{ij}, a^{ij}, a, $\Gamma^\rho_{\alpha\beta}$, $b_{\alpha\beta}$, b^ρ_α are defined as in* (2.1)–(2.3). *All the functions $a_{ij}, \ldots, b^\rho_\alpha \in \mathcal{C}^0(\bar{\omega})$ are identified with functions in $\mathcal{C}^0(\bar{\Omega}^+)$. Then there exist constants $C > 0$ (all denoted by the same symbol) such that*

$$||g_{\alpha\beta}(\varepsilon) - a_{\alpha\beta}||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon, \tag{5.5}$$

$$||g^{\alpha\beta}(\varepsilon) - a^{\alpha\beta}||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon, \tag{5.6}$$

$$g_{i3}(\varepsilon) = g^{i3}(\varepsilon) = \delta_{i3}, \tag{5.7}$$

$$||g(\varepsilon) - a||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon, \tag{5.8}$$

$$||\Gamma^\rho_{\alpha\beta}(\varepsilon) - \Gamma^\rho_{\alpha\beta}||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon, \tag{5.9}$$

$$||\Gamma^3_{\alpha\beta}(\varepsilon) - b_{\alpha\beta}||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon, \tag{5.10}$$

$$||\Gamma^\rho_{\alpha 3}(\varepsilon) + b^\rho_\alpha||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon, \tag{5.11}$$

$$\Gamma^3_{\alpha 3}(\varepsilon) = \Gamma^p_{33}(\varepsilon) = 0. \tag{5.12}$$

Proof. The proof can be found in [9, Lemma 3.1] and completed in [10, Lemma 3.1]. The main argument is the fact that $\mathbf{g}_\alpha(\varepsilon) = \mathbf{a}_\alpha + \varepsilon x_3 \partial_\alpha \mathbf{a}_3$ and $\mathbf{g}_3(\varepsilon) = \mathbf{a}_3$. □

Lemma 5.2. *There exist constants g_0, g_1, such that*

$$0 < g_0 \leq g(\varepsilon) \leq g_1, \quad \forall \varepsilon \in]0, \varepsilon_0], \ \forall x \in \bar{\Omega}^+, \tag{5.13}$$

$$0 < g_0 \leq g \leq g_1, \quad \forall x \in \bar{\Omega}^-. \tag{5.14}$$

Proof. (5.14) follows from the continuity of the strictly positive function g on $\bar{\Omega}^-$. (5.13) follows from (2.5) and (5.8). □

Let us define the 6×6 matrix $\mathbf{G}(0)$ by

$$\mathbf{G}(0) = \left(\begin{array}{c|c} \mathbf{G}_T(0) & 0 \\ \hline 0 & \mathbf{G}_N(0) \end{array} \right),$$

where

$$\mathbf{G}_T(0) = \begin{pmatrix} a^{11}a^{11} & \sqrt{2}a^{11}a^{12} & a^{12}a^{12} \\ \sqrt{2}a^{11}a^{12} & a^{11}a^{22} + a^{12}a^{12} & \sqrt{2}a^{12}a^{22} \\ a^{12}a^{12} & \sqrt{2}a^{12}a^{22} & a^{22}a^{22} \end{pmatrix},$$

and

$$\mathbf{G}_N(0) = \begin{pmatrix} a^{11}a^{33} & a^{12}a^{33} & 0 \\ a^{12}a^{33} & a^{22}a^{33} & 0 \\ 0 & 0 & a^{33}a^{33} \end{pmatrix}.$$

From Lemma 5.1 we easily deduce that there exists a constant $C > 0$, such that

$$||(\mathbf{G}(\varepsilon))_{ij} - (\mathbf{G}(0))_{ij}||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon, \tag{5.15}$$

where the 6×6 matrix $\mathbf{G}(\varepsilon)$ is defined in a obvious way.

Lemma 5.3. *There exist two constants $c_G > 0$ and $C_G > 0$ independent of ε, such that*

$$\boldsymbol{\tau} \cdot \mathbf{G_N}(\varepsilon)\boldsymbol{\tau} \geq c_G ||\boldsymbol{\tau}||^2, \quad \forall \varepsilon \in [0, \varepsilon_0], \ \forall x \in \bar{\Omega}^+, \ \forall \boldsymbol{\tau} \in \mathbb{R}^3. \tag{5.16}$$

$$\boldsymbol{\tau} \cdot \mathbf{G_N}\boldsymbol{\tau} \geq c_G ||\boldsymbol{\tau}||^2, \quad \forall x \in \bar{\Omega}^-, \ \forall \boldsymbol{\tau} \in \mathbb{R}^3. \tag{5.17}$$

$$\boldsymbol{\sigma} \cdot \mathbf{G_N}(\varepsilon)\boldsymbol{\tau} \leq C_G ||\boldsymbol{\sigma}|| ||\boldsymbol{\tau}||, \quad \forall \varepsilon \in [0, \varepsilon_0], \ \forall x \in \bar{\Omega}^+, \ \forall \boldsymbol{\tau} \in \mathbb{R}^3. \tag{5.18}$$

$$\boldsymbol{\sigma} \cdot \mathbf{G_N}\boldsymbol{\tau} \leq C_G ||\boldsymbol{\sigma}|| ||\boldsymbol{\tau}||, \quad \forall x \in \bar{\Omega}^-, \ \forall \boldsymbol{\tau} \in \mathbb{R}^3. \tag{5.19}$$

Proof. We only detail the proof of (5.16). From (2.7) we deduce that for each $\varepsilon > 0$, there exists $c_G(\varepsilon) > 0$, such that

$$\boldsymbol{\tau} \cdot \mathbf{G_N}(\varepsilon)\boldsymbol{\tau} \geq c_G(\varepsilon) ||\boldsymbol{\tau}||^2,$$

for all $x \in \bar{\Omega}^+$ and all $\boldsymbol{\tau} \in \mathbb{R}^3$.

$\mathbf{G}_N(0)$ is clearly symmetric, positive definite and uniform with respect to $x \in \bar{\Omega}^+$. Therefore, there exists a constant $c_{G0} > 0$, such that

$$\boldsymbol{\tau} \cdot \mathbf{G_N}(0)\boldsymbol{\tau} \geq c_{G0}||\boldsymbol{\tau}||^2,$$

for all $x \in \bar{\Omega}^+$ and all $\boldsymbol{\tau} \in \mathbb{R}^3$.

The continuity of the mapping

$$(x, \varepsilon, \boldsymbol{\tau}) \in \bar{\Omega}^+ \times [0, \varepsilon_0] \times \mathcal{B} \to \boldsymbol{\tau} \cdot \mathbf{G}(\varepsilon)(x)\boldsymbol{\tau},$$

where $\mathcal{B} = \{\boldsymbol{\tau} \in \mathbb{R}^3, ||\boldsymbol{\tau}|| = 1\}$, and the compacity of the domain lead to the existence of a constant c_G, such that relation (5.16) holds for $0 \leq \varepsilon \leq \varepsilon_0$. □

Lemma 5.4. *There exists a constant $C > 0$, such that*

$$||s(\varepsilon) - s(0)||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon, \quad \forall \varepsilon > 0, \tag{5.20}$$

where $s(0) = \text{trace}(\mathbf{D}_T\mathbf{G}_T(0))$.

There exist two constants s_0 and s_1, such that

$$0 < s_0 \leq s(\varepsilon) \leq s_1, \quad \forall x \in \bar{\Omega}^+, \ \forall \varepsilon > 0. \tag{5.21}$$

Proof. The scaled matrices $\mathbf{H}_T(\varepsilon), \mathbf{P}_T(\varepsilon), \mathbf{S}(\varepsilon)$ are defined in an obvious way on $\bar{\Omega}^+$ for all $\varepsilon > 0$. Since

$$\begin{aligned} s(\varepsilon) &= \text{trace}(\mathbf{S}(\varepsilon)) = \text{trace}(\mathbf{P}_T(\varepsilon)^T\mathbf{D}_T\mathbf{P}_T(\varepsilon)) \\ &= \text{trace}(\mathbf{D}_T\mathbf{P}_T(\varepsilon)\mathbf{P}_T(\varepsilon)^T) = \text{trace}(\mathbf{D}_T\mathbf{G}_T(\varepsilon)), \end{aligned}$$

we deduce from (5.15) that

$$||s(\varepsilon) - \text{trace}(\mathbf{D}_T\mathbf{G}_T(0))||_{0,\infty,\bar{\Omega}^+} \leq C\varepsilon.$$

In order to infer (5.21), it remains to show that

$$s(0) = \text{trace}(\mathbf{D}_T\mathbf{G}_T(0)) > 0, \quad \forall x \in \bar{\Omega}^+.$$

As for $\mathbf{G}(0)$, $\mathbf{H}(0)$ is defined in an obvious way using the functions a_{ij}. $\mathbf{H}(0)$ is symmetric, positive definite and uniform with respect to $x \in \bar{\Omega}^+$. We proceed as in Sec. 3.1. There exists an invertible matrix $\mathbf{P}_0$, such that

$$\begin{aligned} \mathbf{P}_0^T\mathbf{H}_T(0)\mathbf{P}_0 &= \mathbf{I}, \\ \mathbf{P}_0^T\mathbf{D}_T\mathbf{P}_0 &= \text{diag}(s_0, 0, 0), \end{aligned}$$

with $s_0 > 0$, $\forall x \in \bar{\Omega}^+$. Since $\mathbf{G}_T(0)^{-1} = \mathbf{H}_T(0) = (\mathbf{P}_0\mathbf{P}_0^T)^{-1}$, it is clear that

$$s(0) = \text{trace}(\mathbf{D}_T\mathbf{G}_T(0)) = \text{trace}(\mathbf{D}_T\mathbf{P}_0\mathbf{P}_0^T) = \text{trace}(\mathbf{P}_0^T\mathbf{D}_T\mathbf{P}_0) = s_0.$$

□

6. Asymptotic Analysis

In this section, we establish our main result. The goal is to pass to the limit as $\varepsilon \to 0$ in the scaled variational mixed formulation (5.2)–(5.4), in order to derive the asymptotic formulation and obtain the announced boundary conditions on the surface S. This is achieved in two steps. In Sec. 6.1, we obtain several *a priori* estimations on the sequences, $(\mathbf{u}(\varepsilon))_{\varepsilon>0}$ and $(\sigma(\varepsilon))_{\varepsilon>0}$, presented in Lemma 6.1 through Lemma 6.4. All these estimations are then used in Sec. 6.2 in which we let $\varepsilon \to 0$ to obtain the limit formulation, which is presented in Theorem 6.6. Eventually, we show in Theorem 6.7 how the solution of the asymptotic problem can be explicitely computed in Ω^+ and deduce boundary conditions on S.

6.1. *A priori estimations on* $(\mathbf{u}(\varepsilon))_{\varepsilon>0}$ *and* $(\boldsymbol{\sigma}(\varepsilon))_{\varepsilon>0}$

Lemma 6.1. *Let* $(\mathbf{u}(\varepsilon),\ \sigma(\varepsilon))$ *be the solution to problem* (5.2)–(5.4). *There exist constants* $C_1, C_2 > 0$, *such that for all* $\varepsilon \in]0, \varepsilon_0]$,

$$\left[\sum_{\alpha,\beta} ||\tilde{e}_{\alpha\beta}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-}\right]^{1/2} \leq C_1 \left[\sum_{\alpha,\beta} ||\tilde{\sigma}_{\alpha\beta}(\varepsilon)||^2_{0,\Omega^-}\right]^{1/2} \tag{6.1}$$

and

$$\begin{aligned}&[2||e_{13}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-} + 2||e_{23}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-} + ||e_{33}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-}]^{1/2}\\ &\quad \leq C_2[2||\sigma_{13}(\varepsilon)||^2_{0,\Omega^-} + 2||\sigma_{23}(\varepsilon)||^2_{0,\Omega^-} + ||\sigma_{33}(\varepsilon)||^2_{0,\Omega^-}]^{1/2}.\end{aligned} \tag{6.2}$$

Proof. In (5.3), let us choose $\tau_{ij} = 0$ in Ω^+ and $\tilde{\tau}_{\alpha\beta} = \tilde{e}_{\alpha\beta}(\mathbf{u}(\varepsilon))$ in Ω^-.

$$\int_{\Omega^-} \tilde{\boldsymbol{\sigma}}_T(\varepsilon) \cdot (\mathbf{Q}_T)^T \mathbf{A}_T \mathbf{Q}_T \tilde{\mathbf{e}}_T(\mathbf{u}(\varepsilon)) \sqrt{g} dx = \int_{\Omega^-} \tilde{\mathbf{e}}_T(\mathbf{u}(\varepsilon)) \cdot \tilde{\mathbf{e}}_T(\mathbf{u}(\varepsilon)) \sqrt{g} dx.$$

Using (5.14), (3.8) and Cauchy–Schwarz's inequality we obtain

$$\begin{aligned}&\int_{\Omega^-} \tilde{\boldsymbol{\sigma}}_T(\varepsilon) \cdot (\mathbf{Q}_T)^T \mathbf{A}_T \mathbf{Q}_T \tilde{\mathbf{e}}_T(\mathbf{u}(\varepsilon)) \sqrt{g} dx\\ &\quad \leq \sqrt{g_1} C_A \left[\sum_{\alpha,\beta} ||\tilde{\sigma}_{\alpha\beta}(\varepsilon)||^2_{0,\Omega^-}\right]^{1/2} \left[\sum_{\alpha,\beta} ||\tilde{e}_{\alpha\beta}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-}\right]^{1/2}.\end{aligned}$$

With (5.14) and (3.8) we have

$$\int_{\Omega^-} \tilde{\mathbf{e}}_T(\varepsilon)(\mathbf{u}(\varepsilon)) \cdot \tilde{\mathbf{e}}_T(\varepsilon)(\mathbf{u}(\varepsilon)) \sqrt{g} dx \geq \sqrt{g_0} \sum_{\alpha,\beta} ||\tilde{e}_{\alpha\beta}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-}$$

and we conclude that the first inequality is verified.

The second inequality is proved in the same way choosing $\tau_{i3} = e_{i3}(\mathbf{u}(\varepsilon))$ in Ω^-, and using (5.17) and (5.19). □

Lemma 6.2. *Let* $(\mathbf{u}(\varepsilon),\ \sigma(\varepsilon))$ *be the solution to problem* (5.2)–(5.4). *There exist positive constants* $C_3,\ C_4,\ C_5,\ C_6$ *and* C_7, *such that for all* $\varepsilon \in]0, \varepsilon_0]$,

$$||\tilde{\sigma}_{\alpha\beta}(\varepsilon)||_{0,\Omega^-} \leq C_3, \tag{6.3}$$

$$||\sigma_{i3}(\varepsilon)||_{0,\Omega^-} \leq C_4, \tag{6.4}$$

$$||\tilde{\sigma}_{11}(\varepsilon)||_{0,\Omega^+} \leq C_5\sqrt{\frac{2\mu_e\varepsilon + s_1}{\varepsilon}}, \tag{6.5}$$

$$||\tilde{\sigma}_{\alpha 2}(\varepsilon)||_{0,\Omega^+} \leq C_6, \tag{6.6}$$

$$||\sigma_{i3}(\varepsilon)||_{0,\Omega^+} \leq C_7. \tag{6.7}$$

Proof. Let us choose $\tilde{\tau}_{\alpha\beta} = \tilde{\sigma}_{\alpha\beta}(\varepsilon)$, $\tilde{\tau}_{i3} = \tilde{\sigma}_{i3}(\varepsilon)$ in (5.3) and $\mathbf{v} = \mathbf{u}(\varepsilon)$ in (5.4). We obtain

$$A(\varepsilon)(\sigma(\varepsilon),\ \sigma(\varepsilon)) = L(\mathbf{u}(\varepsilon)).$$

From (5.13), (5.14), (3.7) and (5.21), we deduce

$$\begin{aligned} A(\varepsilon)(\sigma(\varepsilon),\ \sigma(\varepsilon)) \geq \sqrt{g_0}c_A\Bigg(& \sum_{\alpha\beta} ||\tilde{\sigma}_{\alpha\beta}(\varepsilon)||^2_{0,\Omega^-} \\ & + 2||\sigma_{13}(\varepsilon)||^2_{0,\Omega^-} + 2||\sigma_{23}(\varepsilon)||^2_{0,\Omega^-} + ||\sigma_{33}(\varepsilon)||^2_{0,\Omega^-}\Bigg) \\ & + \frac{\varepsilon}{2\mu_e\varepsilon + s_1}\sqrt{g_0}||\tilde{\sigma}_{11}(\varepsilon)||^2_{0,\Omega^+} \\ & + \frac{1}{2\mu_e}\sqrt{g_0}\big(2||\tilde{\sigma}_{12}(\varepsilon)||^2_{0,\Omega^+} + ||\tilde{\sigma}_{22}(\varepsilon)||^2_{0,\Omega^+}\big) \\ & + \frac{\hat{c}_G}{2\mu_e}\sqrt{g_0}\big(2||\sigma_{13}(\varepsilon)||^2_{0,\Omega^+} + 2||\sigma_{23}(\varepsilon)||^2_{0,\Omega^+} + ||\sigma_{33}(\varepsilon)||^2_{0,\Omega^+}\big). \end{aligned}$$

Cauchy–Schwarz's inequality gives

$$L(\mathbf{u}(\varepsilon)) \leq \sqrt{g_1}\left[\sum_i ||f^i||^2_{0,\Omega^-}\right]^{1/2}\left[\sum_i ||u_i(\varepsilon)||^2_{0,\Omega^-}\right]^{1/2}.$$

From the three-dimensional Korn inequality in curvilinear coordinates [8], we deduce that there exists a constant $C = C(\Omega^-, \mathbf{\Psi}, \Gamma_l^- \cup \Gamma^-) > 0$, such that

$$\left[\sum_i ||u_i(\varepsilon)||^2_{0,\Omega^-}\right]^{1/2} \leq C\left[\sum_{i,j} ||e_{ij}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-}\right]^{1/2}.$$

There exists an ε-independent constant $C_Q > 0$ (which is a norm of matrix Q on Ω^-), such that

$$\begin{aligned} \sum_{i,j} ||e_{ij}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-} \leq C_Q \sum_{\alpha,\beta} & ||\tilde{e}_{\alpha\beta}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-} + 2||e_{13}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-} \\ & + 2||e_{23}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-} + ||e_{33}(\mathbf{u}(\varepsilon))||^2_{0,\Omega^-}, \end{aligned}$$

and using Lemma 6.1, we obtain

$$\sum_i ||u_i(\varepsilon)||^2_{0,\Omega^-} \leq C^2C_1^2C_Q \sum_{\alpha,\beta} ||\tilde{\sigma}_{\alpha\beta}(\varepsilon)||^2_{0,\Omega^-} + C^2C_2^2\big[2||\sigma_{13}(\varepsilon)||^2_{0,\Omega^-} + 2||\sigma_{23}(\varepsilon)||^2_{0,\Omega^-} + ||\sigma_{33}(\varepsilon)||^2_{0,\Omega^-}\big].$$

This eventually leads to

$$\left[\sum_i ||u_i(\varepsilon)||^2_{0,\Omega^-}\right]^{1/2} \leq \sqrt{\max(C^2C_1^2C_Q, C^2C_2^2)}\left[\sum_{\alpha,\beta} ||\tilde{\sigma}_{\alpha\beta}(\varepsilon)||^2_{0,\Omega^-} + 2||\sigma_{13}(\varepsilon)||^2_{0,\Omega^-} + 2||\sigma_{23}(\varepsilon)||^2_{0,\Omega^-} + ||\sigma_{33}(\varepsilon)||^2_{0,\Omega^-}\right]^{1/2},$$

which completes the proof. □

It is worth noticing here the particular form of estimate (6.5) in the preceding lemma. This estimate is sufficient since in the limit process we will only use the fact that $\sqrt{\varepsilon}||\tilde{\sigma}_{11}(\varepsilon)||_{0,\Omega^+}$ is bounded as $\varepsilon \to 0$ (see the proof of Theorem 6.6 at the end of the paper).

Lemma 6.3. *Let* $(\mathbf{u}(\varepsilon),\ \sigma(\varepsilon))$ *be the solution to problem* (5.2)–(5.4). *There exist three constants* C_8, C_9 *and* $C_{10} > 0$, *such that*

$$\left[\sum_{\alpha,\beta} ||\tilde{e}_{\alpha\beta}(\varepsilon)(\mathbf{u}(\varepsilon))||^2_{0,\Omega^+}\right]^{1/2} \leq C_8\frac{1}{\varepsilon}\left[\left(\frac{\varepsilon}{2\mu_e\varepsilon + s_0}\right)^2 ||\tilde{\sigma}_{11}(\varepsilon)||^2_{0,\Omega^+} + \left(\frac{1}{2\mu_e}\right)^2 ||\tilde{\sigma}_{12}(\varepsilon)||^2_{0,\Omega^+} + \left(\frac{1}{2\mu_e}\right)^2 ||\tilde{\sigma}_{22}(\varepsilon)||^2_{0,\Omega^+}\right]^{1/2}, \tag{6.8}$$

$$\left[\sum_{\alpha} ||e_{\alpha 3}(\varepsilon)(\mathbf{u}(\varepsilon))||^2_{0,\Omega^+}\right]^{1/2} \leq C_9\frac{1}{\varepsilon}\left[\sum_{\alpha} ||\sigma_{\alpha 3}(\varepsilon)||^2_{0,\Omega^+}\right]^{1/2}, \tag{6.9}$$

$$||e_{33}(\varepsilon)(\mathbf{u}(\varepsilon))||_{0,\Omega^+} \leq C_{10}\frac{1}{\varepsilon}||\sigma_{33}(\varepsilon)||_{0,\Omega^+}. \tag{6.10}$$

Proof. In (5.3), let us choose successively:
$\tau_{ij} = 0$ in Ω^-, $\tilde{\tau}_{\alpha\beta} = \tilde{e}_{\alpha\beta}(\varepsilon)(\mathbf{u}(\varepsilon))$ and $\tau_{i3} = 0$ in Ω^+,
$\tau_{ij} = 0$ in Ω^-, $\tilde{\tau}_{\alpha\beta} = 0$, $\tau_{33} = 0$ and $\tau_{\alpha 3} = e_{\alpha 3}(\varepsilon)(\mathbf{u}(\varepsilon))$ in Ω^+,
$\tau_{ij} = 0$ in Ω^-, $\tilde{\tau}_{\alpha\beta} = 0$, $\tau_{\alpha 3} = 0$ and $\tau_{33} = e_{33}(\varepsilon)(\mathbf{u}(\varepsilon))$ in Ω^+. □

Lemma 6.4. *Let* $(\mathbf{u}(\varepsilon),\ \sigma(\varepsilon))$ *be the solution to problem* (5.2)–(5.4). *There exist constants* C_{11} *and* $C_{12} > 0$, *such that*

$$||\partial_3 u_3(\varepsilon)||_{0,\Omega^+} \leq C_{11}, \tag{6.11}$$

$$||\partial_3 u_\alpha(\varepsilon)||_{0,\Omega^+} \leq C_{12}. \tag{6.12}$$

Proof. Since $e_{33}(\varepsilon)(\mathbf{u}(\varepsilon)) = \frac{1}{\varepsilon}\partial_3 u_3(\varepsilon)$, we directly deduce from estimates (6.10) and (6.7) that

$$||\partial_3 u_3(\varepsilon)||_{0,\Omega^+} \leq C_{11}.$$

The following relation holds

$$||\partial_3 u_\alpha(\varepsilon)||^2_{0,\Omega^+} = \varepsilon ||\partial_3 u^\varepsilon_\alpha||^2_{0,\Omega^+_\varepsilon}.$$

It is possible to extend $\mathbf{u}^\varepsilon$ by 0 to the ε-independent domain $\Omega^+_{\varepsilon_0}$ and apply Korn's inequality in curvilinear coordinates [8]. We deduce

$$||\partial_3 u^\varepsilon_\alpha||^2_{0,\Omega^+_\varepsilon} \leq C \sum_{i,j} ||e^\varepsilon_{ij}(\mathbf{u}^\varepsilon)||^2_{0,\Omega^+_\varepsilon},$$

with

$$||e^\varepsilon_{ij}(\mathbf{u}^\varepsilon)||^2_{0,\Omega^+_\varepsilon} = \varepsilon ||e_{ij}(\varepsilon)(\mathbf{u}(\varepsilon))||^2_{0,\Omega^+}.$$

We therefore have

$$\begin{aligned} ||\partial_3 u_\alpha(\varepsilon)||^2_{0,\Omega^+} \leq\ & \hat{C}\varepsilon^2 \sum_{\alpha,\beta} ||\tilde{e}_{\alpha\beta}(\varepsilon)(\mathbf{u}(\varepsilon))||^2_{0,\Omega^+} \\ & + 2C\varepsilon^2 \sum_{\alpha} ||e_{\alpha 3}(\varepsilon)(\mathbf{u}(\varepsilon))||^2_{0,\Omega^+} \\ & + C\varepsilon^2 ||e_{33}(\varepsilon)(\mathbf{u}(\varepsilon))||^2_{0,\Omega^+}, \end{aligned}$$

and we conclude using Lemmas 6.3 and 6.2 in order to bound the righthand side of the previous inequality. □

6.2. *Asymptotic analysis as* $\varepsilon \to 0$

Let us introduce the functional spaces V_3, $\mathbf{V}^*$ and Σ^*:

$$V_3(\Omega^+) = \left\{ v \in L^2(\Omega^+),\ \frac{\partial v}{\partial x_3} \in L^2(\Omega^+),\ v = 0 \text{ on } \Gamma^+_l \cup \Gamma^+_{\mathbf{u}} \right\},$$

$$\begin{aligned} \mathbf{V}^* = \{ & \mathbf{v},\ \mathbf{v}^- \in (H^1(\Omega^-))^3,\ \mathbf{v}^+ \in (V_3(\Omega^+))^3, \\ & \mathbf{v} = 0 \text{ on } \Gamma^- \cup \Gamma_l \cup \Gamma^+_{\mathbf{u}},\ \mathbf{v}^- = \mathbf{v}^+ \text{ on } S \}. \end{aligned}$$

$V_3(\Omega^+)$ and $\mathbf{V}^*$ are Hilbert spaces with the norms

$$||v||_{V_3(\Omega^+)} = \left\| \frac{\partial v}{\partial x_3} \right\|_{0,\Omega^+},$$

$$||\mathbf{v}||_{\mathbf{V}^*} = \left[\sum_i ||v_i||^2_{1,\Omega^-} + \left\| \frac{\partial v_i}{\partial x_3} \right\|^2_{0,\Omega^+} \right]^{1/2}.$$

It is possible to define the trace $v_{|\partial\Omega^-} \in H^{1/2}(\partial\Omega^-) \subset L^2(\partial\Omega^-)$ of $v \in H^1(\Omega^-)$ on the boundary $\partial\Omega^-$ of Ω^-. The trace on $\partial\Omega^+$ of an element $v \in V_3(\Omega^+)$ can also be defined and particularly $v_{|S} \in L^2_{\rm loc}(S)$ (see Theorem B.2 of the Appendix B). Σ^* is the Hilbert space defined by

$$\Sigma^* = \{\tau = (\tau_{ij}),\ \tau_{ij} = \tau_{ji},\ \tau_{ij} \in L^2(\Omega) \text{ for } (i,j) \neq (1,1),\ \tau_{11} \in L^2(\Omega^-)\}.$$

The following notations are used in the limit scaled variational mixed formulation:

$$\begin{aligned}
A^*(\sigma,\tau) &= A^-(\sigma,\tau) + A^{*+}(\sigma,\tau),\\
A^{*+}(\sigma,\tau) &= \int_{\Omega^+} \left[\frac{1}{\mu_e}\tilde\sigma_{12}\tilde\tau_{12} + \frac{1}{2\mu_e}\tilde\sigma_{22}\tilde\tau_{22} + \frac{1}{2\mu_e}\boldsymbol\sigma_N \cdot \mathbf{G}_N(0)\boldsymbol\tau_N\right]\sqrt{a}dx,\\
B^*(\mathbf{v},\tau) &= B^-(\mathbf{v},\tau) + B^{*+}(\mathbf{v},\tau),\\
B^{*+}(\mathbf{v},\tau) &= \int_{\Omega^+} [(\partial_3\mathbf{v})_N \cdot \mathbf{G}_N(0)\boldsymbol\tau_N]\sqrt{a}dx,
\end{aligned}$$

where the vector $(\partial_3\mathbf{v})_N = \left(\frac{1}{\sqrt{2}}\partial_3 v_1, \frac{1}{\sqrt{2}}\partial_3 v_2, \partial_3 v_3\right)^T$. In the remaining part of this paper the arrows $\to$ and $\rightharpoonup$ denote strong and weak convergence as $\varepsilon \to 0$, respectively.

Lemma 6.5. *Let* $(\mathbf{u}(\varepsilon),\ \sigma(\varepsilon))$ *be the solution to the scaled variational mixed formulation* (5.2)–(5.4)*. Then, there exists a subsequence, still denoted by* $(\mathbf{u}(\varepsilon),\ \sigma(\varepsilon))$ *for convenience, and there exists* $(\mathbf{u}^*, \sigma^*) \in \mathbf{V}^* \times \Sigma^*$*, such that*

$$\tilde\sigma_{11}(\varepsilon) \rightharpoonup \tilde\sigma^*_{11} \quad \text{in } L^2(\Omega^-), \tag{6.13}$$

$$\tilde\sigma_{\alpha 2}(\varepsilon) \rightharpoonup \tilde\sigma^*_{\alpha 2} \quad \text{in } L^2(\Omega), \tag{6.14}$$

$$\sigma_{i3}(\varepsilon) \rightharpoonup \sigma^*_{i3} \quad \text{in } L^2(\Omega), \tag{6.15}$$

$$\mathbf{u}(\varepsilon) \rightharpoonup \mathbf{u}^* \quad \text{in } \mathbf{V}^*. \tag{6.16}$$

Proof. Points (6.13)–(6.15) are direct consequences of Lemma 6.2.

Let us prove (6.16). From (6.1) and (6.3), we deduce that $\tilde e_{\alpha\beta}(\mathbf{u}(\varepsilon))$ is bounded in $L^2(\Omega^-)$. From (6.2) and (6.4), we deduce that $e_{i3}(\mathbf{u}(\varepsilon))$ is bounded in $L^2(\Omega^-)$. Therefore, $e_{ij}(\mathbf{u}(\varepsilon))$ is bounded in $L^2(\Omega^-)$, and Korn's inequality (see [8]) applied on Ω^- yields to the boundedness of $u_i(\varepsilon)$ in $H^1(\Omega^-)$. From Lemma 6.4, we deduce that $u_i(\varepsilon)$ is bounded in $V_3(\Omega^+)$. Consequently, there exists a subsequence $u_i(\varepsilon) \rightharpoonup u_i^*$ in $H^1(\Omega^-) \cup V_3(\Omega^+)$.

Since $u_i(\varepsilon) = 0$ on $\Gamma^- \cup \Gamma_l \cup \Gamma^+_{\mathbf{u}}$, $u_i^* = 0$ on $\Gamma^- \cup \Gamma_l \cup \Gamma^+_{\mathbf{u}}$. Since $u_i^+(\varepsilon)(x_1,x_2,0) = u_i^-(\varepsilon)(x_1,x_2,0)$ in $L^2_{\rm loc}(S)$ and $u_i^-(\varepsilon)(x_1,x_2,0) \in H^{1/2}(S)$, we have that $u_i^+(\varepsilon)(x_1,x_2,0) = u_i^-(\varepsilon)(x_1,x_2,0)$ in $H^{1/2}(S)$ and therefore in $L^2(S)$. Thus, we obtain that $u_i^{*+} = u_i^{*-}$ a.e. on S and $\mathbf{u}^* \in \mathbf{V}^*$. □

Theorem 6.6. $(\mathbf{u}^*, \sigma^*)$ *solves the scaled mixed variational problem*:

$$\mathbf{u}^* \in \mathbf{V}^*, \quad \sigma^* \in \Sigma^*, \tag{6.17}$$

$$A^*(\sigma^*,\tau) = B^*(\mathbf{u}^*,\tau), \quad \forall \tau \in \Sigma, \tag{6.18}$$

$$B^*(\mathbf{v},\sigma^*) = L(\mathbf{v}), \quad \forall \mathbf{v} \in \mathbf{V}. \tag{6.19}$$

Proof. The result is obtained by passing to the limit as $\varepsilon \to 0$ in (5.2)–(5.4).

(i) The terms $A^-(\sigma(\varepsilon),\tau)$, $B^-(\mathbf{u}(\varepsilon),\tau)$ and $B^-(\mathbf{v},\sigma(\varepsilon))$:

Using Lemma 6.5, it is clear that

$$A^-(\sigma(\varepsilon),\tau) \to A^-(\sigma^*,\tau),$$
$$B^-(\mathbf{u}(\varepsilon),\tau) \to B^-(\mathbf{u}^*,\tau),$$
$$B^-(\mathbf{v},\sigma(\varepsilon)) \to B^-(\mathbf{v},\sigma^*).$$

(ii) The term $A^+(\varepsilon)(\sigma(\varepsilon),\tau)$:

From (5.8) (cf. Lemma 5.1), we know that $\sqrt{g(\varepsilon)} \to \sqrt{a}$ in $\mathcal{C}^0(\bar{\Omega}^+)$. From Lemma 6.2, we deduce that $\sqrt{\varepsilon}\tilde{\sigma}_{11}(\varepsilon)$ is bounded in $L^2(\Omega^+)$ for $0 < \varepsilon \leq \varepsilon_0$ and since $\frac{\sqrt{\varepsilon}}{s(\varepsilon)+2\mu_e\varepsilon} \to 0$ in $\mathcal{C}^0(\bar{\Omega}^+)$,

$$\int_{\Omega^+} \frac{\varepsilon}{s(\varepsilon)+2\mu_e\varepsilon}\tilde{\sigma}_{11}(\varepsilon)\tilde{\tau}_{11}\sqrt{g(\varepsilon)}\,dx \to 0.$$

Then, using (6.6), (6.7) (cf. Lemma 6.2) and (5.15), we conclude that

$$A^+(\varepsilon)(\sigma(\varepsilon),\tau) \to A^{*+}(\sigma^*,\tau).$$

(iii) The term $B^+(\varepsilon)(\mathbf{v},\sigma(\varepsilon))$:

(5.9) and (5.10) (cf. Lemma 5.1) lead to

$$\begin{aligned} e_{\alpha\beta}(\varepsilon)(\mathbf{v}) &= \frac{1}{2}(\partial_\alpha v_\beta + \partial_\beta v_\alpha) - \Gamma^p_{\alpha\beta}(\varepsilon)v_p \to \frac{1}{2}(\partial_\alpha v_\beta + \partial_\beta v_\alpha) - \Gamma^\sigma_{\alpha\beta}v_\sigma - b_{\alpha\beta}v_3 \\ &= e_{\alpha\beta}(0)(\mathbf{v}), \end{aligned}$$

in $L^2(\Omega^+)$ for all $\mathbf{v} \in (H^1(\Omega^+))^3$. Since $\sqrt{\varepsilon}\tilde{\sigma}_{11}(\varepsilon)$, $\tilde{\sigma}_{12}(\varepsilon)$ and $\tilde{\sigma}_{22}(\varepsilon)$ are bounded in $L^2(\Omega^+)$,

$$\int_{\Omega^+} \varepsilon\tilde{\mathbf{e}}_T(\varepsilon)(\mathbf{v}) \cdot \tilde{\boldsymbol{\sigma}}_T(\varepsilon)\sqrt{g(\varepsilon)}dx \to 0.$$

We recall that $e_{\alpha 3}(\varepsilon)(\mathbf{v}) = \frac{1}{2}(\partial_\alpha v_3 + \frac{1}{\varepsilon}\partial_3 v_\alpha) - \Gamma^\sigma_{\alpha 3}(\varepsilon)v_\sigma$. Using (5.11) (cf. Lemma 5.1), we deduce that

$$\varepsilon e_{\alpha 3}(\varepsilon)(\mathbf{v}) \to \frac{1}{2}\partial_3 v_\alpha,$$

in $L^2(\Omega^+)$ for all $\mathbf{v} \in (H^1(\Omega^+))^3$. We also have

$$\varepsilon e_{33}(\varepsilon)(\mathbf{v}) \to \partial_3 v_3,$$

in $L^2(\Omega^+)$ for all $\mathbf{v} \in (H^1(\Omega^+))^3$. Therefore, we conclude that

$$\int_{\Omega^+} \varepsilon\mathbf{e}_N(\varepsilon)(\mathbf{v}) \cdot \mathbf{G}_N(\varepsilon)\boldsymbol{\sigma}_N(\varepsilon)\sqrt{g(\varepsilon)}dx \to \int_{\Omega^+} (\partial_3\mathbf{v})_N \cdot \mathbf{G}_N(0)\boldsymbol{\sigma}^*_N\sqrt{a}dx,$$

and

$$B^+(\varepsilon)(\mathbf{v},\sigma(\varepsilon)) \to B^{*+}(\mathbf{v},\sigma^*).$$

(iv) The term $B^+(\varepsilon)(\mathbf{u}(\varepsilon),\tau)$:

Let us show that $\varepsilon\partial_\alpha u_i(\varepsilon) \rightharpoonup 0$ in $L^2(\Omega^+)$. From (6.5), (6.6) (cf. Lemma 6.2) and (6.8) (cf. Lemma 6.3), we deduce that $\varepsilon\tilde{e}_{\alpha\beta}(\varepsilon)(\mathbf{u}(\varepsilon))$ is bounded in $L^2(\Omega^+)$. Therefore, $\varepsilon e_{\alpha\beta}(\varepsilon)(\mathbf{u}(\varepsilon))$ is also bounded in $L^2(\Omega^+)$. Since $\Gamma^p_{\alpha\beta}(\varepsilon)$ is bounded in $\mathcal{C}^0(\bar{\Omega}^+)$ and $u_p(\varepsilon)$ is bounded in $L^2(\Omega^+)$, we deduce from

$$\varepsilon e_{\alpha\beta}(\varepsilon)(\mathbf{u}(\varepsilon)) = \varepsilon\left(\frac{1}{2}(\partial_\alpha u_\beta(\varepsilon) + \partial_\beta u_\alpha(\varepsilon)) - \Gamma^p_{\alpha\beta}(\varepsilon)u_p(\varepsilon)\right),$$

that $\varepsilon\partial_1 u_1(\varepsilon)$, $\varepsilon\partial_2 u_2(\varepsilon)$ and $\varepsilon(\partial_1 u_2(\varepsilon) + \partial_2 u_1(\varepsilon))$ are bounded in $L^2(\Omega^+)$. In the same way,

$$\varepsilon e_{\alpha 3}(\varepsilon)(\mathbf{u}(\varepsilon)) = \frac{1}{2}(\varepsilon\partial_\alpha u_3(\varepsilon) + \partial_3 u_\alpha(\varepsilon)) - \varepsilon\Gamma^\sigma_{\alpha 3}(\varepsilon)u_\sigma(\varepsilon)$$

is bounded in $L^2(\Omega^+)$ and since $\partial_3 u_i(\varepsilon)$ (cf. Lemma 6.4) and $\varepsilon\Gamma^\sigma_{\alpha 3}(\varepsilon)u_\sigma(\varepsilon)$ are bounded in $L^2(\Omega^+)$, this implies that $\varepsilon\partial_\alpha u_3(\varepsilon)$ is bounded in $L^2(\Omega^+)$. We then apply the *classical* Korn inequality to $e(\mathbf{u})$ on Ω^+ to obtain the boundedness of $\varepsilon\partial_1 u_2(\varepsilon)$ and $\varepsilon\partial_2 u_1(\varepsilon)$. To sum up, $\varepsilon\partial_j u_i(\varepsilon)$ is bounded in $L^2(\Omega^+)$.

Hence $\varepsilon u_i(\varepsilon)$ is bounded in $H^1(\Omega^+)$ and there exists a subsequence, still denoted by $\varepsilon u_i(\varepsilon)$, which converges weakly to some v_i in $H^1(\Omega^+)$. The trace of v_i on $\Gamma^+_{\mathbf{u}}$ is 0 since the trace of $u_i(\varepsilon)$ on $\Gamma^+_{\mathbf{u}}$ is 0. Moreover, $\varepsilon\partial_3 u_\alpha \to 0$ in $L^2(\Omega^+)$ and therefore $\partial_3 v_i = 0$ a.e in Ω^+. We conclude that $v_i = 0$ and that $\varepsilon\partial_\alpha u_i(\varepsilon) \rightharpoonup 0$ in $L^2(\Omega^+)$.

As a consequence $\varepsilon e_{\alpha\beta}(\varepsilon)(\mathbf{u}(\varepsilon)) \rightharpoonup 0$ and therefore $\varepsilon\tilde{e}_{\alpha\beta}(\varepsilon)(\mathbf{u}(\varepsilon)) \rightharpoonup 0$, $\varepsilon\big(\frac{1}{2}(\partial_\alpha u_3(\varepsilon) - \Gamma^\rho_{\alpha 3}(\varepsilon)u_\rho(\varepsilon)\big) \rightharpoonup 0$ in $L^2(\Omega^+)$. Eventually,

$$B^+(\varepsilon)(\mathbf{u}(\varepsilon),\tau) \to B^{*+}(\mathbf{u}^*,\tau). \qquad \square$$

Theorem 6.7. *In the domain Ω^+, the displacement field $\mathbf{u}^*$ is given by*

$$u^*_\alpha(x_1,x_2,x_3) = \frac{1}{\mu_e}\sigma^{*-}_{\alpha 3}(x_1,x_2,0)(x_3-1), \quad a.e\ in\ \Omega^+, \tag{6.20}$$

$$u^*_3(x_1,x_2,x_3) = \frac{1}{2\mu_e}\sigma^{*-}_{33}(x_1,x_2,0)(x_3-1) \quad a.e\ in\ \Omega^+. \tag{6.21}$$

Proof. In (6.18), let us choose $\tau = 0$ in Ω^-, $\tau_{\alpha\beta} = 0$ in Ω^+ and $\boldsymbol{\tau}_N = \frac{1}{\sqrt{a}}\mathbf{G}_N(0)^{-1}\big[\frac{1}{2\mu_e}\boldsymbol{\sigma}^*_N - (\partial_3\mathbf{u}^*)_N\big]$ in Ω^+. This leads to

$$\int_{\Omega^+}\left\|\frac{1}{2\mu_e}\boldsymbol{\sigma}^*_N - (\partial_3\mathbf{u}^*)_N\right\|^2 dx = 0,$$

that is to say,

$$\frac{1}{\mu_e}\sigma^*_{\alpha 3} - \partial_3 u^*_\alpha = 0, \quad \text{in } L^2(\Omega^+), \tag{6.22}$$

$$\frac{1}{2\mu_e}\sigma^*_{33} - \partial_3 u^*_3 = 0, \quad \text{in } L^2(\Omega^+). \tag{6.23}$$

In (6.19), let us choose $\mathbf{v} = 0$ in Ω^- and $\mathbf{v} \in (\mathcal{D}(\Omega^+))^3$ in Ω^+. Then

$$\int_{\Omega^+} [(\partial_3 \mathbf{v})_N \cdot \mathbf{G}_N(0) \boldsymbol{\sigma}_N^*] \sqrt{a} dx = 0 = -\int_{\Omega^+} [(\mathbf{v})_N \cdot \mathbf{G}_N(0) \partial_3 \boldsymbol{\sigma}_N^*] \sqrt{a} dx,$$

where the vector $(\mathbf{v})_N = \left(\frac{1}{\sqrt{2}} v_1, \frac{1}{\sqrt{2}} v_2, v_3\right)^T$.

It follows that

$$\partial_3 \boldsymbol{\sigma}_N^* = 0, \tag{6.24}$$

in $(\mathcal{D}'(\Omega^+))^3$ and therefore in $(L^2(\Omega^+))^3$.

From (6.22)–(6.24), we deduce that $\partial_3 \partial_3 u_i^* = 0$ in $L^2(\Omega^+)$. Since the trace of u_i^* on $\Gamma_{\mathbf{u}}^+$ is 0, we obtain

$$u_i^*(x_1, x_2, x_3) = c_i(x_3 - 1), \quad \text{a.e in } \Omega^+,$$

where

$$c_i = -u_i^{*+}(x_1, x_2, 0) = \partial_3 u_i^*(x_1, x_2, x_3). \tag{6.25}$$

From (6.22), (6.24) and (6.25), we deduce that the trace of σ_{i3}^+ on $\partial\Omega^+$ belongs to $L^2(\partial\Omega^+)$. Also

$$c_\alpha = \frac{1}{\mu_e} \sigma_{\alpha 3}^{*+}(x_1, x_2, 0) = -u_\alpha^{*+}(x_1, x_2, 0) \quad \text{in } L^2(S),$$
$$c_3 = \frac{1}{2\mu_e} \sigma_{33}^{*+}(x_1, x_2, 0) = -u_3^{*+}(x_1, x_2, 0) \quad \text{in } L^2(S).$$

It remains to be shown that $\sigma_{i3}^{*+} = \sigma_{i3}^{*-}$ on S.

We first show that $\boldsymbol{\sigma}_N^{*+} = 0$ on Γ_σ^+.

In (6.19) let us choose $\mathbf{v} \in \mathbf{K}$, such that $\mathbf{v} \in (H^1(\Omega^+))^3$, $\mathbf{v} = 0$ on $S \cup \Gamma_{\mathbf{u}}^+ \cup \Gamma_l^+$ and $\mathbf{v} = 0$ in Ω^-. We obtain using Green's formula

$$B^{*+}(\mathbf{v}, \sigma^*) = 0,$$
$$= -\int_{\Omega^+} (\mathbf{v})_N \cdot \mathbf{G}_N(0)(\partial_3 \boldsymbol{\sigma}_N^*) \sqrt{a} dx + \int_{\Gamma_\sigma^+} (\mathbf{v})_N \cdot \mathbf{G}_N(0) \boldsymbol{\sigma}_N^* \sqrt{a} dx.$$

Using (6.24) results in

$$\int_{\Gamma_\sigma^+} (\mathbf{v})_N \cdot \mathbf{G}_N(0) \boldsymbol{\sigma}_N^* \sqrt{a} dx = 0, \quad \forall \mathbf{v} \in (L^2(\Gamma_\sigma^+))^3,$$

which implies

$$\boldsymbol{\sigma}_N^{*+} = 0, \quad \text{in } (L^2(\Gamma_\sigma^+))^3. \tag{6.26}$$

Let us now transform Eq. (6.19),

$$B^-(\mathbf{v}, \sigma^*) + B^{*+}(\mathbf{v}, \sigma^*) = L(\mathbf{v}), \quad \forall \mathbf{v} \in \mathbf{K},$$

using Green's formula (and (6.26)) and going back to cartesian coordinates. This gives

$$B^{*+}(\mathbf{v},\sigma^*) = \int_{\Omega^+} [(\partial_3 \mathbf{v})_N \cdot \mathbf{G}_N(0)\boldsymbol{\sigma}_N^*]\sqrt{a}dx,$$
$$= -\int_{\Omega^+} (\mathbf{v})_N \cdot \mathbf{G}_N(0)(\partial_3 \boldsymbol{\sigma}_N^*)\sqrt{a}dx - \int_S (\mathbf{v})_N^+ \cdot \mathbf{G}_N(0)\boldsymbol{\sigma}_N^{*+}\sqrt{a}dx_1\,dx_2.$$

Using (6.24) results in

$$B^{*+}(\mathbf{v},\sigma^*) = -\int_S (\mathbf{v})_N^+ \cdot \mathbf{G}_N(0)\boldsymbol{\sigma}_N^{*+}\sqrt{a}dx_1 dx_2,$$

and going back to cartesian coordinates

$$B^{*+}(\mathbf{v},\sigma^*) = -\int_{\hat{S}} \hat{\mathbf{v}}^+ \cdot \hat{\sigma}^{*+}\mathbf{n}\,ds.$$

Moreover, we have

$$B^-(\mathbf{v},\tau) = \int_{\Omega^-} [\tilde{\mathbf{e}}_T(\mathbf{v}) \cdot \tilde{\boldsymbol{\sigma}}_T^* + \mathbf{e}_N(\mathbf{v}) \cdot \mathbf{G}_N\boldsymbol{\sigma}_N^*]\sqrt{g}dx,$$
$$= \int_{\hat{\Omega}^-} \hat{e}(\hat{\mathbf{v}})\colon \hat{\sigma}^*\,d\hat{x}.$$

Since $\mathrm{div}(\hat{\sigma}(\varepsilon)) = \hat{\mathbf{f}} \in (L^2(\hat{\Omega}^-))^3$, $\mathrm{div}(\hat{\sigma}^*)$ belongs to $(L^2(\hat{\Omega}^-))^3$ and $\hat{\sigma}^*$ belongs to $H(\mathrm{div},\hat{\Omega}^-)$ (see Appendix B). Therefore, we can define $\hat{\sigma}\mathbf{n}_{|\hat{S}} \in H^{-1/2}(\hat{S})$ and we have Green's formula

$$B^-(\mathbf{v},\tau) = -\int_{\hat{\Omega}^-} \mathrm{div}(\hat{\sigma}^*) \cdot \hat{\mathbf{v}}\,d\hat{x} + \langle \hat{\sigma}^{*-}\mathbf{n}, \hat{\mathbf{v}}^-\rangle_{(H^{-1/2}(\hat{S}))^3,(H^{1/2}(\hat{S}))^3}.$$

Eventually, since

$$L(\mathbf{v}) = \int_{\Omega^-} f^i v_i \sqrt{g}dx = \int_{\hat{\Omega}^-} \hat{\mathbf{f}} \cdot \hat{\mathbf{v}}\,d\hat{x},$$

we obtain

$$-\int_{\hat{S}} \hat{\mathbf{v}} \cdot \hat{\sigma}^{*+}\mathbf{n}\,ds \; + \langle \hat{\sigma}^{*-}\mathbf{n}, \hat{\mathbf{v}}\rangle_{(H^{-1/2}(\hat{S}))^3,(H^{1/2}(\hat{S}))^3} = 0, \quad \forall \hat{\mathbf{v}} \in (L^2(\hat{S}))^3.$$

Therefore, $\hat{\sigma}^{*-}\mathbf{n} = \hat{\sigma}^{*+}\mathbf{n}$ in $(H^{-1/2}(\hat{S}))^3$ but since $\hat{\sigma}^{*+}\mathbf{n} \in (L^2(\hat{S}))^3$ the equality holds in $(L^2(\hat{S}))^3$. In curvilinear coordinates this reads $\boldsymbol{\sigma}_N^{*+}(x_1,x_2,0) = \boldsymbol{\sigma}_N^{*-}(x_1,x_2,0)$ in $(L^2(S))^3$ and the proof is complete. □

To conclude, let us show that the limit displacement and stress tensor fields satisfy in $\hat{\Omega}^-$ the equation of the elasticity problem (1.6) announced in the introduction of the paper. The result is expressed in the cartesian coordinate system.

Theorem 6.8. $\hat{\mathbf{u}}^*$ *and* $\hat{\boldsymbol{\sigma}}^*$ *satisfy*:

$$\begin{cases} \operatorname{div}(\hat{\sigma}^*) + \hat{\mathbf{f}} = 0 & \text{a.e in } \hat{\Omega}^-, \\ \hat{\sigma}^* = \lambda \operatorname{trace}(\hat{e}(\hat{\mathbf{u}}^*))I + 2\mu\hat{e}(\hat{\mathbf{u}}^*) & \text{a.e in } \hat{\Omega}^-, \\ \hat{\mathbf{u}}^* = 0 & \text{a.e on } \hat{\Gamma}^- \cup \hat{\Gamma}_l^-, \\ \hat{\sigma}^*\mathbf{n} = -2\mu_e\hat{u}_n^*\mathbf{n} - \mu_e\hat{\mathbf{u}}_T^* & \text{a.e on } \hat{S}. \end{cases} \tag{6.27}$$

Proof. Since $\mathbf{u}^* \in \mathbf{V}^*$ it is clear that $\hat{\mathbf{u}}^* = 0$ a.e on $\hat{\Gamma}^- \cup \hat{\Gamma}_l^-$. Choosing $x_3 = 0$ in (6.20) and (6.21) of Theorem 6.7, we deduce that

$$\mu_e u_\alpha^* \mathbf{g}^\alpha + 2\mu_e u_3^* \mathbf{g}^3 = -\sigma_{\alpha 3}^* \mathbf{g}^\alpha - \sigma_{33}^* \mathbf{g}^3 \quad \text{a.e on } S,$$

which is exactly the boundary condition expected on $\hat{S}$ expressed in curvilinear coordinates. Let us now obtain the stress-strain compartment equation. Choosing $\tau^+ = 0$ in (6.18) of Theorem 6.6 leads to

$$A^-(\sigma^*, \tau) = B^-(\mathbf{u}^*, \tau).$$

Going back to cartesian coordinates, this equation reads

$$\int_{\hat{\Omega}^-} \hat{A}_{ijkl}\hat{\sigma}_{kl}^*\hat{\tau}_{ij}\, d\hat{x} = \int_{\hat{\Omega}^-} \hat{e}_{ij}(\hat{\mathbf{u}}^*)\hat{\tau}_{ij}\, d\hat{x},$$

where

$$\hat{A}_{ijkl} = \frac{1+\nu}{2E}(\delta_{ik}\delta_{jl} + \delta_{jk}\delta_{il}) - \frac{\nu}{E}\delta_{ij}\delta_{kl}.$$

Since this holds for all $\tau_{ij} = \tau_{ji} \in L^2(\Omega^-)$, we obtain that $\hat{A}_{ijkl}\hat{\sigma}_{kl}^* = \hat{e}_{ij}(\hat{\mathbf{u}}^*)$ a.e in $\hat{\Omega}^-$. This relation can also be written

$$\hat{e}(\hat{\mathbf{u}}^*) = \frac{1+\nu}{E}\hat{\sigma}^* - \frac{\nu}{E}\operatorname{trace}(\hat{\sigma}^*)I,$$

which is equivalent to

$$\hat{\sigma}^* = \lambda \operatorname{trace}(\hat{e}(\hat{\mathbf{u}}^*))I + 2\mu\hat{e}(\hat{\mathbf{u}}^*).$$

Eventually, in order to obtain the equilibrium equation, one may choose in (6.19) of Theorem 6.6 $\mathbf{v}$ such that $\mathbf{v}^+ = 0$, $\mathbf{v} \in (H^1(\Omega^-))^3$ and $\mathbf{v} = 0$ on $\Gamma^- \cup \Gamma_l^- \cup S$. □

It should be noted that this last problem is wellposed. One can easily deduce this by formulating a mixed variational formulation (in cartesian coordinates) and check that assumptions of Theorem 1.2, p. 47 of the book by Brezzi and Fortin [5] are satisfied.

Appendix A

In this first appendix, we recall a result concerning the simultaneous reduction of two quadratic forms.

Let $\mathbf{A}$ be a symmetric, positive definite $n \times n$ matrix and $\mathbf{B}$ a symmetric $n \times n$ matrix. Using the matrix $\mathbf{A}$, one can define the scalar product $(.,.)_{\mathbf{A}}$ on $\mathbb{R}^n$ by

$$(x, y)_{\mathbf{A}} = X^T\mathbf{A}Y, \quad \forall x, y \in \mathbb{R}^n,$$

where X and Y are the $n \times 1$ matrices of x and y in the canonical basis. We define the quadratic form $q_{\mathbf{B}}$ by

$$q_{\mathbf{B}}(x) = X^T \mathbf{B} X, \quad \forall x \in \mathbb{R}^n.$$

There exists a unique linear operator $f\colon \mathbb{R}^n \to \mathbb{R}^n$, which is symmetric for the scalar product $(.,.)_{\mathbf{A}}$, such that

$$q_{\mathbf{B}}(x) = (x, f(x))_{\mathbf{A}}, \quad \forall x \in \mathbb{R}^n.$$

Let $\mathbf{C}$ be the matrix of f in the canonical basis. We have

$$X^T \mathbf{B} X = X^T \mathbf{A}\mathbf{C} X, \quad \forall X \in \mathbb{R}^n$$

and therefore $\mathbf{AC} = \mathbf{B}$. Since $\mathbf{C}$ is the matrix of a symmetric linear operator, it is diagonalizable in a basis which is orthonormal with regard to the scalar product $(.,.)_{\mathbf{A}}$. Hence, there exist an invertible matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{D}$, such that

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \mathbf{I}_n, \tag{A.1}$$

$$\mathbf{P}^{-1} \mathbf{C} \mathbf{P} = \mathbf{D}. \tag{A.2}$$

From (A.1) we deduce that $\mathbf{A}^{-1} = \mathbf{P}\mathbf{P}^T$ and replacing $\mathbf{C}$ by $\mathbf{A}^{-1}\mathbf{B}$ in (A.2), we deduce that $\mathbf{P}^{-1}\mathbf{P}\mathbf{P}^T\mathbf{B}\mathbf{P} = \mathbf{D}$. To sum up, we have that

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \mathbf{I}_n \quad \text{and} \quad \mathbf{P}^T \mathbf{B} \mathbf{P} = \mathbf{D}.$$

Appendix B

In this appendix, we recall two traces theorems. Let Ω be a Lipschitz continuous open subset of $\mathbb{R}^3$. Let us define the Hilbert space $H(\mathrm{div}, \Omega)$ by

$$H(\mathrm{div}, \Omega) = \{\mathbf{v} \in (L^2(\Omega))^3;\ \mathrm{div}(\mathbf{v}) \in L^2(\Omega)\}.$$

Theorem B.1. *The mapping $\gamma_n\colon \mathbf{v} \to \mathbf{v} \cdot \mathbf{n}_{|\partial\Omega}$ is a linear continuous operator from $H(\mathrm{div}, \Omega)$ into $H^{-1/2}(\partial\Omega)$.*

For a proof the reader is referred to Theorem 2.5, p. 27 of the book by Girault and Raviart [11].

For $1 \le i \le 3$, let $a_i\colon \Omega \to \mathbb{R}$ be C^1 functions such that $\sum_{i=1}^3 \partial_i a_i$ is bounded. Let us define the Hilbert space H by

$$H = \left\{ \phi \in L^2(\Omega);\ \sum_{i=1}^3 a_i \partial_i \phi \in L^2(\Omega) \right\}.$$

The following result holds.

Theorem B.2. *Assume the functions a_i satisfy the previous hypothesis. Then for $S \subset \partial\Omega$ a part of the boundary of positive measure, the mapping $\gamma_S\colon \phi \to \phi_{|S}$ is a linear continuous operator from H into $L^2_{\mathrm{loc}}\left(S, \sum_{i=1}^3 a_i n_I \, d\sigma\right)$, where $\mathbf{n}$ is the outward normal.*

For a proof the reader is referred to Bardos [2], p. 205.

References

[1] I. Babuška and A. K. Aziz, Survey lectures on the mathematical foundations of the finite element method, in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, ed. A. K. Aziz (Academic Press, New York, 1979).

[2] C. Bardos, Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; application à l'équation de transport, *Ann. Scient. École Norm. Sup.* **3**(4) (1970) 185–233.

[3] G. Bayada, M. Chambat and K. Lhalouani, Asymptotic analysis of a thin-layer device with tresca's contact law in elasticity, *Math. Models Methods Appl. Sci.* **22**(10) (1999) 811–836.

[4] F. Brezzi, On the existence, uniqueness and the approximation of saddle points problems arising from Lagrangian multipliers, *RAIRO Anal. Numer.* **8** (1974) 129–151.

[5] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods* (Springer-Verlag, New York, 1991).

[6] D. Caillerie, A. Mourad and A. Raoult, Towards a fiber-based constitutive law for the myocardium, in *Modeling and Simulation for Computer Aided Medicine and Surgery (MS4CMS), ed. Thiriet*, ESAIM Proceedings, Vol. 12, 2002, pp. 25–30.

[7] P. G. Ciarlet, *Mathematical Elasticity, Vol. II*: *Plates and Shells* (North-Holland, 1996).

[8] P. G. Ciarlet, *Mathematical Elasticity, Vol. III: Theory of Shells* (North-Holland, 2000).

[9] P. G. Ciarlet and V. Lods, Asymptotic analysis of linearly elastic shells. I: Justification of membrane shell equations, *Arch. Rational Mech. Anal.* **136** (1996) 119–161.

[10] P. G. Ciarlet, V. Lods and B. Miara, Asymptotic analysis of linearly elastic shells. II: Justification of flexural shell equations, *Arch. Rational Mech. Anal.* **136** (1996) 163–190.

[11] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations. Theory and Algorithms,* Springer Series in Computational Mathematics (Springer-Verlag, 1986).

[12] F. Krasucki and S. Lenci, Yield design of bonded joints, *Eur. J. Mech. A Solids*, **19**(4) (2000) 649–667.

[13] P. Pebay, T. Baker and J. Pousin, Dynamic meshing for finite element based segmentation of cardiac imagery, in *Fifth World Congress on Computational Mechanics. Vienna*, 2002.

[14] Q. C. Pham, Segmentation et mise en correspondance en imagerie cardiaque multimodale conduites par un modèle anatomique bi-cavités du coeur. PhD thesis, Institut National Polytechnique de Grenoble (2002).

[15] Q. C. Pham, F. Vincent, P. Clarysse, J. Pousin, P. Croisille and K. Toivo, Spatio-temporal segmentation of the heart ventricles from MRI using a 3-D elastic active region model, *IEEE Trans. Medical Imaging*, submitted, 2002.

[16] F. Vincent, P. Clarysse, P. Croisille and I. E. Magnin, An elastic-based region model and its application to the estimation of the heart deformation in tagged MRI, in *ICIP-2000*, Vancouver, BC, Canada (2000), pp. 629–632.