Emulation statistique de champs de vent à haute résolution

Liyun Guelton

THÈSE / Télécom Bretagne
pour obtenir le grade de
Docteur de Télécom Bretagne
En accréditation conjointe avec l'Ecole Doctorale Sicma
Mention: Sciences et Technologies de l'Information et de la Communication

présentée par
Liyun He Guelton
préparée dans le département Signal et communications
Laboratoire Labsticc

Émulation statistique de champs de vent à haute résolution

Thèse soutenue le 20 juin 2014
devant le jury composé de :

Valérie Monbet
Professeure, Université de Rennes 1 / présidente

Sylvie Thiria
Professeure, Université Pierre et Marie Curie - Paris VI / rapporteur

Thomas Corpetti
Directeur de recherche, LETG-Rennes Costel / rapporteur

Pierre Ailliot
Maître de conférences, Université de Bretagne Occidentale / examinateur

Bertrand Chapron
Chercheur, Ifremer - Plouzané / examinateur

Jean Tournadre
Chercheur (HDR), Ifremer – Plouzané / directeur de thèse

Ronan Fablet
Professeur, Télécom Bretagne / invité

N° d'ordre : 2014telb0319

Sous le sceau de l'Université Européenne de Bretagne

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Sicma

Émulation statistique de champs de vent à haute résolution

Thèse de Doctorat

Mention : STIC

Présentée par **Liyun He Guelton**

Laboratoire : LOS / IFREMER
Laboratoire : Lab-STICC / SC / Télécom Bretagne

Directeur de thèse : Jean Tournadre

Soutenue : le 20 juin 2014

Jury :

Mme Sylvie Thiria, Professeur, LOCEAN / Université de Paris VI (Rapporteur)
M. Thomas Corpetti, Directeur de recherche, CNRS (Rapporteur)
Mme Valérie Monbet, Professeur, Université de Rennes 1 (Président)
M. Pierre Ailliot , Maitre de conférence, Université de Brest (Examinateur)
M. Jean Tournadre, HDR, Ifremer (Directeur de thèse)
M. Bertrand Chapron, Chercheur, Ifremer (Encadrant)
M. Ronan Fablet, Professeur, Télécom Bretagne (Encadrant)

À notre lumière et à notre joie,
Lucie et Laëtitia

Remerciements

J'adresse mes remerciements à toutes les personnes qui m'ont accompagnée tout au long de cette thèse. Une thèse, c'est un travail très personnel, mais dans un contexte collectif. J'ai beaucoup apprécié le temps passé dans mon laboratoire à l'Ifremer et j'adresse sincèrement mes remerciements à tout LOS pour m'avoir offert un cadre agréable pour effectuer ce double travail, à la fois sur mon sujet et sur soi-même.

L'aide de mes encadrants Bertrand Chapron et Ronan Fablet a été précieuse pour mener le sujet jusqu'au bout. Bertrand m'a fait découvrire le domaine de la géophysique et m'a accordé les moyens nécessaires à la réalisation de mes travaux. Merci Ronan pour ton exigence et pour ton aide dans le domaine de l'apprentissage statistique.

Je suis également très reconnaissante à Sylvie Thiria et Thomas Corpetti d'avoir accepté d'être les rapporteurs de cette thèse et de s'être investis dans la relecture du manuscrit, ainsi qu'à tous les autres membres de jury : Valérie Monbet, Pierre Ailliot. Merci Pierre pour ton aide tout au long de ma thèse, et pour les nombreux workshops auxquels tu m'as fait participer.

Je tiens à remercier mon directeur de thèse Jean Tournadre, et aussi Yves Quilfen, pour leur aide sur le manuscrit. Ils ont eu un vrai regard critique sur ce document et leurs remarques et corrections ont permi d'améliorer significativement la qualité du rapport final. Merci Emmanuelle Autret, ma collègue de bureau, fidèle interprète, compagne de tous les jours. Merci Olivier Archer pour ton aide technique, pour ton amitié, et bravo Mr Muscle ! Merci Abderrahim Bentamy pour tes conseils, merci Pierre Queffeulou pour les discussions scientifiques et culinaires. merci Arshad Rawad, c'est décidement toi qui amène le soleil de l'île Maurice jusqu'à la pointe du diable.

Un grand merci aux membres de l'administration, Marie-laure Quentel, Janick Vourch, Monique Larsonneur et Geneviève Larue de Télécom Bretagne, vous rendez facile ce qui nous semble insurmontable !

Merci à mes amis : la famille de Stéphanie Even et Ronan Keryell, Laurent Déjean, Adrien Martin, la famille de Caroline Fontaine et Yves Coudène, Hyunseuk Yoo, Aimée Johansen, Isa et Laurent, Peng Ying mon amie de lycée ; merci à tous mes voisins de l'allée des roses ; merci la famille de Paule Giscos et Marie-Hélène Dougnon.

Merci à mes parents et toute ma famille en Chine, et j'ai une pensée toute particulière pour ma grand-mère. 谢谢我的灰太郎还有我们两个可爱的狼宝宝。

Table des matières

Introduction

1

1.1 Positionnement du problème

Avec les avancées en technologies de télédétection, les observations sur l'océan et sur l'atmosphère ont fait beaucoup de progrès, à la fois en quantité et en qualité. Les satellites permettent d'observer de grandes zones et fournissent des observations avec une bonne résolution spatiale, de l'ordre du kilomètre ou mieux. L'augmentation de la quantité de données concernant les paramètres océaniques comme les vagues, la température de l'eau, les vents en surface de la mer, la couleur de l'eau ou les courants permet d'avoir une meilleure compréhension de la dynamique océanique et de pouvoir analyser les phénomènes physiques à petite échelle.

L'utilisation des observations satellite est de plus en plus fréquente pour décrire les dynamiques océaniques et atmosphériques de façon détaillée, mais leur couverture spatiale partielle les rend parfois inutilisables directement, et surtout leur résolution temporelle grossière et irrégulière fait qu'elles ne sont souvent pas disponibles. Les sorties des modèles numériques ont une résolution temporelle plus élevée et régulière mais leurs résolutions spatiales restent encore assez faibles. En prenant en compte les avantages et les inconvénients de l'une et l'autre, les techniques d'émulation statistique offrent une solution alternative pour obtenir des données « artificielles » avec une Haute Résolution (HR) spatio-temporelle. Le principe de ces techniques est d'émuler les données à HR spatiale à partir des données à Basse Résolution (BR) spatiale mais à HR temporelle, souvent issues des sorties d'un modèle numérique, et de « catalogues » qui consistent en des couples de données historiques co-localisées à basse et à haute résolution spatiale.

Dans ce contexte, le projet ANR GEO-Fluids, démarré en 2010, est dédié aux études et méthodes d'analyse d'écoulements fluides géophysiques. Il est piloté par Étienne Mémin de l'INRIA Rennes et regroupe les équipes INRIA, LMD, Ifremer (équipe Laboratoire d'Océanographie Spatiale (LOS)), *etc*. Cette thèse s'inscrit dans ce projet dans une démarche visant à associer les efforts récents théoriques et numériques de description des dynamiques océaniques et atmosphériques à petite échelle, avec la recherche de méthodologies avancées d'analyse d'observations satellite issues de différents capteurs. L'objectif final

de cette thèse est de développer des solutions originales et efficaces d'interprétation et d'interpolations spatiales utilisant les séries d'observations tout en respectant, si nécessaire et si possible, des contraintes dynamiques précises. Les observations disponibles peuvent caractériser la température, le champ de vent, la couleur de l'eau, *etc*. Les jeux de données sont différents, mais les méthodologies pour l'émulation de leur dynamique aux petites échelles (à HR) peuvent être similaires.

La première partie de cette étude concerne les applications des champs de vent à HR. Les données de vent à HR offrent des prospectives dans différents domaines pour des nombreuses applications. Elles permettent de :

- améliorer la représentation atmosphérique, y compris la visualisation de la circulation générale et la détermination de la vitesse des vents, notamment dans des zones pauvres en données observées, la localisation des fronts, la détection des zones de calme et de bascule de vent qui ne sont pas toujours bien vues par les modèles ;
- aider à mieux illustrer les effets locaux, y compris l'effet des montagnes, les différences de frottement et de turbulence près du sol, les effets de blocage de vent comme les accélérations et les décélérations ;
- aider à améliorer les modèles numériques : la compréhension de la dynamique à HR permet aux météorologues d'étendre et d'améliorer leurs modèles numériques en précision et en résolution, surtout dans les zones côtières où les modèles ont plus de difficultés à reproduire les effets locaux ;
- aider la gestion des risques : les analyses statistiques à HR permettent de mieux évaluer les risques en ingénierie marine, en pollution, en plan de sauvetage et de défense *etc* ;
- améliorer les stratégies commerciales. Par exemple, les vents à HR sont utiles pour aider à la production d'énergie, pour savoir où placer les fermes éoliennes ; les données à HR sont utiles pour les conception de navires à opérer dans l'environnement côtier ;
- aider à valider les modèles d'échange océan-atmosphère ;
- étudier les impacts du changement climatique.

Plus spécifiquement, cette étude consiste à émuler des vents à HR spatiale à partir d'un champ de vent à BR spatiale, issu d'une observation ou d'un modèle, en utilisant un apprentissage sur des situations observées par des mesures satellite à très HR spatiale. Les informations disponibles peuvent ainsi s'appuyer sur les associations historiques entre les champs de vent à BR et les champs de vent à HR. Dans cette étude, les données à BR utilisées proviennent du modèle European Center for Medium-range Weather Forecast (ECMWF) et celles à HR sont issues de mesures Radar à Synthèse d'Ouverture (SAR) du satellite ENVISAT, recueillies par la société Collecte Localisation Satellites (CLS). Ces deux types de données sont associés à une date et à une zone géographique pour former une donnée synchronisée. Les sorties du modèle ECMWF sont disponibles toutes les 6 heures (0 h, 06 h, 12 h, 18 h UTC). La mesure du vent

par SAR repose sur une approche diffusiométrique [Monaldo et al. (2003)] : le vent soufflant sur la surface de la mer crée de la rugosité, le radar émet des ondes vers la surface, et l'intensité des échos rétro-diffusés vers l'instrument est proportionnelle à la rugosité qui est elle-même reliée au vent local. La figure 1.1 donne un exemple d'un couple de données ECMWF et SAR co-localisées autour de la zone Bergen.

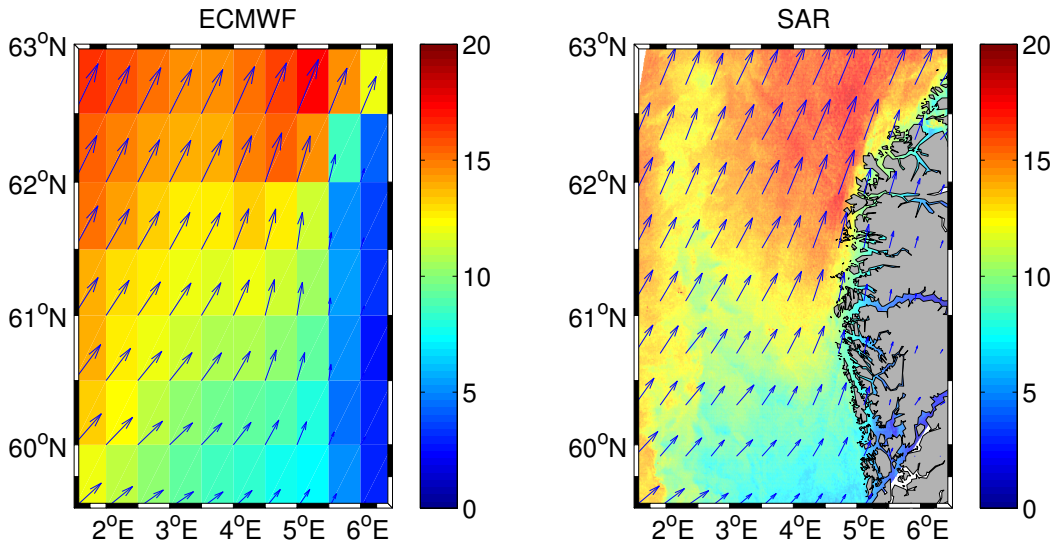
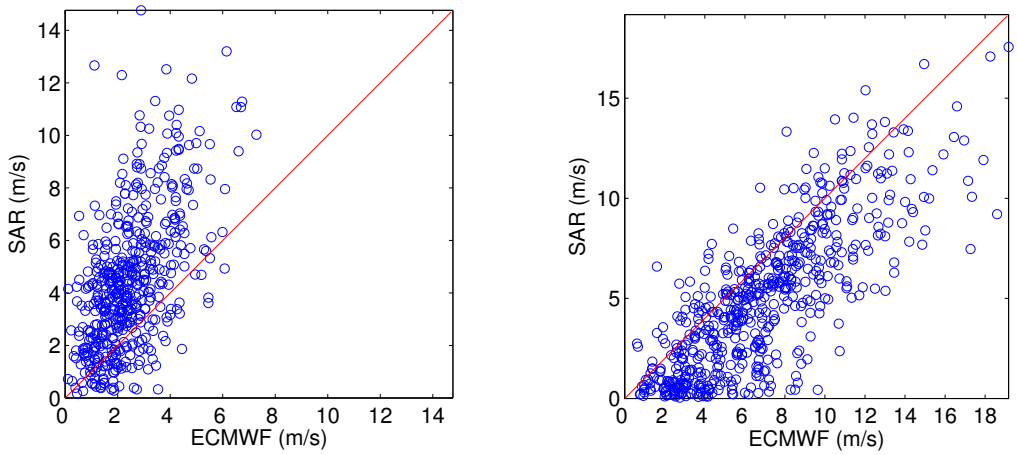


FIGURE 1.1 – *Exemple d'un couple des donnée ECMWF et SAR co-localisées autour de la zone Bergen. Le vent SAR a été acquis à 10h09 le 17 septembre 2005.*

L'objectif principal de l'émulation est de retrouver les dynamiques physiques associées aux petites échelles qui sont absentes à BR. Il existe deux types de dynamique aux petites échelles : les dynamiques conditionnées par les situations à grande échelle, comme celles des fronts et des effets d'accélération et de décélération, des modifications par la présence de continent, *etc.* et les dynamiques détachées de la grande échelle qui sont plus aléatoires. Nous cherchons à retrouver le premier type de dynamique au cours de l'émulation à HR. La figure 1.2 illustre les effets d'accélération (cf. Figure 1.2a) et de décélération (cf. Figure 1.2b) pour lesquels les vents observés par SAR sont systématiquement plus ou systématiquement moins forts que les vents du modèle numérique ECMWF.

Les dynamiques à petite échelle sont principalement liées aux positions géographiques et aux situations des vents. En particulier, par rapport au vent en pleine mer, le vent côtier est fortement influencé par l'orographie, la discontinuité entre la terre et l'océan, et le gradient thermique imposé par la différence de température entre la terre et la mer [Beaucage et al. (2007)]. Dans la zone côtière autour de Bergen, sur la côte ouest de la Norvège, les effets aux petites échelles sont amplifiés par la complexité de la topographie. Cette zone est formée de nombreuses îles, de nombreux *fjords* s'avançant dans les terres et de nombreuses montagnes très accidentées. Elle a été choisie comme cas d'étude en raison d'une part de cette complexité et son intérêt pour évaluer la



(a) *Point (N61.91°, E5.80°)* (b) *Point (N62.07°, E5.22°)*

FIGURE 1.2 – *Diagramme de dispersion de la vitesses des vents entre les données ECMWF et les données SAR : au point (N61.91°, E5.80°) (a) et au point (N62.07°, E5.22°) (b).*

pertinence de l'émulation HR, et d'autre part du fait de la disponibilité d'une base de données significatives d'observations HR SAR.

Les difficultés de l'émulation sont pour beaucoup liées aux imperfections des données et à la synchronisation temporelle entre elles. Si les données à basse et à haute résolution étaient parfaites, les données à BR ne seraient que la moyenne de la HR sur une certaine zone et les mesures à HR représenteraient les effets plus locaux. Dans ce cas, la mise en lien la basse et la haute résolution serait relativement simple. Malheureusement, ni les données du modèle ni les observations satellite ne sont parfaites. Les modèles numériques ont des difficultés pour simuler les zones complexes à cause de la paramétrisation et du côut de calcul. Les observations par télédétection sont connues pour leur amélioration de la résolution spatiale, mais elles sont limitées par la technique de mesure, par les interférences (pluie, présence d'objets autres que la cible, *etc.*), et par la difficulté des inversions des modèles semi-empiriques. L'autre problème de synchronisation, est lié à la différence de la résolution temporelle : ECMWF a une résolution temporelle régulière (toutes les 6 h) et l'heure de la mesure SAR dépend du passage satellite. Ces difficultés exigent une étape de préparation des données importante.

Pour que l'émulation statistique à HR soit performante, on recherche des approches robustes et adaptées qui permettent de lier les informations à BR et à HR. Ce lien est appelé la « fonction de transfert ». Elle doit prendre en compte la particularité des données aux deux résolutions, mais également les contraintes physiques de la zone d'étude. Il existe déjà de nombreux travaux basés sur les méthodes d'apprentissage statistique pour résoudre des problèmes de passage de résolution pour les paramètres météorologiques et géophysiques [Zorita et von Storch (1999), Walmsley et al. (2001), Minvielle (2009), Goubanova et al. (2010)]. Cependant, la fonction de transfert est souvent modélisée comme une fonction linéaire et elle est identique sur toute la zone d'étude. Or cette linéarité

n'est pas toujours établie, surtout dans le cas où les processus physiques sont complexes. Dans notre zone d'étude, les facteurs de modifications à la côte varient d'un point à l'autre et chaque lieu doit faire, si possible, l'objet d'une étude spécifique qui soit à la fois théorique et empirique [Mayençon (1982)]. Dans ce cadre, cette thèse se propose de mettre en œuvre une méthode de régression non-linéaire qui est optimale et plus robuste, Machine à Vecteurs de support pour la Régression (SVR) [Vapnik et al. (1997), Schölkopf et Smola (2001)], par point spécifique. Elle est comparée avec des modèles de référence de la littérature (e.g. modèle analogue [Zorita et von Storch (1999)] et régression linéaire globale [Goubanova et al. (2010)]).

1.2 Démarche générale

L'idée principale de cette thèse est d'associer les avancées récentes dans le domaine de la télédétection et dans le domaine de l'apprentissage statistique en un modèle original permettant l'émulation de vent à très HR. À cette fin, on met en place un système complet pour obtenir les émulations statistiques HR. La démarche générale considère les aspects suivants : l'analyse de l'état de l'art, la préparation des données, l'analyse et l'interprétation des relations entre les données BR et HR, la recherche des modèles d'émulation optimaux à HR et les évaluations expérimentales.

L'analyse de l'état de l'art permet de positionner les travaux de thèse et de proposer une approche d'apprentissage statistique plus performante et mieux adaptée à notre problème. Le chapitre 2 donne la formulation du problème général et un état de l'art des différentes méthodes statistiques existantes pour l'émulation à HR.

Pour la préparation des données, une première phase d'appairage de données consiste à synchroniser les données à basse et à haute résolution, mais la synchronisation de deux données à des résolutions spatio-temporelles différentes pose généralement des problèmes. Les solutions courantes sont des méthodes d'interpolation, pour interpoler les champs à BR aux heures des observations à HR. Cependant, des erreurs supplémentaires sont alors introduites entre basse et haute résolution. Le compromis entre la qualité et la quantité de données apporte une autre contrainte pour la sélection. Ces problèmes sont abordés et traités dans le chapitre 3.3. Une fois les données synchronisées, il reste à identifier et éliminer les données aberrantes.

Les analyses spatiales conjointes permettent de mieux connaître chaque source de données et leurs différences en les comparant les une aux autres afin d'obtenir des éléments interprétables. Elles permettent également d'identifier les contraintes physiques et d'aider à définir les variables explicatives pertinentes. Ceci est abordé dans le chapitre 4.

Ensuite, nous recherchons un modèle générique et optimal qui permette de mettre en relation les basses et hautes résolutions. L'aspect générique signifie

que le modèle conçu dans l'émulateur est indépendant de la nouvelle situation d'entrée et qu'il n'est appris qu'une seule fois. Le coût calculatoire de ces méthodes par apprentissage est principalement un coût « hors ligne », lié à la calibration des modèles. L'aspect optimal signifie que le modèle conçu a de meilleures capacités prédictives par rapport aux autres méthodes existantes et qu'il est robuste, dans le sens où il est moins sensible aux variables explicatives utilisées.

L'apprentissage du modèle engendre deux problèmes : la définition des variables explicatives et la définition des méthodes d'apprentissage de la fonction de transfert. Pour l'apprentissage, compte tenu des résultats des analyses spatiales, les questions suivantes peuvent se poser pour la recherche des solutions : faut-il utiliser une seule fonction de transfert sur toute la zone d'étude (modèle global) ou une fonction de transfert par point spécifique (modèle local) ? Quelle est le type de relation intrinsèque entre la basse et la haute résolution ? Régression linéaire ou non linéaire ? Ces questions orientent l'analyse des données et le choix des modèles. La définition des variables explicatives permet de proposer les variables les plus pertinentes possibles pour la prédiction à HR. Cette étape peut être appliquée indépendamment du choix des méthodes d'apprentissage à l'aide de l'analyse des données. Elle peut aussi être appliquée selon les performances d'une méthode d'apprentissage. Le chapitre 5 introduit de manière détaillée les différents modèles d'apprentissage statistique envisagés.

L'étape de validation expérimentale permet d'évaluer et de comparer les différents modèles d'émulation, pour pouvoir finalement proposer un modèle adapté à notre problème. Les évaluations du modèle proposé peuvent être faites sur une série de points caractéristiques. Les cas particuliers permettent d'utiliser les situations concrètes pour illustrer les performances de l'émulateur. Les propriétés statistiques sur l'ensemble des exemples émulés dans le catalogue permettent de voir la cohérence et la robustesse de l'émulateur. Cette dernière étape est présentée dans le chapitre 6.

Méthodologie et état de l'art 2

Ce chapitre introduit la formulation du problème et l'état de l'art des techniques de *downscaling* statistique. Cela permet de positionner ces travaux et de proposer une nouvelle approche plus performante et mieux adaptée à notre problème. Dans cette thèse, les modèles de *downscaling* considérés sont représentés de manière cohérente et unifiée dans la formalisme de Machine à Vecteurs de support pour la Régression (SVR).

Les modèles SVR sont proposés pour leurs propriétés d'optimisation et de généralisation par rapport aux méthodes de régression plus classiques, comme par exemple les méthodes analogues et la méthode de Régression Linéaire Multiple (MLR).

2.1 Formulation du problème

Pour une situation à BR donnée, on cherche à estimer la situation à HR qui aurait été la même que celle mesurée par SAR si une mesure avait été faite. Notons X_{BR} et Y_{HR} respectivement les situations à BR et à HR attendue. L'objectif de l'émulation à HR peut s'écrire comme suit :

$$Y_{\mathrm{HR}} = f(X_{\mathrm{BR}}) \tag{2.1}$$

où f est la fonction de transfert (ou fonction de régression).

La fonction de transfert est déterminée à partir d'une base de données d'apprentissage qui est un ensemble de paires de BR et HR permettant de fournir une série de couples de réalisations de X_{BR} et Y_{HR} — $(x_i, y_i)_{1 \leq i \leq n}$. n est le nombre de réalisation. Le but est d'obtenir une description de la relation sous-jacente entre X_{BR} et Y_{HR}. Une technique courante pour trouver la fonction de transfert f optimale, notée f^*, est de minimiser les erreurs de prédiction entre les estimations $f(X_{\mathrm{BR}})$ et les Y_{HR} observés pour l'ensemble des données d'apprentissage (x_i, y_i).

2.1.1 Variable expliquée

La variable de sortie Y désigne la variable à prédire, la réponse ou la variable expliquée. Ici, nous considérons des vecteurs vent de composantes zonale et méridienne u et v. Chaque composante est émulée séparément et reconstruite à HR. Chaque point sur la grille à HR fait l'objet d'une émulation

indépendante : on est donc dans le cas de l'univariée ($Y \subset \mathbb{R}^1$). La réponse Y est alors composée d'une seule colonne d'échantillons de la composante zonale ou méridienne en un point de la grille à HR.

2.1.2 Variables explicatives

Supposons qu'il y ait N couples d'entrée-sortie (ou échantillons) (x_i, y_i), chaque échantillon x_i est composé d'un vecteur de variables explicatives appartenant à $\mathbb{R}^d$ et s'écrit :

$$\begin{pmatrix} X^1 & X^2 & \cdots & X^d \end{pmatrix}$$

Chaque colonne de X^j désigne toutes les valeurs de la variable explicative j pour l'ensemble des échantillons. Notons $\mathbf{X}$ la matrice de dimension $N \times d$, où N est le nombre d'échantillons et d est le nombre de variables explicatives.

La figure 2.1 illustre un exemple de variables explicatives données par les informations à BR autour du point d'intérêt. Par convention, les coordonnées d'un point sur la grille à HR sont notées (p, q) et les coordonnées du point à BR le plus proche en distance du (p, q) sont notées (p', q').

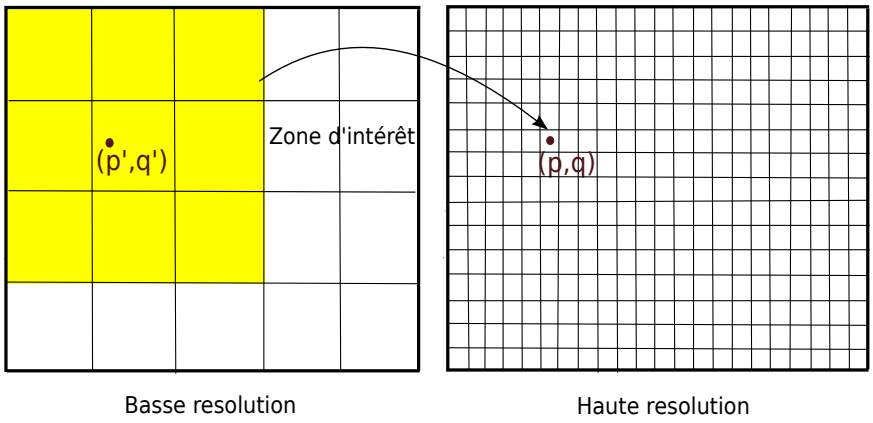


FIGURE 2.1 – *Fenêtre des informations à BR autour du point (p', q'). (p, q) est le point émulé sur la grille à HR. La boîte de couleur jeune correspond à une fenêtre de taille 1 centrée sur le point (p', q').*

Supposons que les informations à BR au point (p', q') et aux 8 points les plus proches sont utilisées. Les variables explicatives du i^{e} échantillon $x_i(p, q)$ peuvent alors s'écrire :

$$\begin{aligned}(\, z_i(p'-1, q'-1)\,, z_i(p', q'-1)\,, z_i(p'+1, q'-1),\\ z_i(p'-1, q')\,, z_i(p', q')\,, z_i(p'+1, q'),\\ z_i(p'-1, q'+1)\,, z_i(p', q'+1)\,, z_i(p'+1, q'+1)\,)\end{aligned}$$

où $z_i(p', q')$ est la i^{e} BR au point (p', q') et les autres z_i la BR aux points voisins.

L'information à BR z_i au point (p', q') est un vecteur. Par exemple, si on considère les deux composantes u et v, on obtient $z = (u, v)$. Le nombre de variables explicatives est égal au nombre de points utilisés multiplié par la

dimension du vecteur z qui est le nombre d'informations utilisées en chaque point.

2.1.3 Normalisation des variables explicatives

Pour éviter de privilégier certaines variables explicatives, la normalisation permet de ramener toutes les variables explicatives sur la même échelle. Cela est utile et presque obligatoire quand les variables n'ont pas la même unité. Un autre avantage de normaliser les variables est que cela permet ensuite de relativiser plus facilement l'importance que l'on veut donner à chaque variable.

Il existe différentes méthodes de normalisation. La normalisation la plus simple est de redimensionner les valeurs de chaque variable explicative X^k sur le même intervalle fermé $[a\ b]$:

$$x_i^k = (x_i^k - \min(X^k)) \frac{b-a}{\max(X^k) - \min(X^k)} + a \tag{2.2}$$

où $\min(X^k)$ et $\max(X^k)$ sont respectivement la valeur minimale et la valeur maximale pour la k^e variable explicative.

2.2 État de l'art

2.2.1 *Downscaling* statistique

Dans le domaine de l'océanographie et de la météorologie, le problème de recherche d'une fonction de transfert entre la basse et la haute résolution rentre dans la catégorie du *downscaling*. En effet, le *downscaling* est une technique de réduction d'échelle : c'est un processus qui permet de mettre en relation les informations climatiques à grande échelle et celles à petite échelle pour résoudre des problèmes de changement d'échelle [Wilby et Wigley (1997), von Storch et al. (2000), Wilby et al. (2004), Biaou, Angelbert (2004)]. Il est parfois aussi appelé méthode de « désagrégation ». Dans le domaine du traitement d'image, on le nomme la technique de « super résolution ».

Il existe deux types de méthodes de *downscaling* : le *downscaling* dynamique et le *downscaling* statistique. Le *downscaling* dynamique utilise la circulation générale comme condition limite. Il est par exemple utilisé dans la prévision météorologique opérationnelle, où les propriétés atmosphériques sont calculées sur une grille plus fine pour résoudre les équations de mouvement et de thermodynamique [Spak et al. (2007)], contrairement au *downscaling* statistique qui est basé sur une modélisation empirique entre les informations à grande échelle et celles à petite échelle. La démarche privilégiée dans cette thèse entre dans cette deuxième catégorie.

Le *downscaling* statistique est notamment utilisé pour obtenir le changement climatique à l'échelle régionale à partir de la circulation générale et dans les applications de la prévision de la précipitation à HR [Wilby et al. (2004),

Wilby et Wigley (1997), Zorita et al. (1995)]. Il existe différentes méthodes de *downscaling* statistique : classification météorologique [Kidson et Thompson (1998), Corte-Real et al. (1999), Cassou et al. (2011)] ; modèle de régression [Huth (1999), Walmsley et al. (2001), Cheng et al. (2008), Goubanova et al. (2010)] et générateurs météorologiques [Katz (1996), Wilks (1999)]. Vautard (1990) utilisent de multiples régimes de temps pour l'analyse des précurseurs et successeurs. Walmsley et al. (2001) utilisent et comparent les méthodes de MLR et d'arbre de régression pour prédire le vent *offshore* à partir des paramètres météorologiques le long des côtes du Danemark. Rooy et Kok (2004) combinent des méthodes physiques et de régression pour obtenir les vents à terre.

Le choix des approches de *downscaling* statistique est motivé par plusieurs éléments. Un des plus grands avantages du *downscaling* statistique est qu'il est particulièrement utile pour un environnement hétérogène avec une géographie complexe ou avec un fort gradient environnemental tel que dans une île, en montagne ou dans un contexte continent/mer [Wilby et al. (2004)], où les processus physiques sont difficiles à modéliser directement [Christensen et al. (2007)]. Pour ce dernier type de configuration, Benestad et al. (2008) a montré que les méthodes statistiques constituent une approche pragmatique pour modéliser les paramètres locaux à partir des informations climatiques à grande échelle.

Le faible côut de mise en œuvre d'une méthode statistique est un autre avantage du *downscaling* statistique. Par rapport au *downscaling* dynamique, le coût de calcul du *downscaling* statistique est en effet beaucoup moins important. Le *downscaling* statistique est souvent composé de deux phases : la phase d'apprentissage et la phase de prédiction. La phase d'apprentissage peut nécessiter des coûts calculatoires importants, mais il s'agit généralement d'une étape réalisée en amont « hors-ligne ». Une fois la fonction de transfert apprise, le temps de calcul est généralement négligeable pendant la phase de prédiction.

Cependant, le *downscaling* statistique a plusieurs contraintes comparativement à la résolution numérique des dynamiques géophysiques en jeu. La qualité de la méthode dépend de la qualité et de la représentativité de la base de données qui sert à l'apprentissage. Les méthodes statistiques supposent que la variable à prédire et les variables explicatives ont une relation stationnaire qui n'évolue pas avec le temps. Cette condition peut ne pas être vérifiée dans un contexte de changement climatique rapide, par exemple.

2.2.2 Apprentissage statistique

L'apprentissage statistique est défini comme une technique où l'on cherche à déduire les dépendances fonctionnelles (régularités) à partir d'un ensemble d'exemples d'apprentissage, tels que des paires de données, de motifs, d'observations [Kecman (2001)]. On peut en premier lieu distinguer la régression linéaire des méthodes non-linéaires. Au cours des 20 dernières années, de nombreuses méthodes et techniques de régression non-linéaires ont été proposées.

On peut notamment citer les méthodes de plus proches voisins [Zorita et al. (1995), Martin et al. (1996)], les forêts aléatoires [Breiman (2001)], les modèles Machine à Vecteurs de support pour la Régression (SVR) [Vapnik et al. (1997)] ou encore les réseaux de neurones [Lau (1991), Dreyfus et al. (2011)]. Ces approches sont très populaires dans de nombreux domaines : physique, chimie, biologie, finance, sociologie, *etc*.

Dans le domaine des sciences de l'océan et de l'atmosphère, les méthodes dites analogues et les méthodes de régression linéaire sont souvent utilisées [Zorita et von Storch (1999), Walmsley et al. (2001), Goubanova et al. (2010)]. Les méthodes analogues correspondent à des méthodes de régression du type k-plus proches voisins. Les méthodes de régression multi-modèles connaissent depuis quelques années également un essor grandissant [Ben Ticha (2007), Minvielle (2009)]. Pour ces trois différents types de modèle de régression, quelques références choisies sont introduites de manière détaillée.

2.2.2.1 Méthodes analogues

Elles sont souvent utilisées comme première référence pour les méthodes de *downscaling* [Martin et al. (1996), Zorita et von Storch (1999), Timbal et McAvaney (2001)], par exemple, pour la reconstruction de la précipitation locale en fonction d'autres paramètres météorologiques et pour la reconstruction de la température de la surface de mer. Elle peut atteindre une très bonne performance si on dispose d'une grande base de données, représentative de toute la variabilité des processus étudiés.

L'approche « analogue » est une régression simple. Le terme « proche » implique une distance vectorielle entre deux échantillons : pour un nouvel échantillon x, les distances avec tous les échantillons $\{x_s\}$ dans la base de données sont calculées ; les réponses d'un ou plusieurs échantillons les plus proches sont utilisées pour calculer la nouvelle réponse.

Zorita et von Storch furent parmi les premiers à explorer des approches de *downscaling* statistique pour des problématiques environnementales, plus particulièrement de prédictions de précipitation. Dans leur article de 1999, ils utilisent le plus proche voisin pour obtenir les précipitations journalières et mensuelles sur la période 1901-1989 pour 92 stations dans la péninsule Ibérique. La base de données utilisée est les données de pression de surface de mer du National Center for Environmental Prediction (NCEP) (5°) sur la période 1951-1989. L'article introduit et compare différentes méthodes : méthode analogue, méthode de régression linéaire, méthode de classification et réseaux de neurones. Ils concluent que la méthode analogue obtient de très bons résultats par rapport aux 3 autres méthodes et qu'elle reproduit les bons niveaux de variabilité des variables locales.

2.2.2.2 Méthode de Régression Linéaire Multiple (MLR)

Dans la méthode de MLR, la variable à prédire est linéaire en ses paramètres et en ses variables explicatives. La régression linéaire est déjà développée avant que l'on utilise l'ordinateur pour les statistiques. Elle continue à être utilisée pour des raisons de simplicité et de description adéquate et interprétable de la façon dont les entrées affectent les sorties.

L'article de Goubanova et al. (2010) utilise les méthodes de *downscaling* statistique pour établir le lien entre les données à grande échelle et les données à petite échelle. Une fois le lien établi, les sorties du modèle IPSL-CM4 sont utilisées comme variables explicatives pour prédire les anomalies de vent journalier à petite échelle. Les données de reconstruction sont ensuite utilisées pour estimer les impacts régionaux dus au changement climatique.

Dans cet article, les données utilisées pour l'apprentissage et les données utilisées pour la prédiction ne sont pas les mêmes. Dans la phase d'apprentissage, les données des réanalyses du NCEP (2.5°) et les données de vent du diffusiomètre de QuikSCAT (0.5°) pour la période 2000-2008 sont utilisées comme données à grande échelle et données à petite échelle, respectivement. Pour diminuer les dimensions de l'espace d'entrée et de sortie, la technique des Fonctions Orthogonales Empiriques (EOF) est utilisée. Les premiers 20 EOF des anomalies des données des réanalyses sur la zone S5°–40° et W70°–90° sont utilisées comme variables explicatives (d'entrée) et les premiers 10 EOF des anomalies des vents Quick Scatterometer (QuikSCAT) sur la zone S0°–55° et W62°–122° sont utilisées comme variables expliquées (de sortie). La méthode de régression linéaire est utilisée pour établir le lien entre les EOF dans l'espace d'entrée et les EOF dans l'espace de sortie. Dans la phase de prédiction, les résultats de la simulations du modèle IPSL-CM4 sont projetés dans l'espace EOF de l'espace d'entrée obtenue dans la phase d'apprentissage. Ensuite, les prédictions sont calculées avec les nouveaux coefficients et le modèle déjà appris.

Les estimations d'une augmentation de la moyenne de la vitesse de vent sur la côte restent cohérentes avec d'autres analyses et d'autres études récentes. Les résultats montrent néanmoins de plus en plus de difficultés d'émulation à l'approche de la côte, à travers les coefficients de corrélation entre les champs de vent reconstruits et observés par satellite qui diminuent progressivement du large vers la côte Pérou-Chili. La diminution va de 0.8 à 0.5, voire 0.2 par endroits. On peut également remarquer qu'il reste peu courant d'utiliser dans la phase de prédiction des données différentes de celles utilisées dans la phase d'apprentissage, ce qui revient à supposer que les différentes données sont complètement cohérentes.

La thèse de Minvielle (2009) propose une méthode de désagrégation hybride, basée sur une décomposition en régimes de temps, permettant de reconstruire,

selon une fonction de transfert estimée à partir des observations, les séries journalières des variables atmosphériques pour les modèles de forçage sur la région Atlantique nord et Atlantique tropical.

La reconstruction statistique est constituée de 3 phases. D'abord, une phase de classification automatique (*k-means*) regroupe les jours qui ont des caractéristiques atmosphériques proches dans la même classe selon les critères de régimes de temps sur l'atlantique. Le centre de chaque classe est appelé « centroïde ». Ensuite, les distances de chaque jeu de données aux centroïdes de toutes les classes sont calculées et utilisées comme variables explicatives. Finalement, en chaque point de grille, une régression linéaire est utilisée pour apprendre la relation entre les paramètre de surface à HR et les distances euclidiennes :

$$y(t) = \sum_{k=1}^{m} \omega_k d_k(t) + b \tag{2.3}$$

où m est le nombre de classe.

Cette méthode peut être vue comme une approche de réduction de dimension : on remplace les variables explicatives par un vecteur de distances. Deux problèmes peuvent être identifiés : premièrement, la relation linéaire entre les hautes résolutions (température, humidité, *etc*) et les distances de chaque BR aux centroïdes est difficile à évaluer ; deuxièmement, l'utilisation d'inflation de la variance après la phase de régression par cas analogues fait que l'approche a les mêmes inconvénients qu'une approche analogue.

Pour la première remarque, autant il est difficile d'évaluer la relation linéaire entre la température et l'humidité de surface avec les distances euclidiennes entre les vents à BR et les centroïdes, autant il est possible de prouver que les vents à HR sont plus probablement linéairement proportionnés aux distances euclidiennes quadratiques (et non les distances euclidiennes comme proposé). Pour le démontrer, supposons que l'on cherche à prédire le vent à HR (u_h, v_h) en un point en donnant la circulation générale — le vent à BR au même point — (u, v). Pour que la méthode fonctionne, au minimum 3 classes, et donc 3 centroïdes, $\{u_i^\circ, v_i^\circ\}, i = 1, \cdots, 3$ sont nécessaires. Les distances entre un point particulier à BR et les centroïdes s'écrivent :

$$\begin{cases} d(x, x_1^\circ) = \sqrt{(u - u_1^\circ)^2 + (v - v_1^\circ)^2} & (1) \\ d(x, x_2^\circ) = \sqrt{(u - u_2^\circ)^2 + (v - v_2^\circ)^2} & (2) \\ d(x, x_3^\circ) = \sqrt{(u - u_3^\circ)^2 + (v - v_3^\circ)^2} & (3) \end{cases} \tag{2.4}$$

En résolvant l'équation 2.4, on obtient :

$$\begin{aligned} u &= \alpha_0 + \alpha_1 d(x, x_1^\circ)^2 + \alpha_2 d(x, x_2^\circ)^2 + \alpha_3 d(x, x_3^\circ)^2 \\ v &= \beta_0 + \beta_1 d(x, x_1^\circ)^2 + \beta_2 d(x, x_2^\circ)^2 + \beta_3 d(x, x_3^\circ)^2 \end{aligned} \tag{2.5}$$

où les coefficients α et β ne dépendent que des cordonnées des centroïdes. Si l'on suppose que le vent à HR est proportionnel au vent à BR, on peut en déduire que le vent à HR a une relation linéaire avec les distances quadratiques.

La deuxième étape de la thèse de Minvielle (2009) est l'utilisation des paramètres reconstruits à HR pour forcer un modèle d'océan de plus HR. Ainsi, des simulations océaniques ont été réalisées avec deux modèles d'océan, le modèle Nord-Atlantique au 1/4° NATL4 sur la période 1975-2001 et le modèle global 1/2°ORCA05 sur la période 1958-2002. L'analyse de ces simulations montre une bonne représentation de l'état océanique moyen et de la variabilité interannuelle, avec une sous-estimation de variance en accord avec les biais du forçage et une tendance en température partiellement capturée.

2.2.2.3 Méthode de régression multi-modèles

Les approches multi-modèles reposent sur une première étape de classification. On associe à chaque classe un modèle de régression à HR, souvent de type MLR [Walmsley et al. (2001)]. La classification permet d'expliciter la structure d'un ensemble de données et de regrouper les données qui ont des comportements similaires dans des sous-ensembles. La classification peut simplifier les modèles de régression et finalement améliorer la performance des modèles. Donnons quelques exemples de régression multi-modèles :

L'article de Walmsley et al. (2001) propose des méthodes statistiques pour prédire les vitesses de vent *offshore* le long des côtes Danoises. Les mesures météorologiques terrestres, comme la vitesse et la direction du vent, la température, le gradient de température, l'heure du mesure *etc*. à deux endroits proches de la côte danoise sont utilisées comme variables explicatives. Les données sont d'abord distribuées dans 9 classes différentes en fonction de l'intensité du vent. Dans chaque classe, les données sont encore séparées en deux parties : une partie pour l'apprentissage du modèle statistique (70%) et l'autre partie pour les évaluations du modèle (30%). Pour chaque partie des données d'apprentissage, les deux approches d'apprentissage statistique sont comparées : Régression Linéaire Multiple (MLR) et arbre de régression (Classification and Regression Trees).

En raison des dépendances très linéaires entre les données terrestres et les données de vent *offshore*, l'article obtient de très bons résultats pour les deux approches, avec une légère amélioration pour la méthode d'arbre de régression. Les coefficients de corrélations varient de 0.94 à 0.98 et les Root-Mean-Square Error (RMSE) varient de $0.6\,\mathrm{m\,s^{-1}}$ à $1.1\,\mathrm{m\,s^{-1}}$.

La thèse de Ben Ticha (2007) propose une première phase de classification pour regrouper des champs de vent similaires en terme de comportement spatial du flux. Cela revient à calculer la distance de la direction entre les deux champs de vent à BR spatiale. Les champs de vent similaires sont classés dans la même classe. Le jeu de données est donc réduit à un certain nombre de cas représentatifs. Une deuxième phase d'association consiste à associer les données à HR aux situations à BR. Chaque classe est donc associée au moins

à une donnée à HR. Cette association permet d'extraire les structures à haute fréquence qui sont absentes à BR. Les structures obtenues sont injectées dans un nouveau champ de vent d'entrée à BR spatiale pour reconstruire un champ de vent à HR synthétisée. Les champs de vent synthétisés sont finalement utilisés pour établir la cartographie du potentiel éolien.

On peut remarquer deux inconvénients dans cette approche : premièrement, le nombre de points de la zone d'étude est relativement petit (9 points pour une zone de $6.0° \times 6.0°$). Pour plus de points ou pour une zone plus grande, il est plus difficile d'utiliser la même approche pour classer les vents par leurs directions. Deuxièmement, cette approche revient à injecter des variabilités à HR constantes pour chaque classe de vent. Comme le critère de classification est basé sur les directions de vent à BR, cela signifie que chaque classe contient des vents à des intensités différentes. Injecter des variabilités constantes par classe revient à supposer que les variabilités à HR ne changent pas en fonction de l'intensité de vent. Cette hypothèse n'est pas toujours vraie (cf. Chapitre 4, la section 4.3.2).

On peut notamment remarquer que les erreurs de reconstruction restent très élevées, un biais de 100% (par rapport au champ de référence) peut même apparaître dans certaines zones de reconstruction.

2.3 Modèle proposé

Pour aborder en toute généralité les différents modèles, nous proposons un modèle de *downscaling* statistique générique de la forme suivante :

$$f(x) = \sum_{s=1}^{n} c_s \mathrm{K}(x, x_s) + b \tag{2.6}$$

où $c \in \mathbb{R}^n$ est le vecteur de coefficients de régression, n étant le nombre de données de référence. b est le biais et K est une fonction de noyau. Cette forme générique permet de formuler dans un cadre unifié les différents modèles de régression à l'aide de noyau. Ils ne diffèrent ensuite que par la paramétrisation retenue c et la méthode de calibration associée K.

À la différence de la formulation classique d'un modèle de régression, qui formule la variable à prédire y en fonction de toutes les variables explicatives $(X^1\ X^2\ \cdots\ X^d)$ explicitement, le modèle proposé dans l'équation 2.6 exprime la variable y à l'aide d'une fonction de noyau, qui est elle même une fonction de similarité entre deux échantillons de x. Cette nouvelle formulation engendre les composantes clés du modèle de *downscaling* suivant :

Choix des données de référence Les données de référence x_s peuvent être tous les échantillons de x, ou quelques échantillons de x, ou encore les barycentres/centroïdes des sous-ensembles de x.

Choix du noyau La fonction de noyau mesure la similarité entre la variable d'entrée x et la donnée de référence x_s. Dans le cas d'une régression

linéaire, la fonction de noyau K prend la forme d'un produit scalaire entre x et x_s qui s'écrit de la façon suivante :

$$\mathrm{K} = \langle x, x_s \rangle$$

et, dans le cas d'une régression non-linéaire, la variable x peut être transformée par une application non linéaire $\Phi(x)$ dans un autre espace où il existe une solution linéaire au problème. Il suffit que le noyau choisi soit un noyau de « Mercer » qui vérifie $K(x, x_i)$ continu, symétrique et défini positif [Vapnik et al. (1997), Schölkopf et Smola (2001)]. L'équation (5.30) donne un exemple de noyau non-linéaire du type gaussien.

$$K(x, x_s) = \exp(-\gamma \|x - x_s\|^2) \tag{2.7}$$

Paramétrisation des coefficients Les paramètres c sont des réponses accordées à chaque référence x_s. Ils peuvent être fixés *a priori* ou ajustés tels que le modèle obtenu présente les qualités d'apprentissage et de généralisation requises [Dreyfus et al. (2011)].

Nous reformulons les modèles considérés dans les exemples de l'état de l'art à l'aide de ces éléments-clés :

– Pour les modèles basés sur une régression linéaire utilisés dans Walmsley et al. (2001), Minvielle (2009), Goubanova et al. (2010), la fonction de régression peut être écrite sous la forme :

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{c}_s \langle x, x_s \rangle \tag{2.8}$$

avec une fonction de noyau correspondant au produit scalaire entre x et x_s :

$$\mathrm{K} = \langle x, x_s \rangle$$

Goubanova et al. (2010) utilisent les coefficients de projection dans l'espace EOF de l'ensemble des échantillons de x comme données de référence $\{x_s\}$, tandis que Minvielle (2009) se restreint à quelques points, les centroïdes de ses régimes de temps.
Les coefficients de régression sont généralement obtenus par minimisation d'une fonction de coût des moindres carrés, qui est la somme des distances quadratiques entre la variable y estimée par modèle et celle mesurée. Les estimations des coefficients c sont finalement données par :

$$\hat{c} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}$$

– Pour les modèles basés sur les méthodes analogues utilisés dans Zorita et von Storch (1999), la fonction de régression prend la forme :

$$\hat{f}(x) = \sum_{s=1}^{n} \hat{c}_s \, g(x, x_s) \tag{2.9}$$

La fonction de noyau K correspond à une mesure de similarité quelconque g entre la variable x et la donnée de référence x_s. Pour l'approche du plus proche voisin, $\mathrm{K}(x, x_j) = 1$, si $\forall k \neq j,\ d(x, x_j) < d(x, x_k)$ et $\mathrm{K}(x, x_j) = 0$ sinon. Pour l'approche de type k-plus proches voisins, on peut choisir une fonction exponentielle par exemple, avec $\mathrm{K}(x, x_i) = \exp(-\frac{\|x-x_i\|^2}{2\sigma^2})$. En général, l'ensemble des échantillons de x est utilisé comme données de référence. Les paramètres $\hat{c}$ sont fixés *a priori* par l'utilisateur. Ils correspondent généralement directement aux réponses de chaque donnée de référence normalisées :

$$\frac{y_s}{\sum_{s=1}^{n} \mathrm{K}(x, x_s)}$$

En résumé, les modèles de régression MLR utilisent un noyau linéaire avec les données de référence fixées par l'utilisateur qui peuvent être l'ensemble des échantillons ou les centroïdes des sous-ensembles, et ses paramètres sont généralement estimés par la méthode des moindres carrés. Le noyau pour les méthodes analogues peut être une fonction de mesure de similarité quelconque, tous les échantillons sont utilisés comme données de référence et les paramètres sont fixés *a priori* par l'utilisateur.

Dans ce contexte, l'approche Machine à Vecteurs de support pour la Régression (SVR) qui optimise simultanément le choix des données de référence et les coefficients de régression est proposée. Les modèles SVR sont construits sur la base de la théorie de l'apprentissage de Vapnik en s'appuyant sur le principe de minimisation du risque structurel [Vapnik et al. (1997)]. Les paramètres sont obtenus en minimisant un critère des moindres carrés pénalisés avec une marge douce, puis en résolvant ce problème d'optimisation sous contraintes [Schölkopf et Smola (2001), Basak et al. (2007)]. Grâce à l'autorisation de certaines erreurs (marge douce), la fonction de régression f ne dépend que d'une sous-partie des données d'apprentissage, appelées *Vectors de Support* [Guermeur et Paugam-Moisy (1999)]. Une liste de fonction de noyau peut être choisie tant qu'elle vérifie $K(x, x_i)$ continu, symétrique et défini positif [Vapnik et al. (1997)]. Pour un problème de régression non-linéaire, un noyau non-linéaire, par exemple un noyau gaussien (Radial Basis Function), peut être envisagé. Nous détaillons l'approche SVR dans le chapitre 5 en section 5.3.3.

Le tableau 2.1 donne une synthèse de différents modèles de régression en fonction des trois composantes clés. Le chapitre 5 introduit de manière détaillée ces différents modèles de régression.

Tableau 2.1 – *Synthèse de différents modèles de régression en fonction des trois composantes clés.*

Méthode	Choix des références $\{x_s\}$	Paramètre c_s	Noyau K
MLR	Toutes ou Centroïdes	Estimateurs des moindres carrées	$K = \langle x, x_s \rangle$
Analogue	Toutes	Constant	Exemple d'un noyau gaussien $K(x, x_s) = \exp(-\gamma \lvert x - x_s \rvert^2)$
SVR	Vectors de Support	Multiplicateurs de Lagrange	Noyer Mercer

Données expérimentales 3

Les méthodes d'apprentissage statistique reposent toutes sur une base de données historiques. Pour le problème qui nous intéresse, elle contient les données de vent à basse et haute résolution dont nous détaillons les caractéristiques dans ce chapitre. Une rapide présentation des ces données en section 3.1 permet de mieux comprendre l'origine de leurs différences. La zone d'étude choisie est introduite en section 3.2. La section 3.3 décrit ensuite l'appairage des données à BR et à HR, l'identification et la suppression de données aberrantes, et l'élimination des couples de données incohérents.

3.1 Présentation des données utilisées

Les données BR utilisées sont les sorties du modèle numérique ECMWF et celles à HR sont issues d'observations satellitaires SAR. Pour mieux comprendre les sources de différences entre ces deux types de données, nous présentons d'abord les concepts d'« échelle » et de « résolution » et ensuite les principes des modèles numériques et des observations par télédétection.

3.1.1 Concepts d'échelle et de résolution

3.1.1.1 Échelle

Le concept d'échelle aide à comprendre la différence entre les données à différentes résolutions.

Chaque phénomène océanique ou atmosphérique a une échelle minimale à partir de laquelle il commence à être observé et une échelle maximale, la taille maximale du phénomène. Ces phénomènes sont classés en fonction de leurs échelles spatiales et temporelles. Or, les petites échelles spatiales vont souvent de pair avec les petites échelles temporelles et les grandes échelles spatiales vont souvent de pair avec les grandes échelles temporelles [Orlanski (1975)].

Historiquement, les météorologues ont étudié la dynamique atmosphérique à deux échelles : la macro-échelle qui correspond à une échelle spatiale de plus d'un millier de kilomètres et à une échelle temporelle de l'ordre d'une semaine ; et la micro-échelle qui a une échelle spatiale de quelques mètres et une échelle temporelle de l'ordre de quelques minutes. Néanmoins, il existe de nombreux phénomènes intermédiaires qui n'étaient pas classés en raison des difficultés

d'observation. En 1975, Orlanski (1975) a ajouté une catégorie intermédiaire entre la micro-échelle et la macro-échelle : la méso-échelle. Elle permet de caractériser les phénomènes qui ont une échelle spatiale et temporelle médiane.

Finalement, les phénomènes océaniques et atmosphériques sont divisés en 3 grandes catégories : macro-échelle, méso-échelle et micro-échelle :

- la macro-échelle ou synoptique se réfère aux événements synoptiques sur une échelle de plusieurs milliers de kilomètres, comme les fronts chauds et froids, les dépressions, les anticyclones, les ouragans et les grands courants marins ;
- le terme « méso-échelle » semble avoir été premièrement mentionné par Ligda (1951) pour décrire les phénomènes à l'échelle plus petite que l'échelle synoptique, comme ceux de la circulation atmosphérique générale, mais plus grande que la micro-échelle. Selon la définition du glossaire de l'American Meteorological Society (AMS), la méso-échelle est en rapport à des phénomènes atmosphériques ayant des échelles horizontales allant de quelques dizaines à plusieurs centaines de kilomètres, par exemple les orages, les lignes de grains, les fronts, les bandes de précipitations dans les cyclones tropicaux et extra-tropicaux, et des systèmes générés par la topographie comme les ondes de montagne ou les brises de mer et de terre ;
- la micro-échelle caractérise, en météorologie et en océanographie, ce qui se passe à l'échelle du nuage individuel ou du vortex dans l'océan local, soit moins de 2 km à l'échelle horizontale. Il s'agit de phénomènes de très petites échelles comme les tornades, les turbulences, etc.

Le tableau 3.1 donne quelques ordres de grandeur pour chaque échelle. Comme vu précédemment, plus l'échelle spatiale est petite, plus l'échelle temporelle est petite.

Tableau 3.1 – *Catégories d'échelles spatiales et temporelles pour la dynamique atmosphérique [Orlanski (1975)].*

Échelle	Sous-catégorie	Échelle temporelle	Échelle horizontale
Macro-échelle	α (Planétaire)	$>$ quelques jours	$>$ 2000 km
	β (Synoptique)		
Méso-échelle	α	quelques jours	2000 km
	β	$\uparrow$	$\uparrow$
	γ	1 h	2 km
Micro-échelle	α	1 h	$<$ 2 km
	β	$\uparrow$	
	γ	1 s	

Cette classification s'applique également aux vents. Les vents d'ouest et les alizés sont des vents à l'échelle planétaire ; les orages et les brises maritimes sont des vents à méso-échelle ; la plus petite échelle de mouvement d'air est la micro-échelle, caractéristique des vents très locaux, souvent chaotiques, notam-

ment les rafales et les tourbillons de poussière. Le tableau 3.2 fournit quelques exemples de vents avec leur échelle et leur densité maximale respective.

Tableau 3.2 – *Quelques exemples de vent à différentes échelles.*

Perturbation	Échelle	Durée	Vent maximum
Cyclone extra-tropical	500–2000 km	3–15 jours	$55\,\mathrm{m\,s^{-1}}$
Anticyclone	500–2000 km	3–15 jours	$10\,\mathrm{m\,s^{-1}}$
Ouragan	300–2000 km	1–7 jours	$90\,\mathrm{m\,s^{-1}}$
Cyclone tropical	300–1500 km	3–15 jours	$33\,\mathrm{m\,s^{-1}}$
Dépression tropicale	300–1000 km	5–10 jours	$17\,\mathrm{m\,s^{-1}}$
Rafale	10–300 km	0.5–6 h	$35\,\mathrm{m\,s^{-1}}$
Méso-cyclone	10–100 km	0.5–6 h	$60\,\mathrm{m\,s^{-1}}$
Coup de vent méso-échelle	4–20 km	10–60 min	$60\,\mathrm{m\,s^{-1}}$
Coup de vent micro-échelle	1–4 km	2–15 min	$70\,\mathrm{m\,s^{-1}}$

3.1.1.2 Résolution

Pour pouvoir étudier un phénomène, il faut choisir une échelle adaptée, d'au moins la taille minimum du phénomène pour que celui-ci soit observable. On parle alors de portée. Néanmoins, les observations disponibles sont souvent discrètes, c'est à dire qu'il n'existe qu'une seule valeur descriptive du phénomène, quantitative ou qualitative, pour une certaine durée ou sur une certaine zone. On parle alors de résolution. Si la résolution est supérieure à la taille maximale du phénomène, celui-ci risque de ne plus être observable. Ainsi, meilleure est la résolution, meilleure sera la qualité des observations des phénomènes à petite échelle. Pour cette raison, on utilise souvent le terme « données à petite échelle » pour désigner des données de HR et le terme « données à grande échelle » pour désigner des données de BR.

Les données océanographiques et météorologiques sont souvent associées à une date et à un endroit. Dans le domaine temporel, elles peuvent être instantanées ou moyennées sur une durée. La « résolution temporelle » correspond au temps écoulé entre deux mesures consécutives. Dans le domaine spatial, elles peuvent être en un point ou moyennées sur une zone. La volume de la surface définit la résolution spatiale. En traitement d'images, la « résolution spatiale » est définie par un nombre de pixels par unité de longueur de la structure à numériser. Plus le nombre de pixels par unité est élevé, plus la distance entre deux pixels est petite, plus la dimension de la matrice de l'image est grande, et plus le degré de détail de l'image est élevé.

Les termes « fin » ou « grossier » appliqués à la qualité de la résolution sont assez ambigus. Une résolution « fine » désigne une « haute » résolution et une résolution « grossière » désigne une « basse » résolution. Pour éviter cette ambiguïté, les termes « haute » et « basse » résolution sont utilisés.

3.1.1.3 Échelle de champs de vent ECMWF et SAR

Les données à BR utilisées sont issues du modèle européen ECMWF. Sa résolution spatiale est de 0.5° en longitude et en latitude. La résolution temporelle est de 6 h. Les vents à HR sont issus des images SAR prises par le satellite ENVISAT. Les données SAR atteignent une résolution spatiale de l'ordre de 1 km avec une résolution temporelle irrégulière. La précision du vent SAR est de l'ordre de $2\,\mathrm{m\,s^{-1}}$ en intensité et de 25° en direction [Monaldo et al. (2003)]. Étant donné l'échantillonnage de l'océan par un satellite défilant, une zone donnée de l'océan est échantillonnée irrégulièrement en temps et en espace. À un temps donné, la zone peut n'être que partiellement couverte par le SAR. Les directions initiales pour l'estimation du vent SAR sont souvent données par les modèles numériques comme ECMWF. Cela explique que les différences entre ECMWF et SAR en un point apparaissent plus suivant l'intensité du vent que suivant sa direction.

Le rapport de la résolution spatiale entre la haute et BR est d'environ 50×50. Selon la catégorisation d'échelles donnée en section 3.1.1, la résolution de l'ECMWF se situe à la méso-échelle. Théoriquement, sa résolution est suffisante pour capturer certains phénomènes comme les rafales, les méso-cyclones, et d'autres phénomène à échelle encore plus grande comme les dépressions ou les anticyclones. La résolution du SAR se situe dans le catégorie de micro-échelle, ce qui signifie que sa résolution permet d'appréhender des phénomènes beaucoup plus locaux, comme les coups de vent, les turbulences locales, certains systèmes générés par la topographie comme les ondes de montagne ou les brises de mer et de terre.

3.1.2 Données du modèle numérique

Les prévisions numériques reprennent les lois physiques régissant le comportement des fluides et utilisent des modèles mathématiques de l'atmosphère et de l'océan pour prédire les conditions atmosphériques à partir des conditions initiales (le début de la prédiction). Elles ne deviennent réalistes qu'à partir des années 50 avec l'apparition de l'informatique. La Prévision Numérique du Temps (PNT) repose sur le choix d'équations mathématiques offrant une approximation réaliste du comportement de l'atmosphère et de l'océan. Ces équations sont ensuite résolues numériquement, pour obtenir une simulation des états futurs des conditions atmosphériques [Hamilton (1996)]. L'approche classique pour obtenir des équations possédant une valeur prédictive consiste à résoudre une ou plusieurs équations différentielles contenant la variable temporelle. Dans les cas les plus commodes, la solution exprime les variables à prévoir en fonction du temps et des conditions initiales. Il suffit alors d'alimenter cette équation avec les valeurs numériques requises pour obtenir une solution dite exacte.

Pour obtenir une bonne prévision, il faut tenir compte des phénomènes qui

sont plus petits que la résolution du modèle (phénomènes dits sous-maille). La représentation de l'influence moyenne à grande échelle des phénomènes de la petite échelle est appelée paramétrisation. Les phénomènes sous-maille les plus souvent paramétrés par les concepteurs des modèles sont :

- la convection verticale, dont font partie les orages ;
- la physique des nuages ;
- les effets radiatifs atmosphériques ;
- l'interaction surface-air. Par exemple, les échanges de chaleur et d'humidité entre la surface et l'atmosphère ;
- l'effet des montagnes et des irrégularités du terrain. Par exemple, l'effet de blocage du vent et les ondes atmosphériques en aval des montagnes.

Les mailles du modèle numérique sont souvent trop grandes pour « décrire » explicitement ces phénomènes plus petits. Ils sont pris en compte par le biais d'algorithmes spécifiques qui simulent leur influence moyenne à l'intérieur des mailles du modèle.

Bien que la Prévision Numérique du Temps (PNT) représente le plus grand succès de la météorologie, elle n'est que partiellement efficace dans son application. La PNT rencontre trois obstacles principaux :

- la paramétrisation physique du modèle. Il est difficile de prendre en compte tous les phénomènes ;
- la dynamique atmosphérique très sensible, dans certaines conditions, à la moindre fluctuation ;
- la représentativité à cause de la faible résolution spatiale. Plus la résolution est faible, moins le modèle est apte à bien représenter les phénomènes de moyenne et petite échelle.

Malgré les obstacles, la PNT a l'avantage de fournir des données à une résolution temporelle régulière et une couverture spatiale complète. En général, elle fournit de bonnes informations sur la circulation générale.

Le centre européen pour les prévisions météorologiques à moyen terme (ECMWF) fournit plusieurs types de données issues des modèles numériques. Dans cette étude, les données d'analyses sont utilisées. Les données d'analyses combinent les données de prévisions à court terme avec les observations pour produire le meilleur ajustement des deux. Les données sont disponibles toutes les 6 H (0 h, 06 h, 12 h, 18 h UTC), avec une résolution spatiale de 0.5°.

3.1.3 Observations par télédétection

Les mesures par télédétection échantillonnent directement les phénomènes locaux souvent inaccessibles à la prévision numérique opérationnelle. Dans cette partie, on introduit les différentes techniques de télédétection pour les mesures de paramètres océaniques et atmosphériques, plus spécifiquement pour les mesures de vent.

Depuis les années 70, les applications de la télédétection se sont beaucoup développées dans les domaines de la météorologie, de la climatologie et de

l'océanographie, et s'accompagnent d'une révolution informatique. Ce type de méthode d'acquisition utilise la mesure des rayonnements électromagnétiques émis ou réfléchis des objets étudiés dans un certain domaine de fréquences (infrarouge, rayonnement visible, micro-ondes). Ceci est rendu possible par le fait que les objets étudiés émettent ou réfléchissent du rayonnement à différentes longueurs d'onde et intensités selon leur état.

Les mesures ou les acquisitions d'informations par télédétection sur un objet ou un phénomène se font à distance par les capteurs. Les types de capteurs utilisés pour l'océanographie sont très variés [Kergomard (2013)]. Les radiomètres utilisant le rayonnement visible analysent la couleur de l'océan, ce qui permet de mesurer la production biologique (plancton) et la turbidité ; les radiomètres infrarouge ou micro-ondes mesurent la température et la salinité de surface de la mer. Les radars embarqués sur des avions ou certains satellites ont l'avantage d'être insensibles aux nuages. Ils permettent d'observer les phénomènes ondulatoires présents sur l'océan, en particulier les vagues. Enfin, certains types particuliers de capteurs, radars-altimètres ou diffusiomètres sont utilisés pour mesurer avec une très grande précision l'altitude de la surface de la mer qui reflète une composante de la dynamique océanique.

Dans notre étude, nous nous intéressons au vent de surface de l'océan dont la mesure est maintenant faite au quotidien depuis l'espace par de nombreux capteurs, soit actifs soit passifs. Le principe de mesure de tous ces capteurs repose sur la variation de la rugosité de la surface de la mer en fonction du vent et sur les interactions entre ondes électromagnétiques et ondes de surface. Nous ne nous intéressons ici qu'aux capteurs actifs. Les capteurs actifs émettent un signal hyperfréquence et recueillent le coefficient de rétrodiffusion σ° de la surface de la mer. Les variations d'amplitude du signal rétro-diffusé dépendent des petites vagues (de l'ordre du cm), qui sont créées par le vent local [Marcos (2002)]. La relation entre le coefficient de rétrodiffusion et le vent peut être établie par un modèle géophysique semi-empirique (*Geophysical Model Function (GMF)*) :

$$\sigma^\circ = B_0 \left[1 + B_1 \cos\phi + B_2 \cos 2\phi\right]^z \tag{3.1}$$

où ϕ est l'angle entre la direction de vent et l'angle de visée du radar ; les coefficients B_0, B_1, B_2 dépendent de l'angle d'incidence et l'intensité de vent normalisée et z est un coefficient de régularisation ($z = 0.625$ dans le CMOD4 [Hersbach et al. (2007)]). Les vents sont estimés en résolvant le problème inverse, c'est-à-dire le calcul du vent en connaissant le coefficient de rétrodiffusion. Une seule mesure du σ° ne suffit pas pour estimer le vent en vitesse et direction. Plusieurs mesures du σ° sous des azimuts différents, ou une information externe est nécessaire.

Les données de vents à HR utilisées dans notre étude sont issues des observations du satellite ENVIronmental SATellite (ENVISAT). ENVISAT est un satellite d'observation de la Terre de l'Agence Spatiale Européenne (ESA) lancé en mars 2002 dont l'objectif est de mesurer de manière continue à différentes

échelles les principaux paramètres environnementaux de la Terre relatifs à l'atmosphère, l'océan, les terres émergées et les glaces. Après d'avoir perdu le contact avec le satellite le 8 avril 2012, l'ESA a annoncé officiellement la fin de mission du satellite ENVISAT le 9 mai 2012. C'est un satellite de très grande taille qui embarque 10 instruments scientifiques comprenant radar, radiomètre et plusieurs spectromètres. Son radar à ouverture synthétique (Advanced SAR (ASAR)) opère en bande C et permet de mesurer les vents de surface de la mer [Curlander et McDonough (1991), McCandless et Jackson (2004)]. Pour l'estimation des données de vent, il existe plusieurs méthodes de traitement des données SAR. Dans cette étude, la vitesse du vent SAR est calculée avec l'algorithme CMOD4 [Stoffelen et Anderson (1997)], et la direction du vent auxiliaire provient du modèle de prévision numérique ECMWF. La résolution spatiale est de l'ordre du kilomètre.

Par rapport aux données du modèle numérique, les données obtenues par télédétection ont l'avantage d'améliorer la résolution spatiale. Il reste néanmoins quelques limitations :

- Le grand laps de temps entre deux passages consécutifs du satellite signifie une BR temporelle. Avec les données de vent SAR issues du satellite ENVISAT, il faut plus de 10 jours en moyenne pour que le satellite repasse au même endroit ;
- La pluie peut faiblement atténuer et réfléchir les signaux micro-ondes : d'un côté, les gouttes atténuent les signaux qui les traversent pendant leur propagation, et par conséquent elles diminuent la vitesse du vent estimé ; d'un autre côté, l'écho radar réfléchi par les gouttes augmente la vitesse du vent estimé ; de plus, la rugosité de la surface de la mer est augmentée par les éclats provoqués par les gouttes de pluie ;
- Les vents étant déduits de la rugosité de la surface de la mer, les glaces, la terre, tous les autres objets sur la surface de la mer peuvent contaminer les mesures du vent et il faut les identifier et les supprimer. Parmi ces éléments de contamination, la terre est facile à identifier, contrairement aux glaces et aux autres objets flottants.

En conclusion, d'un côté, la télédétection offre l'avantage de permettre une vision synoptique des conditions de vent à HR de vastes régions qu'il est impossible d'obtenir par les moyens traditionnels (par exemple : bouées, bateaux). D'un autre côté, les données obtenues par télédétection permettent certaines études à petite échelle impossibles avec les données issues des modèles numériques synoptiques.

3.1.4 Représentation des champs de vent

Chaque champ des données ECMWF ou SAR est représenté sur une grille. Chaque point de la grille est identifié par sa longitude et sa latitude. Pour les champs de vents, une donnée à un point $(\mathrm{lat}_i, \mathrm{lon}_j)$ est un vecteur. D'un point vu théorique, le vent en un point donné est défini comme le vecteur vitesse de

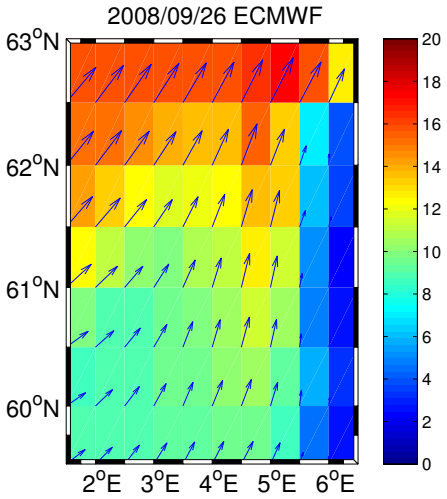


FIGURE 3.1 – *Exemple d'un champ de vent spatial autour de la zone Bergen en Norvège. Les différentes couleurs correspondent à l'intensité de vent en* $\mathrm{m\,s^{-1}}$. *Les flèches indiquent la direction vers laquelle le vent est en train de souffler et leur longueur est proportionnelle à l'intensité de vent.*

la particule atmosphérique localisée en ce point [Le Vourc'h et al. (2001)]. Un vecteur vent est composé soit d'un module (intensité) a et d'une direction θ, soit d'une composante zonale u (l'axe E-W) et d'une composante méridienne v (l'axe N-S).

Il y a trois inconvénients à travailler avec intensités et directions de vent : il est plus difficile de caractériser le bruit ; les erreurs d'émulation peuvent être plus grandes du fait que les bruits sont amplifiés de la même façon que les intensités par rapport aux coordonnées cartésiennes ; et il n'est pas toujours facile de travailler avec les valeurs angulaires.

Par la suite, cette dernière représentation est utilisée, un vecteur au point $(\mathrm{lat}_i, \mathrm{lon}_j)$ s'écrit comme un vecteur (u, v). L'ensemble de tous les vecteurs de vent pris à un instant donné constitue un champ de vent :

$$\begin{bmatrix} (u_{\mathrm{lat}_1,\mathrm{lon}_1}, v_{\mathrm{lat}_1,\mathrm{lon}_1}) & \cdots & (u_{\mathrm{lat}_1,\mathrm{lon}_M}, v_{\mathrm{lat}_1,\mathrm{lon}_M}) \\ \vdots & (u_{\mathrm{lat}_i,\mathrm{lon}_j}, v_{\mathrm{lat}_i,\mathrm{lon}_j}) & \vdots \\ (u_{\mathrm{lat}_N,\mathrm{lon}_1}, v_{\mathrm{lat}_N,\mathrm{lon}_1}) & \cdots & (u_{\mathrm{lat}_N,\mathrm{lon}_M}, v_{\mathrm{lat}_N,\mathrm{lon}_M}) \end{bmatrix}$$

La figure 3.1 présente un exemple de champ de vent ECMWF. Les composantes u et v sont stockées dans deux matrices différentes et elles sont exprimées en mètre par seconde ($\mathrm{m\,s^{-1}}$). Pour cette illustration, il est plus pratique d'utiliser l'intensité et la direction du vent dans une seule figure plutôt que les composantes dans deux figures différents. L'intensité de vent est indiquée par une couleur et la direction est indiquée par une flèche. Dans cet exemple, le vent souffle du sud-ouest au nord-est avec une intensité moyenne de $12\,\mathrm{m\,s^{-1}}$.

3.2 La zone d'étude et ses particularités

La zone retenue pour notre étude est la côte ouest de la Norvège autour de Bergen (cf. Figure 3.2). Elle est choisie pour la complexité de sa topographie. La région de Bergen est formée de nombreux *fjords* s'avançant dans les terres

et de nombreuses montagnes très accidentées. Le point culminant est d'environ 2500 m. On y trouve également beaucoup d'îles proches de la côte. Les caractéristiques de ce type de région sont particulièrement intéressantes pour l'étude des effets locaux. Ils sont souvent non observables ou mal prévus par les modèles numériques et en général mieux observés par les mesures satellitaires.

FIGURE 3.2 – *La boîte rouge indique la zone d'étude — Bergen, dans le sud-ouest de la Norvège. Elle se situe à N*59.5°–63° *et à E*1.5°–6.5°.

Au large, la physique de l'atmosphère est en général conditionnée par la circulation à grande échelle et les modèles et les données satellite sont en meilleur accord comparativement aux zones côtières où les dynamiques atmosphériques et océaniques sont beaucoup plus complexes et de plus petites échelles [Walmsley et al. (2001)]. L'influence de la côte et du relief sur le vent à micro-échelle est très bien expliquée dans l'ouvrage de Mayençon [Mayençon (1982)], qui se résume en :

- la vitesse de vent est plus faible en moyenne sur terre que sur mer ;
- le vent modifie le champ de pression au voisinage des côtes et cela d'autant plus que sa vitesse est plus élevée et que le relief a des dimensions verticales et horizontales plus grandes.

Reprenons le deuxième point : l'action du vent sur la pression au voisinage d'une côte est particulièrement nette s'il existe une chaîne de montagne ou des collines voisines du littoral. L'intensité du phénomène et les distances d'action dépendent principalement de la force du vent ainsi que des dimensions de la montagne.

Seules les données à HR permettent d'appréhender l'influence de la côte et du relief sur le vent à micro-échelle. Dans une certaine mesure, l'émulation de vent au large (*offshore*) est relativement simple et directe puisque les effets dus à la topographie, aux obstacles, et aux changements de rugosité multiples ne s'appliquent généralement pas [Walmsley et al. (2001)]. Dans les zones côtières, la prédiction est également compliquée par les changements de rugosité et de stabilité [Pryor et Barthelmie (1998)]. Les variabilités locales dans les zones

côtières se manifestent également par des effets d'accélération et de décélération.

En résumé, les variabilités entre les différentes résolutions sont naturellement plus élevées à l'approche de la côte. Cette conclusion générale se vérifie également sur les données considérées : la figure 3.3 montre la moyenne de la norme de la différence entre les données ECMWF (interpolation la plus proche sur la grille de SAR) et les données SAR pour la zone Bergen. On observe que cette distance moyenne augmente de $1.4\,\mathrm{m\,s^{-1}}$ jusqu'à $3.0\,\mathrm{m\,s^{-1}}$ à l'approche de la côte Bergen.

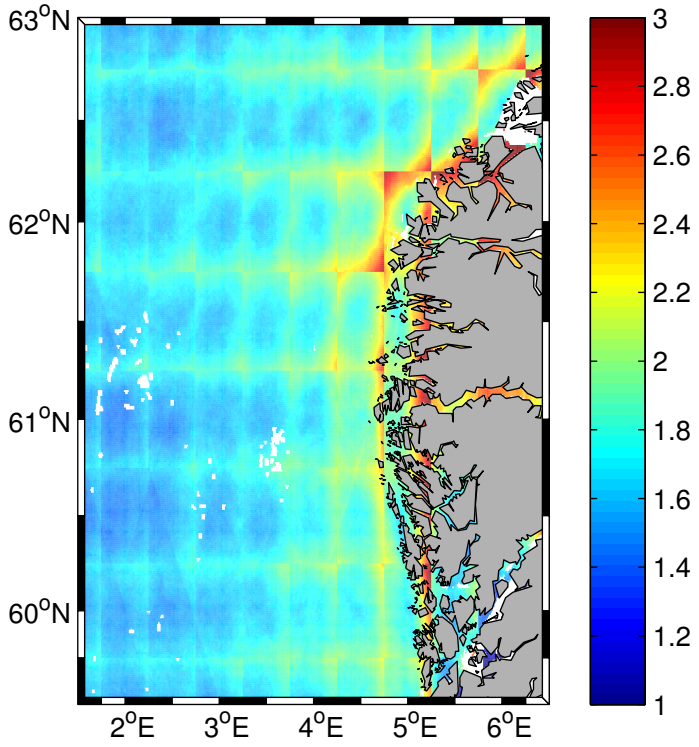


FIGURE 3.3 – *Norme (en* $\mathrm{m\,s^{-1}}$*) de la différence entre les données ECMWF et les données SAR pour la zone Bergen. La distance est calculée pour chaque point sur la grille du SAR avec les 758 paires de deux données co-localisées.*

Les méthodologies d'émulation à HR pourraient différer en fonction des zones choisies. Au large, des méthodes comme le rehaussement de gradient, qui suppose que la variabilité à HR est liée principalement aux phénomènes frontaux, peuvent être utilisées. Dans ce cas, on suppose que la BR est représentative par rapport à la résolution où les donnée se situent. Pour les zones côtières, les variabilités à HR par rapport aux situations à BR peuvent varier énormément d'un endroit à l'autre et il est important d'établir un modèle qui permette de mettre en lien entre les situations à BR et les situations locales pour chaque point spécifique. Dans ce contexte, une méthodologie basée sur un apprentissage statistique est développée et présentée dans le chapitre 5.

3.3 Catalogue de données appariées

Cette partie des travaux consiste à préparer les paires de données historiques à basse et haute résolution co-localisées en temps et espace. Elles sont indispensables pour les approches statistiques car elles fournissent les données d'apprentissage permettant d'établir la relation entre haute et basse résolution. L'ensemble des paires de BR et HR constitue la base de données pour l'apprentissage. On l'appelle ici « catalogue » (cf. Figure 3.4). Le catalogue a un rôle très déterminant. De sa qualité et de sa représentativité dépend très directement la pertinence des modèles d'émulation HR appris.

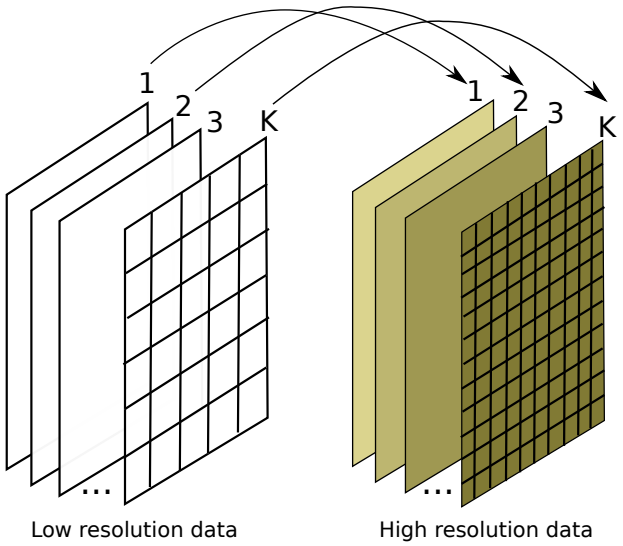


Figure 3.4 – *Paires de haute et basse résolution dans un catalogue*

Pour atteindre ces propriétés, il faut d'abord mettre en place une méthode pour coupler les données à des résolutions différentes. La partie 3.3.2 traite du problème de synchronisation temporelle. Dans la partie 3.3.3, la méthode de sélection des paires de haute et basse résolution qui concernent la même situation de vent est présentée. Les données aberrantes et les points qui contiennent des données aberrantes systématiques sont identifiés et masqués. La partie 3.3.4 donne quelques exemples des données issues du catalogue.

3.3.1 Propriété d'un catalogue

La première propriété, la qualité, exige que les données à haute et basse résolution ne contiennent pas trop d'erreurs et que chaque paire de haute et basse résolution soit synchronisée. Il est normal que les données soient entachées d'erreurs, quel que soit leur résolution et leur moyen d'acquisition. Cependant, il faut que le niveau d'erreur reste acceptable.

Pour la deuxième propriété, un catalogue est d'autant plus représentatif qu'il contient des situations variées. Ces différentes situations permettent aux méthodes statistiques de mieux généraliser la relation entre la haute et la basse résolution.

Une mauvaise qualité ou une mauvaise représentativité d'un catalogue peuvent toutes les deux conduire aux mêmes type de conséquences : elles peuvent fausser les choix des méthodes statistiques ; elles peuvent détériorer la qualité de l'apprentissage voir mettre une méthode statistique en échec. Même si les conséquences sont connues, il est souvent difficile d'obtenir une grande quantité de données, et il est encore plus difficile d'obtenir une grande quantité de données de qualité. Suivant les données disponibles, il est important de bien préparer le catalogue tout en respectant au mieux les deux critères précédents.

3.3.2 Synchronisation des données

Le catalogue est construit de la façon suivante : pour chaque champ de vent à HR, la BR correspondante est choisie comme étant sur la même zone d'étude et la plus proche possible dans le temps. On appelle la première correspondance la « synchronisation en spatial » et la deuxième la « synchronisation en temporel ».

La synchronisation temporelle est un peu compliquée dans notre cas : la résolution temporelle de l'ECMWF est de 6 h et les mesures satellites ne sont pas effectuées à ces heures synoptiques (0 h, 06 h, 12 h, 18 h UTC). Cette différence fait que les deux données ont rarement la même heure. Il existe le même type de problème dans l'assimilation de données météorologiques. Deux solutions sont possibles :

Approche 1 : utiliser les champs de vent ECMWF les plus proches des heures des observations satellitaires ;

Approche 2 : interpoler les champs de vent du modèle aux heures des observations.

Dans la première approche, les hautes et les basses résolutions peuvent avoir une différence de temps d'au plus 3 heures. L'utilisation des champs de vent ECMWF les plus proches des heures d'observation peut introduire des erreurs, qui peuvent même faire apparaître un écart important entre paires de haute et basse résolution. Dans la deuxième approche, l'interpolation peut conduire à un moyennage qui peut dégrader la représentativité, par exemple pour un front bien structuré et net, et introduire des erreurs liées à l'interpolation à l'heure d'observation. Les deux approches ont des avantages et des inconvénients, il est donc nécessaire d'évaluer les erreurs introduites par chacune des approches.

3.3.2.1 Utilisation du champ le plus proche

Rappelons que le fait d'utiliser les champs de vent ECMWF les plus proches aux heures des observations satellite peut introduire jusqu'à 3 h de différence. Dans notre cas, sur la zone Bergen, les différences d'heure entre les observations et les données ECMWF sont d'entre 1 heure et demi et 3 heures. Pour évaluer la différence entre un champ de vent à l'instant t et celui avec

quelques heures de décalage, la technique d'autocorrélation est utilisée. L'idée est de calculer les corrélations croisées d'une série temporelle par elle même, pour obtenir les coefficients de corrélation en décalant de 1 h, 2 h, $\cdots$, N h. Ces coefficients permettent de voir que la variabilité de la série à l'instant t reste la même à l'instant t'.

Pour cela, des séries de vent avec une résolution temporelle la plus fine possible sont nécessaires. La résolution temporelle de ECMWF à 6 h est trop grossière pour pouvoir évaluer l'erreur introduite avec 3 h de décalage maximum. Pour cela, les données de bouée, qui ont une résolution temporelle fine et régulière, sont utilisées. Dans la zone d'étude, une série de vent sur la station Troll à N60.60° et à E3.70° de la Mer du Nord au sud-ouest de la Norvège est utilisée. Cette plate-forme pétrolière fournit l'intensité et la direction du vent en local. Les données sont fournies toutes les 10 minutes par la station mais elles sont moyennées sur une heure pour le produit final. La figure 3.5 donne un exemple d'une série de vent de 2007 à 10 m de la surface de la mer. Son intensité et ses composantes u et v sont illustrées.

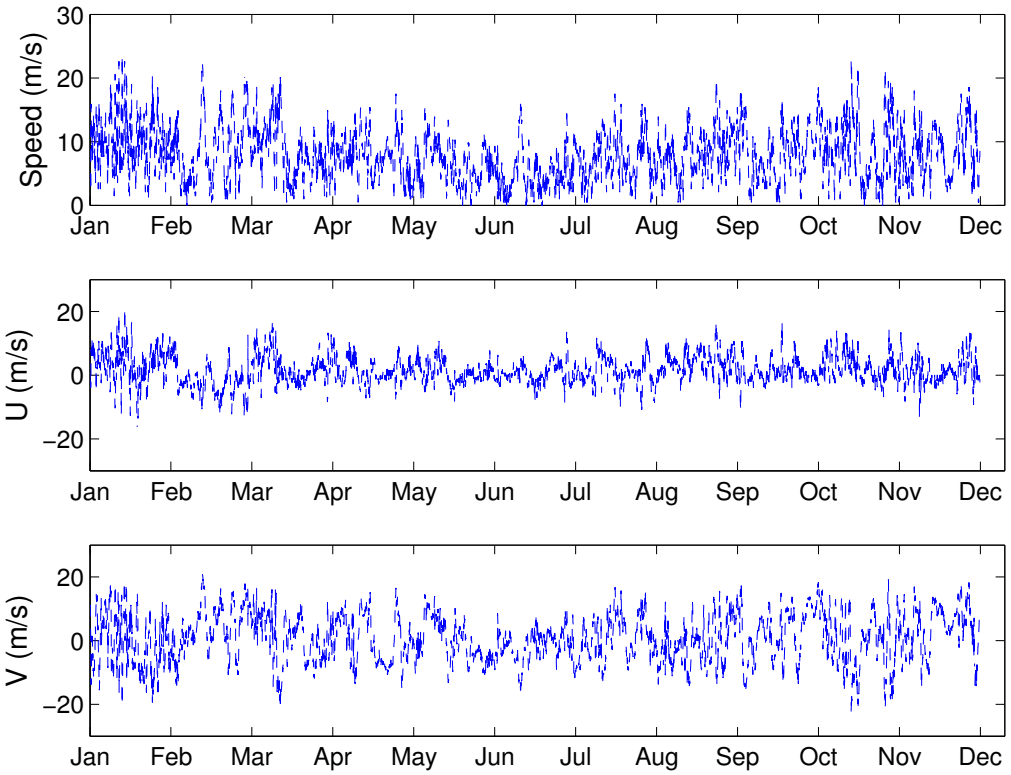


Figure 3.5 – *Série de vent en 2007 sur la station Troll (N60.60°,E3.70°) et ses composantes zonale et méridienne.*

Le tableau 3.3 indique les autocorrélations en intensité du vent, en composante zonale u et en composante méridienne v. Pour une différence temporelle entre 1 h et 3 h, les coefficients varient entre 0.94 et 0.83 pour la composante u et entre 0.97 et 0.91 pour la composante v. On constate que les coefficients sont plus élevés et varient moins pour v. C'est du au fait que la composante méridionale a une plus faible variabilité temporelle que

la composante zonale. Le résultat précédent montre que l'autocorrélation reste forte malgré le décalage temporel (≤ 3 h).

Tableau 3.3 – *Coefficients croisés d'une série temporelle de la station* Troll. Δt *représente le décalage temporel.*

Δt (Heures)	0	1	2	3	4	5	6
Coefficient (intensité)	1.0	0.95	0.91	0.86	0.81	0.76	0.71
Coefficient (U)	1.0	0.94	0.89	0.83	0.78	0.73	0.69
Coefficient (V)	1.0	0.97	0.95	0.91	0.88	0.84	0.81

3.3.2.2 Interpolation linéaire

Une autre approche pour résoudre le problème de synchronisation en temps est d'interpoler les champs de vent du modèle vers les heures des observations. Pour évaluer cette approche, les champs de vent de l'ECMWF ayant une résolution temporelle de 6 h et une série de vent qui a une résolution temporelle plus petite, par exemple des mesures par bouée, sont utilisées. L'idée est d'interpoler les champs ECMWF sur l'endroit et sur les horaires de la série de mesures et ensuite de la comparer avec cette dernière.

La même série prise sur la station Troll (cf. Figure 3.5) est utilisée pour estimer l'erreur de conversion (§ 3.3.2.1). La station fournit les intensités et les directions des vents toutes les heures. On peut en déduire les coordonnées cartésiennes u et v. Pour l'obtention des informations u et v des données ECMWF, une interpolation linéaire est utilisée pour interpoler les données ECMWF au même endroit et aux mêmes horaires que la station Troll.

Le tableau 3.4 affiche les coefficients de corrélation par saison entre la série temporelle mesurée et les données interpolées de l'ECMWF. Les coefficients varient entre 0.90 et 0.94 pour la composante u et entre 0.96 et 0.98 pour la composante v. Les coefficients sont plus élevés au printemps et à l'automne pour la composante u contrairement à ceux pour la composante v.

Tableau 3.4 – *Coefficients de corrélation par saison entre la série temporelle sur la station* Troll *et les données ECMWF qui sont interpolées sur le même endroit et sur les mêmes horaires.*

Saison	Printemps	Été	Automne	Hiver
Coefficient (intensité)	0.92	0.93	0.93	0.95
Coefficient (U)	0.94	0.90	0.93	0.90
Coefficient (V)	0.96	0.98	0.96	0.98

La figure 3.6 montre l'intensité de ces deux séries en fonction de la saison. La figure 3.7 montre un zoom sur une partie du mois janvier pour pouvoir mieux apprécier la différence entre la série interpolée et la série de référence.

À partir de la figure 3.7, on peut remarquer que les données ECMWF (marquées par des étoiles rouges) sont assez cohérentes avec les mesures *in-situ* (ligne bleue). Par contre, les résultats de l'interpolation en temps et en espace

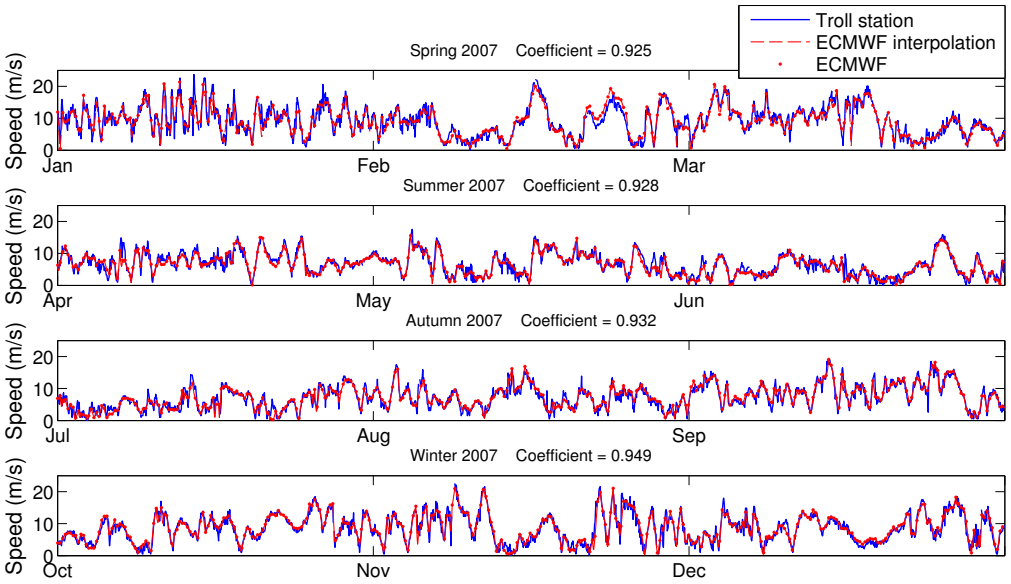


FIGURE 3.6 – *Comparaison entre une série interpolée de ECMWF (rouge étoilé) en temps et en espace et une série de référence (bleu) mesurée sur la station* TROLL *à N60.60° et à E3.70°. La corrélation entre deux série est tracée par saison et les coefficients de corrélation sont affichés en haut de chaque figure.*

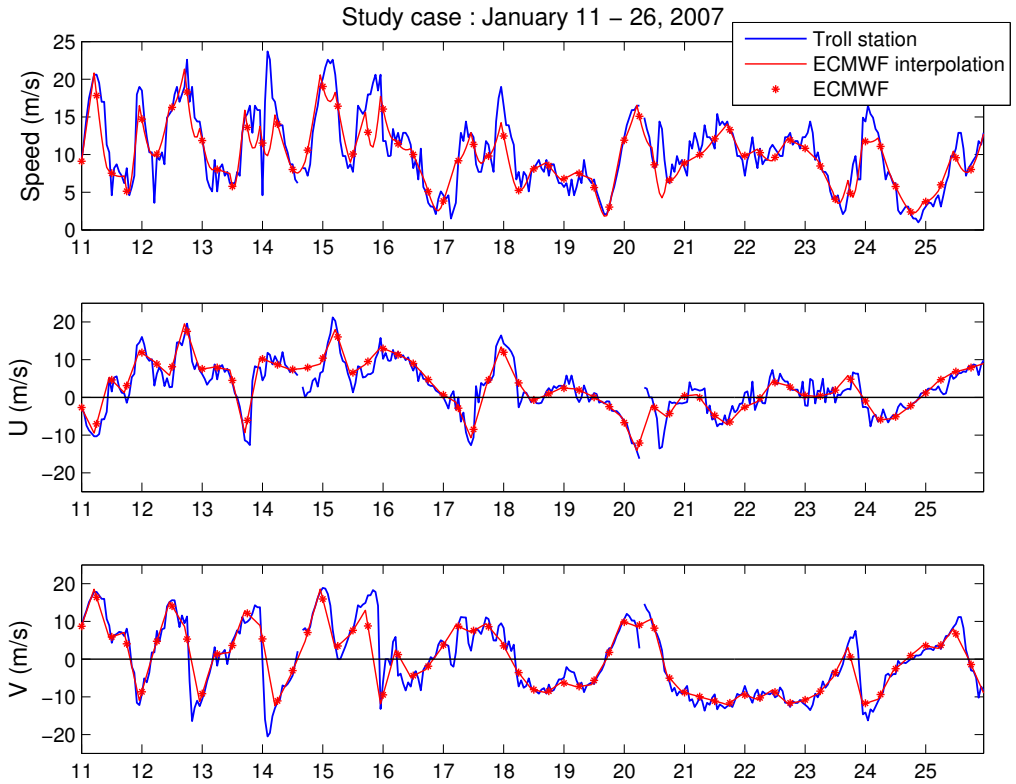
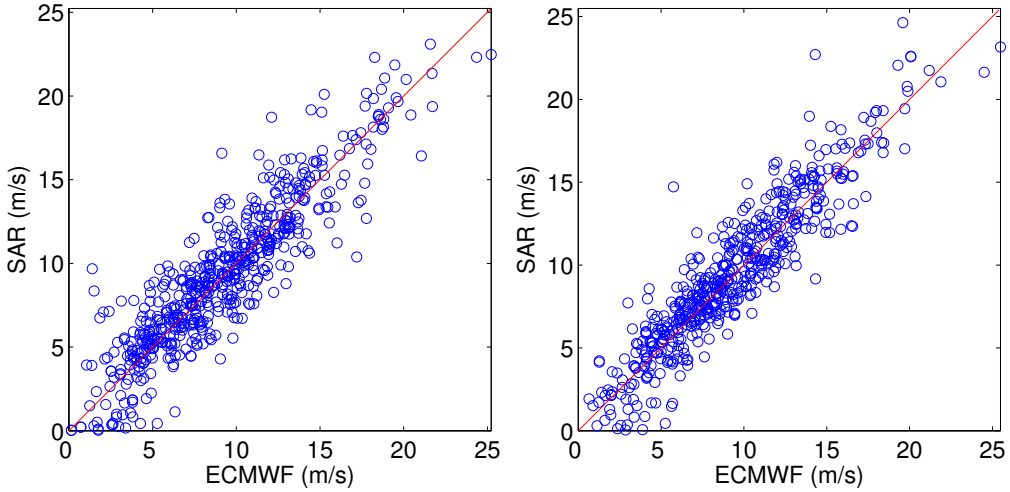


FIGURE 3.7 – *Zoom de la figure 3.6 en janvier. Les comparaison se font en intensité, suivant la composante u et suivant la composante v.*

(ligne rouge étoilée) peuvent être très différents des mesures par endroits. Par exemple, autour du 20 janvier 2007, un changement rapide du vent sur une durée très courte entre 12 h et 18 h fait que les interpolations se trompent complètement, alors que les prévisions de l'ECMWF de 12 h et 18 h sont très proches des mesures.

Compte tenu des erreurs potentiellement introduites par l'interpolation linéaire et de la préférence donnée au fait de conserver des champs structurés, la première approche est préférable. La figure 3.8 illustre les dispersions de la vitesses de vents entre les données SAR et les données ECMWF les plus proches du temps SAR au point (N61.86°, E3.13°) et au point (N61.86°, E3.13°). Après suppression des paires de données à basse et haute résolution trop différentes (§ 3.3.3), les données SAR et les données ECMWF restent très proches, malgré les décalages temporels.



(a) *Point (N61.86°, E3.13°)* (b) *Point (N62.56°, E2.38°)*

FIGURE 3.8 – *Diagramme de dispersion de la vitesses des vents entre les données SAR et les données ECMWF sur les deux points au large (cf. Zone illustrée par figure 3.3) : point (N61.86°, E3.13°) (a), point (N61.86°, E3.13°) (b). L'axe des abscisses est de l'ECMWF et l'axe des ordonnés est du SAR.*

3.3.3 Sélection des données

On cherche à évaluer les deux approches pour résoudre le problème de synchronisation temporelle entre les données SAR et les données ECMWF. La première approche, qui consiste à utiliser directement les champs de vent ECMWF les plus proches aux heures des observations satellite, est sélectionnée pour pouvoir préserver les structures à fort gradient (fronts par exemple) bien nettes. Cependant, les champs ECMWF utilisés risquent d'avoir un grand écart par rapport aux champs d'observations. Pour ne pas dégrader la qualité du catalogue en choisissant cette approche, une étape d'élimination des paires de données trop incohérentes est nécessaire.

Cette étude utilise environ 5 années de données SAR, correspondant à une période allant de 2005 à début 2010. Chaque donnée SAR est couplée avec la donnée ECMWF la plus proche de son heure d'observation, ce qui constitue un catalogue de 860 paires de basse et haute résolution au total. Les paires qui ont de grandes différences doivent être supprimées. Pour ces éliminations, il y a plusieurs éléments à prendre en compte :

- Il faut d'abord analyser les données à basse et haute résolution et identifier les données aberrantes point par point. Une donnée aberrante est définie comme une observation qui se trouve très « loin » des autres observations. Elle est souvent isolée par rapport aux autres points et elle peut être liée à des erreurs techniques ou de mesure, par exemple.
- La sélection d'une paire dans le catalogue doit se fonder sur la comparaison des différences entre BR et HR vis-à-vis des variabilités observées localement, sachant que ces variabilités sont spatialement non-homogènes. Un des plus grands avantages des données SAR est leur amélioration sur ces zones où les données du modèle numérique sont généralement moins bonnes, ce qui fait qu'il est normal que la variabilité entre SAR et ECMWF soit plus importante dans la zone côtière ou *fjord*. Cette variabilité est une information très importante pour l'émulateur et doit être conservée.
- Il est important de choisir un seuil de suppression juste pour un bon compromis entre la qualité du catalogue et la quantité des données restantes. Un seuil de suppression trop rigide risque d'éliminer trop de données et de favoriser davantage le comportement linéaire entre les hautes et les basses résolutions, puisque les données restantes ont peu de différences.

La sélection des données dans le catalogue se fait en prenant en compte tous ces éléments. Les sections suivantes introduisent les critères et méthodes considérés pour chaque aspect.

3.3.3.1 Données aberrantes

Pour avoir une idée globale sur la qualité des données, on trace une carte des moyennes de la norme de la différence entre les données SAR et les données ECMWF. La moyenne au point (p, q) est définie par :

$$\mathrm{dist}_{p,q} = \frac{\sum_{i=1}^{N_{p,q}} \sqrt{(u^{\mathrm{HR}}{}_{p,q,i} - u^{\mathrm{BR'}}{}_{p,q,i})^2 + (v^{\mathrm{HR}}{}_{p,q,i} - v^{\mathrm{BR'}}{}_{p,q,i})^2}}{N_{p,q}} \tag{3.2}$$

où $N_{p,q}$ est le nombre d'observations SAR disponibles et $\langle u^{\mathrm{HR}}{}_{p,q,i}, v^{\mathrm{HR}}{}_{p,q,i} \rangle$ est le i^{e} vecteur de vent SAR au point (p, q). $\langle u^{\mathrm{BR'}}{}_{p,q,i}, v^{\mathrm{BR'}}{}_{p,q,i} \rangle$ est le i^{e} vecteur de vent interpolé (interpolation linéaire) au point (p, q) à partir des données ECMWF de la grille à HR.

La figure 3.9 montre une variabilité plus élevée en zones côtières et dans les *fjords* (zone 1 et zone 3). Pour les points brillants dans la zone 2 et 4, il est possible qu'il existe quelques données aberrantes ou des erreurs systématiques

dans les mesures SAR. Les données ECMWF du modèle sont normalement déjà validées et les données SAR peuvent contenir des données aberrantes si elles n'ont pas fait l'objet d'un pré-traitement spécifique.

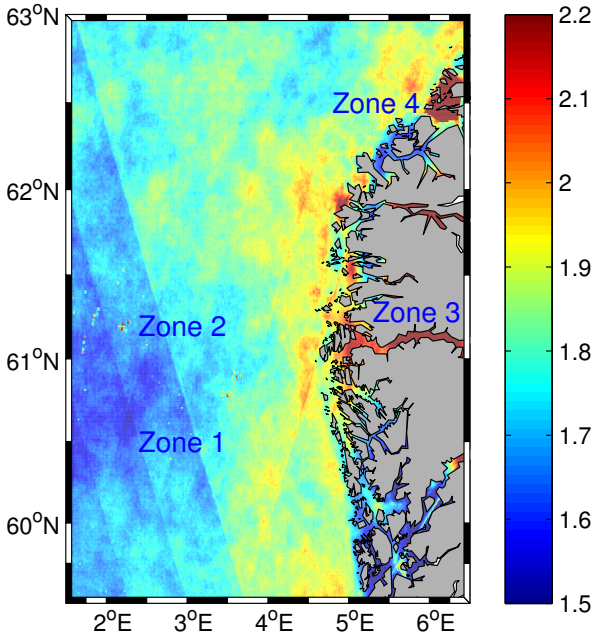
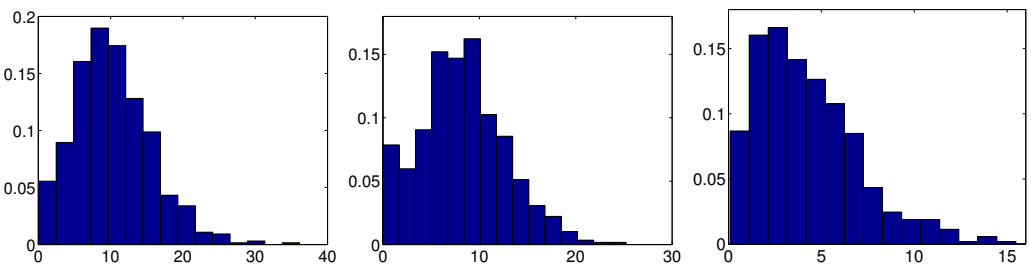


FIGURE 3.9 – *Moyenne (en* $\mathrm{m\,s^{-1}}$*) de la norme de la différence entre SAR et ECMWF pour chaque point de la zone d'étude. La zone 1 représente la zone au large où la différence entre SAR et ECMWF est la moins grande. Les zones 2 et 4 indiquent des points de brillance où il peut exister des différences systématiques. La zone 3 représente les fjords.*

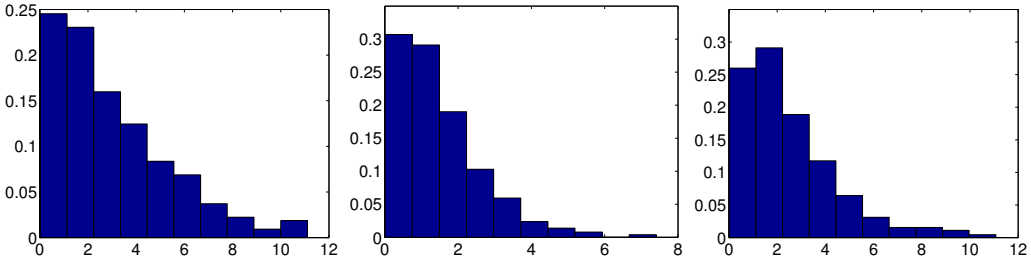
Les histogrammes des intensités des vents SAR au point (N61.21°, E2.27°) de la zone 2, au point (N60.0°, E2.0°) de la zone 1 et au point (N61.14°, E5.99°) de la zone 3 sont tracés figure 3.10. Les vents SAR très forts ($> 25\,\mathrm{m\,s^{-1}}$) sont observés au point (N61.21°, E2.27°) (3.10a). La plage supérieure de la validité des données SAR est fixée à $25\,\mathrm{m\,s^{-1}}$ à cause de la limitation de la Geophysical Model Function (GMF) aux vents forts [Monaldo et al. (2003)]. Les observations des vents supérieures n'ont pas été validées, elles sont donc masquées et ne doivent pas être utilisées.



(a) *Point (N61.21°, E2.27°), zone 2* (b) *Point (N60.0°, E2.0°), zone 1* (c) *Point (N61.14°, E5.99°), zone 3*

FIGURE 3.10 – *Histogramme des intensités des vents* $\mathrm{m\,s^{-1}}$ *en un point dans la zone 1 (b), 2 (a) et 3 (c) respectivement. Les zones sont indiquées sur la figure 3.9*

Malgré l'élimination des données aberrantes, des points qui ont des valeurs élevées dans la zone 2 ou 4 existent toujours. Si on trace le histogramme de la norme de la différence à un point de la zone 2, des erreurs élevées systématiques sont observées par rapport à un point de la zone 1 ou de la zone 3 (cf. Figure 3.11). En comparant ces points avec les positions des plate-formes pétrolières, il est intéressant de noter que la majorité des points se situe dans le voisinage de ces plate-formes.



(a) *Point (N61.17°, E2.19°), zone 2* (b) *Point (N60.0°, E2.0°), zone 1* (c) *Point (N61.14°, E5.99°), zone 3*

FIGURE 3.11 – *Histogramme de la norme de la différence (en* $\mathrm{m\,s^{-1}}$*) entre SAR et ECMWF en un point dans la zone 1, 2 et 3 respectivement. Les zones sont indiquées sur la figure 3.9*

En prenant en compte les plate-formes pétrolières, la présence de la terre et les points qui ont des valeurs systématiquement élevées, on génère un masque de points représenté figure 3.12. Ces points blancs ne sont plus utilisés et il n'y a pas d'émulation.

3.3.3.2 Seuil de suppression

Pour le deuxième élément, en prenant en compte le fait que la variabilité dans la zone côtière est normalement plus élevée qu'au large, une solution est de comparer seulement les vents à HR, dégradés à l'échelle de BR avec les vents à BR au large.

Pour ce dernier élément, il faut d'abord fixer un seuil de suppression. Les paires qui ont une différence supérieure à ce seuil sont éliminées du catalogue. Il y a deux façons de fixer le seuil de suppression : calculer la distance absolue ou calculer la distance relative entre la basse et la haute résolution. La distance absolue de la i^{e} paire est définie par :

$$\alpha_i = \frac{\sum_{(p',q')\in\varphi_i} \sqrt{(u^{\mathrm{HR}'}{}_{p',q',i} - u^{\mathrm{BR}}{}_{p',q',i})^2 + (v^{\mathrm{HR}'}{}_{p',q',i} - v^{\mathrm{BR}}{}_{p',q',i})^2}. * \mathrm{mask}_{\mathrm{coast}}}{K_i} \tag{3.3}$$

où $(u^{\mathrm{HR}'}, v^{\mathrm{HR}'})$ sont les champs de vent SAR dégradés (moyennés) vers la grille à BR et $(u^{\mathrm{BR}}, v^{\mathrm{BR}})$ sont les données ECMWF. φ_i est l'ensemble des points (p', q') où les vents SAR sont disponibles pour la i^{e} donnée de la grille à BR. $\mathrm{mask}_{\mathrm{coast}}$ est le masque de la zone côtière, obtenu en éloignant la côte de 50 km, et K_i est le nombre de points où les vents SAR sont disponibles après avoir appliqué le masque.

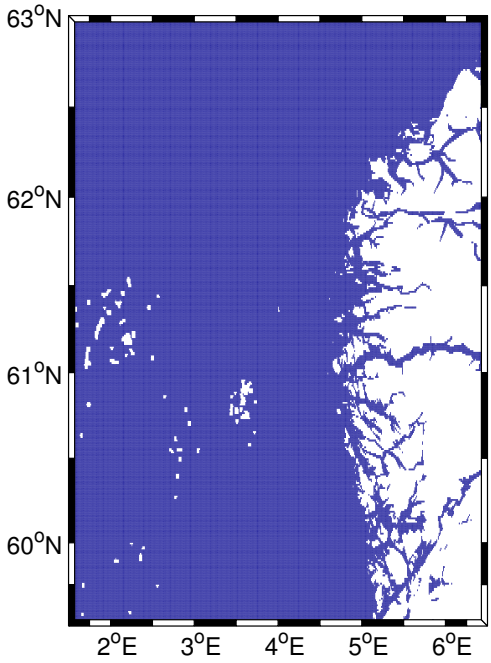


FIGURE 3.12 – *Carte de masque. Le blanc indique les points qui ne sont pas émulés.*

Le problème du critère de distance absolue est qu'il a tendance à plus supprimer les vents forts. Par exemple, une distance de $3\,\mathrm{m\,s^{-1}}$ parait beaucoup pour un vent faible mais moins pour un vent fort. Dans ce cas, une distance relative par rapport à l'intensité en moyenne d'un champ de vent ECMWF est plus adaptée. Elle est donnée par :

$$\beta_i = \frac{\alpha_i}{\overline{\mathrm{ECMWF}_i}} \tag{3.4}$$

$$\text{avec} \qquad \overline{\mathrm{ECMWF}_i} = \frac{\sum_{(p',q')\in\varphi'_i} \sqrt{(u^{\mathrm{BR}}{}_{p',q',i})^2 + (v^{\mathrm{BR}}{}_{p',q',i})^2}. * \mathrm{mask}_{\mathrm{coast}}}{K'_i} \tag{3.5}$$

où $\overline{\mathrm{ECMWF}_i}$ est la moyenne en intensité d'un champ de vent ECMWF sur la zone au large et φ'_i est l'ensemble des points (p', q') où les vents ECMWF sont disponibles pour la i^{e} donnée après avoir appliqué le masque.

Pour une valeur fixée de paramètre β, les paires qui ont une distance relative inférieure à β sont conservées. La figure 3.13 trace le nombre de paires restantes dans le catalogue en fonction du seuil de suppression β. Plus la valeur de β est grande, plus la différence entre les deux données est tolérée, et plus de paires sont conservées dans le catalogue.

Les conséquences d'un seuil de suppression trop faible sont :

- trop de paires de données sont supprimées et il n'y a pas assez de données pour que le catalogue reste représentatif ;
- cela peut favoriser la dépendance linéaire entre la basse et la haute résolution.

En fonction des paires restantes pour chaque seuil β, la moyenne de l'intensité du vent pour chaque donnée ECMWF restante est calculée selon l'équation

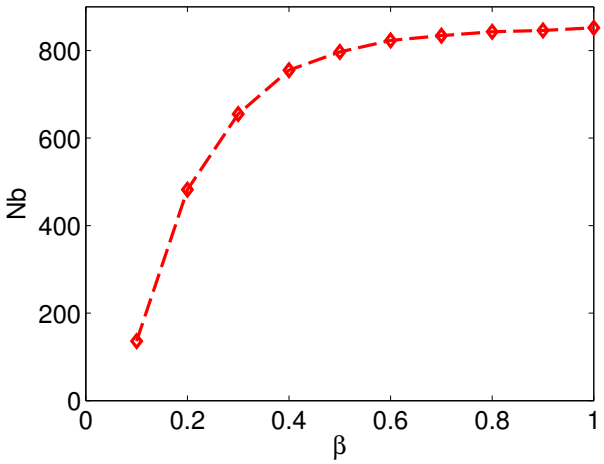
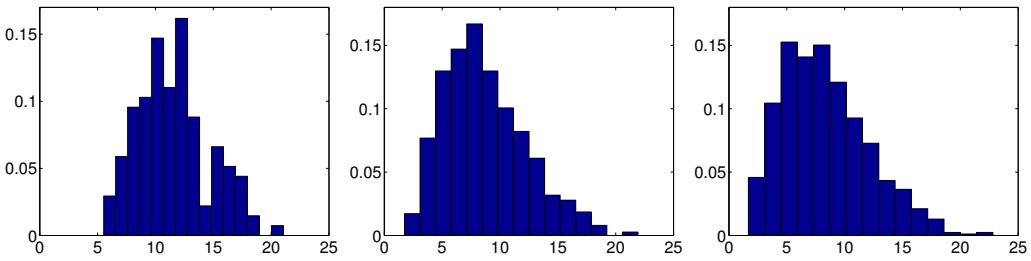


FIGURE 3.13 – *Nombre de paires de données ECMWF et SAR conservées dans le catalogue en fonction du seuil de suppression fixé.*

(3.5). La figure 3.14 illustre les histogrammes des intensités des vents ECMWF restants pour les différents seuils. Cela nous permet d'apprécier la représentativité du catalogue après les suppressions. Quand le seuil est très petit, par exemple, $\beta = 0.1$ (cf. Figure 3.14a), il supprime visiblement beaucoup de vents faibles. Par conséquent, la distribution des vents est modifiée et ses caractéristiques ne sont plus respectées. La distribution des vents pour $\beta = 0.4$ reste proche de celle en conservant presque tout le catalogue ($\beta = 1.0$).



(a) $\beta = 0.1$ (b) $\beta = 0.4$ (c) $\beta = 1.0$

FIGURE 3.14 – *Histogramme de la moyenne de l'intensité du vent* ($\mathrm{m\,s^{-1}}$) *par chaque champ ECMWF restante dans le catalogue en fonction du seuil de suppression* β.

Pour un équilibre entre la qualité et la représentativité du catalogue, le seuil de suppression est finalement fixé à 0.4 ce qui fait un catalogue de 758 paires hautes et basses résolutions.

3.3.4 Exemples du catalogue

La figure 3.15 donne un exemple de paires des deux données issues du catalogue. Les données ECMWF sont disponibles pour tous les points et les données SAR couvrent souvent partiellement la zone d'étude, selon le passage et la fauchée du satellite.

Comme les données SAR ne sont pas disponibles partout, le nombre de données disponibles en chaque point n'est pas le même. La figure 3.16 illustre le nombre de données SAR disponibles pour la grille à HR.

Dagestad et al. (2009) étudient le nombre de données SAR nécessaires en chaque point pour que le catalogue soit représentatif. En calculant la moyenne de l'intensité du vent SAR en fonction du nombre de données disponibles, ils montrent que la courbe de la moyenne se stabilise à partir d'environ 600 données pour tous les points et ils en concluent qu'au delà de cette valeur le catalogue a une bonne représentativité.

La figure 3.16 montre que le nombre de données disponibles est d'environ 450 dans certaines zones *fjords*. L'apprentissage en ces points peut être plus difficile et les erreurs d'émulation peuvent être plus élevées par rapport aux autres points.

3.4 Conclusion

Ce chapitre introduit les données expérimentales, la zone d'étude choisie et la construction de catalogue.

La présentation des sorties du modèle numérique, des observations par télédétection et du concept d'échelle et de résolution permettent de mieux appréhender les données ECMWF et SAR et de comprendre les différences observées entre les données basse et haute résolution.

La zone d'étude choisie est la côte ouest de la Norvège. La différence de la variabilité à HR au large et en zones côtières est d'autant plus grande que les effets locaux sont amplifiés par la complexité de sa topographie, ce qui en fait un cas d'étude très intéressant pour l'émulation à HR.

Pour la préparation du catalogue, plusieurs problèmes ont été abordés :

- la synchronisation temporelle du fait de la différence des résolutions temporelles des données BR et HR ;
- la nécessité de distinguer la variabilité intrinsèque de la variabilité générée par des incohérences entre les sorties du modèle numérique ECMWF et les observations SAR ;
- la nécessite d'un compromis entre la cohérence entre les données BR et HR et la représentativité du catalogue.

La construction du catalogue de paire de données ECMWF et SAR résulte finalement de deux principes :

1. L'association d'une donnée SAR sur la zone d'étude avec la donnée ECMWF disponible la plus proche temporellement ;

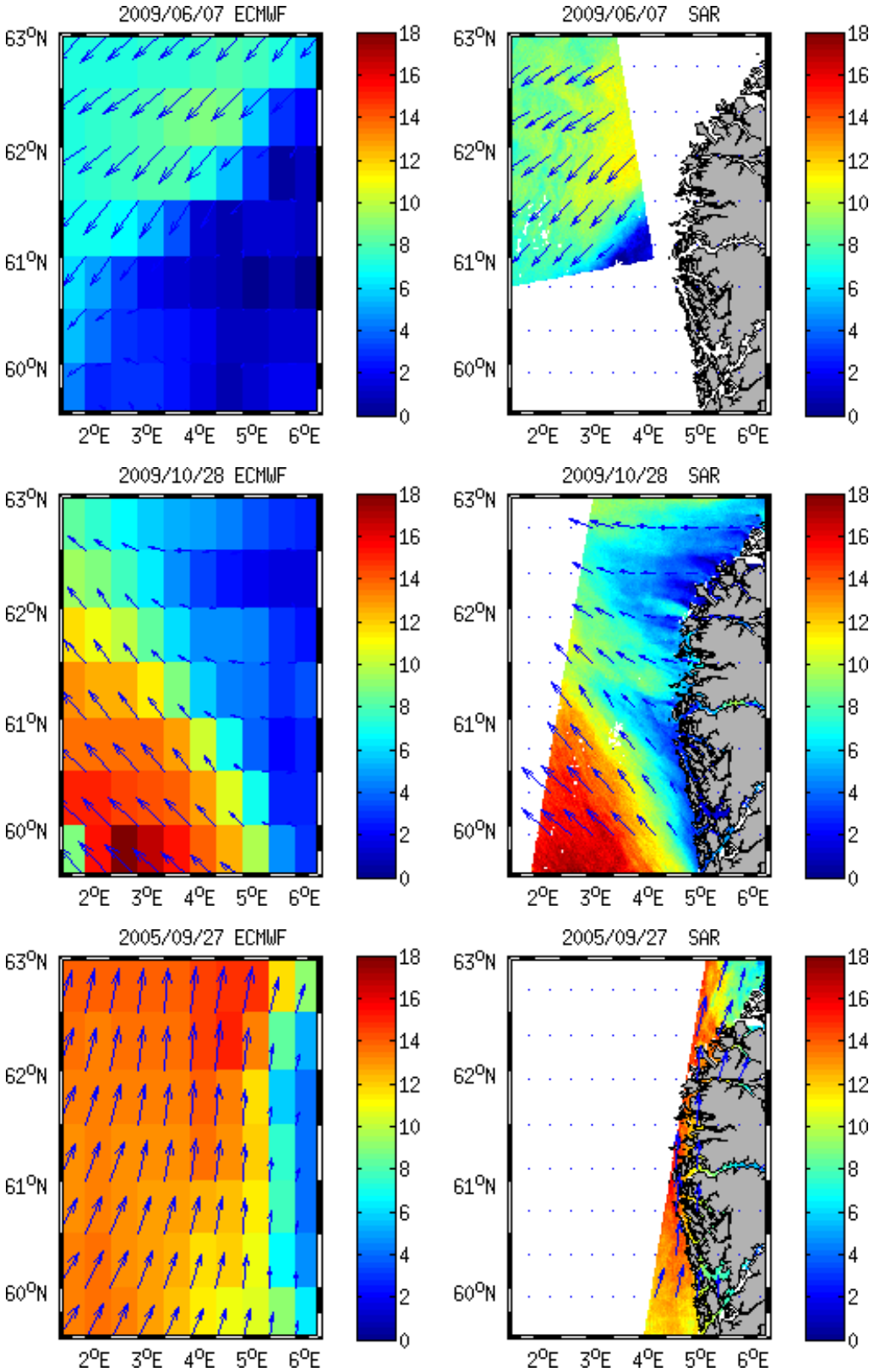


FIGURE 3.15 – *Exemples des paires de l'ECMWF et du SAR. Les flèches indiquent les directions de vent et les couleurs représentent leurs intensités (en* $\mathrm{m\,s^{-1}}$*). La couleur grise indique la terre.*

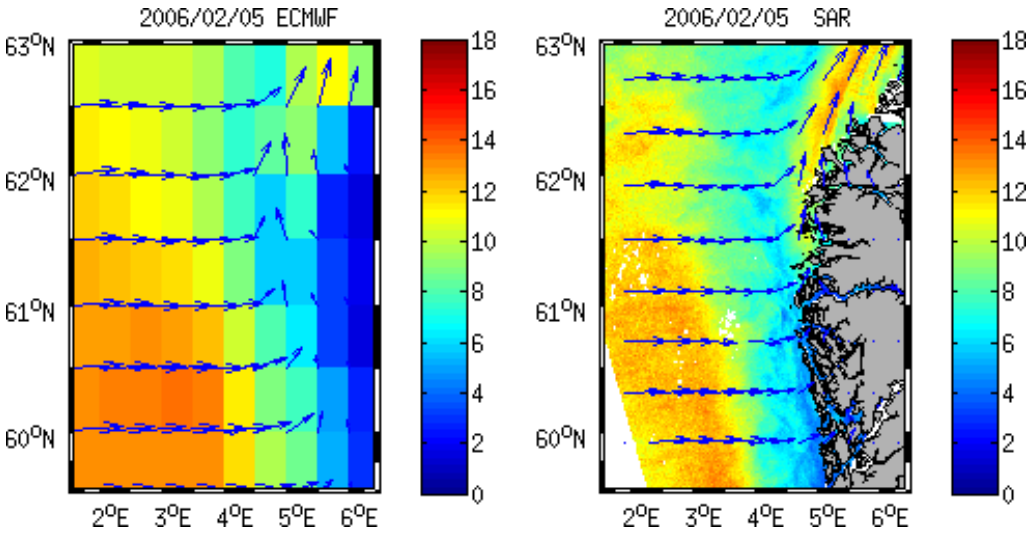


FIGURE 3.15 – *Suite des exemples des paires de l'ECMWF et du SAR.*

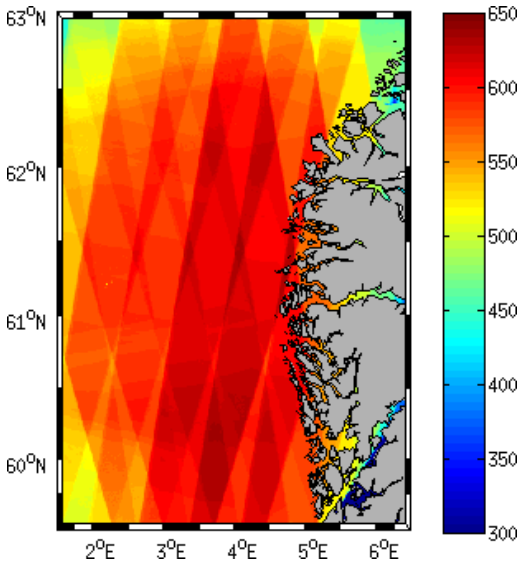


FIGURE 3.16 – *Distribution spatiale du nombre de paires de données ECMWF et SAR pour la zone d'étude considérée.*

2. La sélection de paires de données ECMWF et SAR présentant une similarité supérieure à un certain seuil, similarité évaluée uniquement sur la zone *offshore*. Le choix du seuil de similarité réalise un compromis entre cohérence et représentativité du catalogue.

Analyse statistique conjointe 4

L'objectif de ce chapitre est d'identifier les différences entre les données à basse et haute résolution ainsi que les contraintes physiques à prendre en compte pour l'émulation des champs de vent à HR. L'analyse conjointe entre BR et HR permet également d'illustrer les effets locaux représentés par les données SAR et de montrer l'intérêt des données à HR.

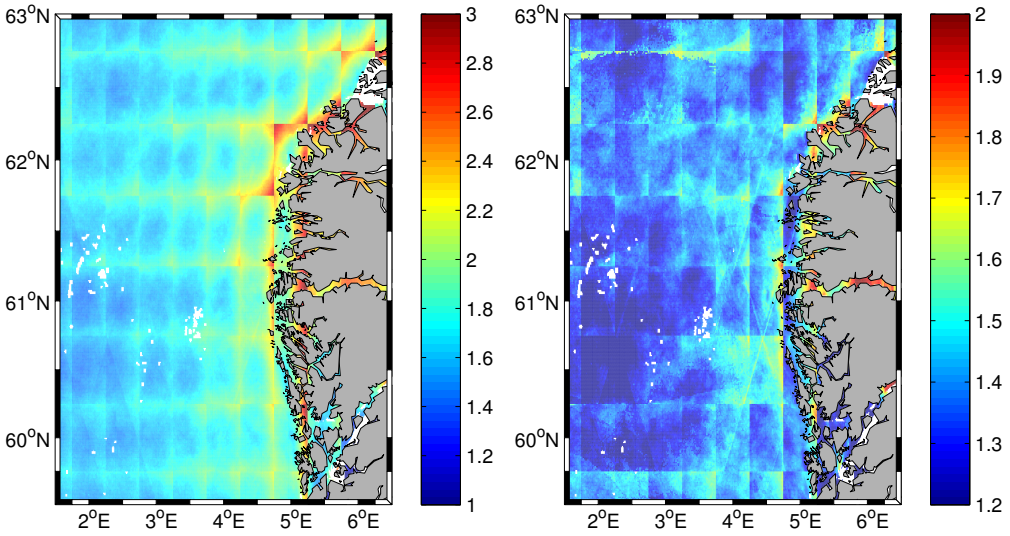
Les paramètres statistiques simples comme, par exemple, la moyenne, l'écart type et la corrélation entre la basse et la haute résolution pour chaque point, permettent d'avoir une première estimation de la distribution de leurs différences. Ensuite, des analyses plus locales, comme la technique de rose des vents pour un point spécifique, permettent d'afficher et de comparer les distributions des directions et des intensités des vents. Un diagramme de dispersion peut illustrer directement la relation entre la basse et la haute résolution. Les analyses de comportement du vent en fonction de la distance de la côte vers l'océan montrent également comment les variabilités entre les deux résolutions évoluent d'un point à l'autre et illustrent par exemple le fait que les données SAR sont plus adaptées pour décrire les dynamiques à petite échelle.

4.1 Analyses globales

Les analyses globales se font sur toute la zone d'étude. La figure 4.1 montre les cartes de moyenne et d'écart type de la norme des différences entre ECMWF et SAR. Elles sont calculées ainsi : pour un point (p, q) de la grille à HR, la norme entre le vecteur vent ECMWF (interpolé au point le plus proche du point (p, q)) et le vecteur vent SAR est calculée pour chaque paire où ECMWF et SAR sont disponibles simultanément. La moyenne et l'écart type sont calculés sur toutes les distances normalisées.

La figure 4.1 montre que la moyenne et l'écart type sont tous les deux plus élevés dans les zones côtières et dans les *fjords*. Un effet de « tuile » peut être observé, correspondant à la grille utilisée par ECMWF. À partir de ces deux cartes, certains points où la moyenne et l'écart type des distances normalisées entre ECMWF et SAR sont élevés peuvent être sélectionnés pour des analyses plus locales (cf. Section 4.2).

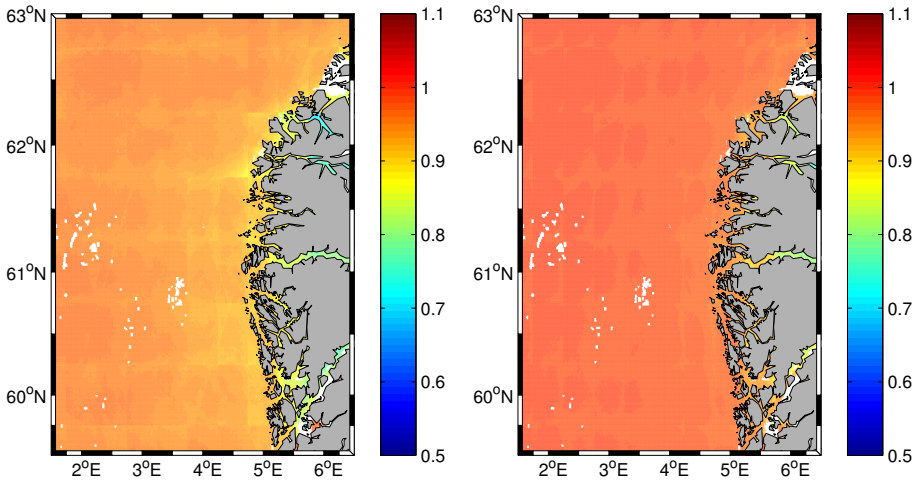
La même tendance peut être observée sur les cartes de coefficient de corrélation entre ECMWF et SAR pour la composante u (cf. Figure 4.2a) et



(a) *Moyenne* (b) *Écart-type*

FIGURE 4.1 – *Moyenne et écart type (en* $\mathrm{m\,s^{-1}}$*) de la norme des différences entre ECMWF et SAR.*

pour la composante v (cf. Figure 4.2b). Les coefficients sont plus élevés pour la composante v, car la composante méridionale est en général moins variable que la composante zonale qui évolue souvent moins vite que la composante zonale.



(a) *Composante u* (b) *Composante v*

FIGURE 4.2 – *Coefficient des corrélations entre ECMWF et SAR.*

Pour analyser la relation entre ECMWF et SAR plus en détail, on sélectionne une série de points de la grille à HR qui sont représentatifs de zones ayant des caractéristiques géographiques différentes et qui ont des variabilités élevées à HR dans les zones côtières par rapport aux analyses globales (§ 4.1).

Pour le premier critère, trois zones principales (cf. Figure 4.3) sont identifiées : zone *fjord*, zone *côtière*, et zone *offshore* (au large). La zone *fjord* est

caractérisée par des vallées glaciaires envahies par la mer, souvent encadrées de hautes montagnes ; la zone côtière est caractérisée par de forts gradients continent/mer, de nombreuses petites îles et presqu'îles ; la zone *offshore* est caractérisée par un environnement relativement simple [Walmsley et al. (2001)].

Pour le deuxième critère, les points 1 à 8 sont également sélectionnés en fonction de la variabilité entre les hautes et basses résolutions. En ces points, la moyenne et l'écart type de la norme de la différence y sont plus élevés et ils sont particulièrement intéressants pour analyser les différences entre les deux données.

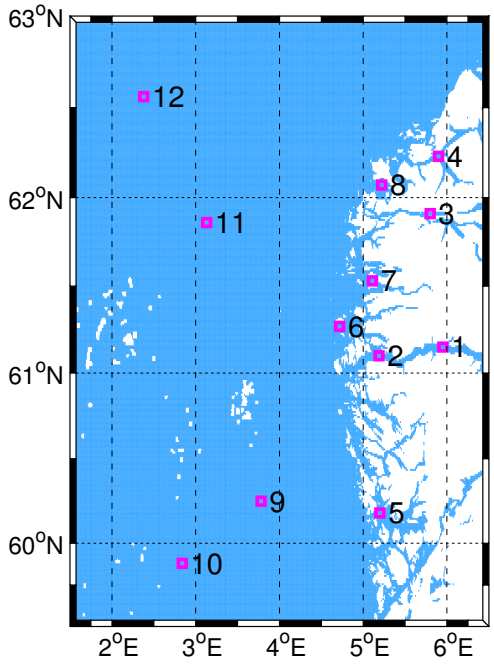


FIGURE 4.3 – *12 points sélectionnés pour les analyses locales.*

Tableau 4.1 – *Coordonnées de 12 points sélectionnés.*

	Point	Latitude (°)	longitude(°)
Zone *fjord*	1	61.15	5.95
	2	61.10	5.19
	3	61.91	5.80
	4	62.23	5.90
Zone côtière	5	60.18	5.20
	6	61.27	4.72
	7	61.53	5.11
	8	62.07	5.22
Zone *offshore*	9	60.25	3.78
	10	59.88	2.84
	11	61.86	3.13
	12	62.56	2.38

Pour toutes les analyses suivantes et pour chacun des 12 points, nous

notons, pour chaque point (p, q) à HR, (p', q') le point à BR le plus proche du point (p, q) en distance. Pour chaque donnée SAR au point (p, q), on utilise donc la donnée ECMWF au point (p', q').

4.2 Analyses locales

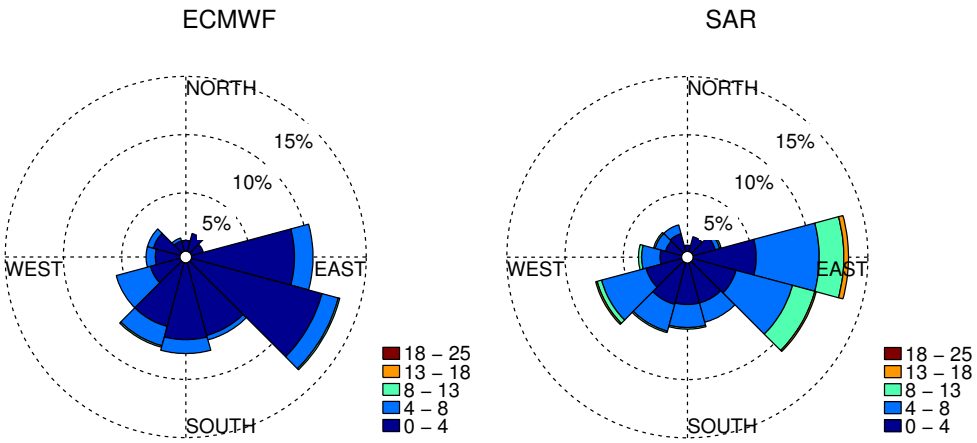
4.2.1 Roses des vents

Une rose des vents indique les pourcentages de données par secteur de direction et la distribution de leurs intensités dans chaque secteur (cf. Figure 4.4). La comparaison des roses des vents entre les données ECMWF et les données SAR peut montrer les différences des deux données ainsi que les différents comportements en fonction des positions géographiques.

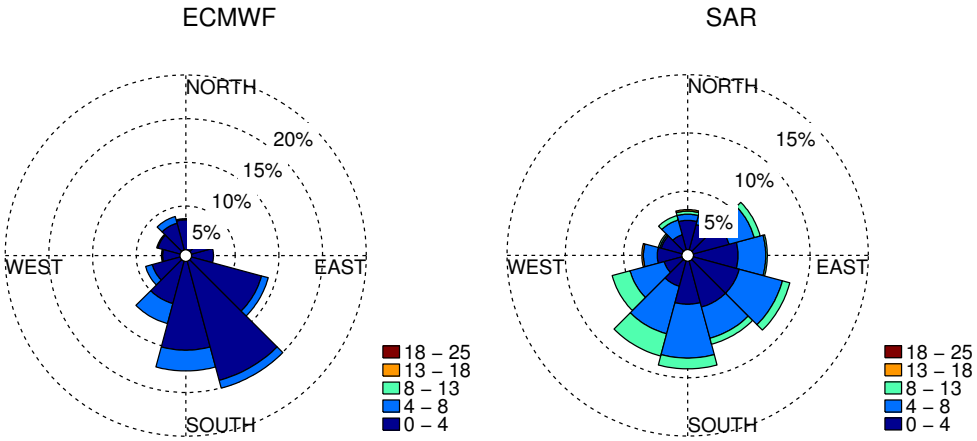
La direction du vent, par convention celle d'où vient le vent, est répartie sur 360° au compas. Le nord est par convention indiqué en haut du diagramme, à 360° (0°). L'est est à 90°, le sud à 180° et l'ouest à 270°. Dans l'exemple de la figure 4.4a, les roses des vents au point 1 avec un pas de 30° indiquent qu'environ 10 % de vent est dans le secteur est pour les données ECMWF contre 13 % pour les données SAR. Au niveau de la distribution de la vitesse du vent dans le même secteur, 8 % et 2 % de vent sont respectivement entre $0\,\mathrm{m\,s^{-1}}$ à $4\,\mathrm{m\,s^{-1}}$ et $4\,\mathrm{m\,s^{-1}}$ à $8\,\mathrm{m\,s^{-1}}$ pour l'ECMWF, ce qui fait une totalité de 10 % de vent d'est. En revanche, la rose des vents pour les données SAR indique 2 % de vent fort entre $8\,\mathrm{m\,s^{-1}}$ à $18\,\mathrm{m\,s^{-1}}$, vent qui n'appairait pas sur les données ECMWF.

Les roses des vents en 9 points parmi les 12 sont choisies pour illustration (cf. Figure 4.4). Les données ECMWF ne sont pas utilisées quand les données co-localisées du SAR manquent : il y a donc autant de données ECMWF que de données SAR en chaque point. Globalement, les différences de roses des vents entre ECMWF et SAR aux points dans les *fjords* (cf. Figure 4.4 – *Zones fjords*) sont plus grandes qu'aux points côtiers (cf. Figure 4.4 – *Zones côtières*) et encore plus grande qu'aux points au large (cf. Figure 4.4 – *Au large*). Aux points 1, 3, 4, il y a plus de vents forts du SAR que de l'ECMWF, contrairement aux points 5, 7, 8. Les roses de vents restent très proches entre ECMWF et SAR pour les points au large (10, 11 et 12).

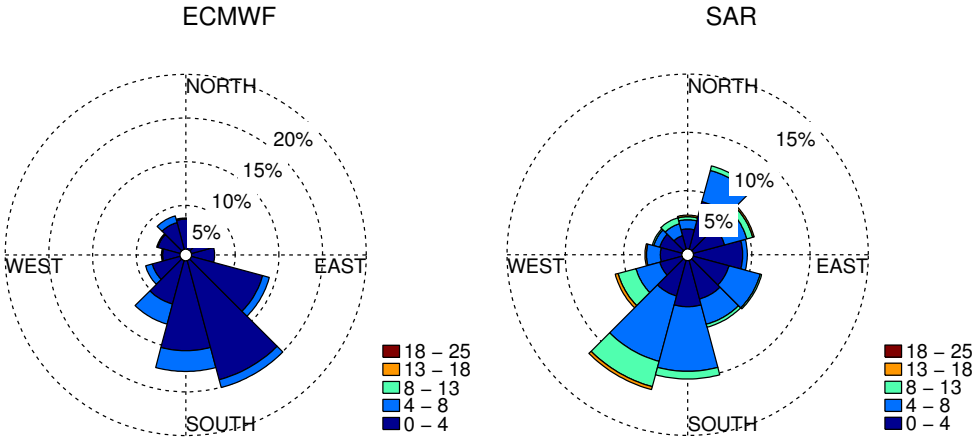
En fonction des conditions géographiques, les vents forts sont plus fréquents dans certaines directions. Lorsqu'une chaîne de montagne ou simplement des collines se trouvent à peu de distance de la côte, le vent est canalisé suivant deux directions privilégiées (et souvent opposées [Mayençon (1982)]), ce qui peut être le cas pour les points 1, 3 et 4 où les observations montrent parfois des vents beaucoup plus forts que les modèles numériques. Ces points se situent au fond des *fjords* entourés par des montagnes de deux côtés, ce qui favorise le phénomène de canalisation du vent. Sur des autres points comme 5, 7 et 8, les observations ont tendance à être plus faibles que les prévisions numériques, car ils sont protégés par la côte, les îles et les presqu'îles. Plus ils



(a) *Point 1*

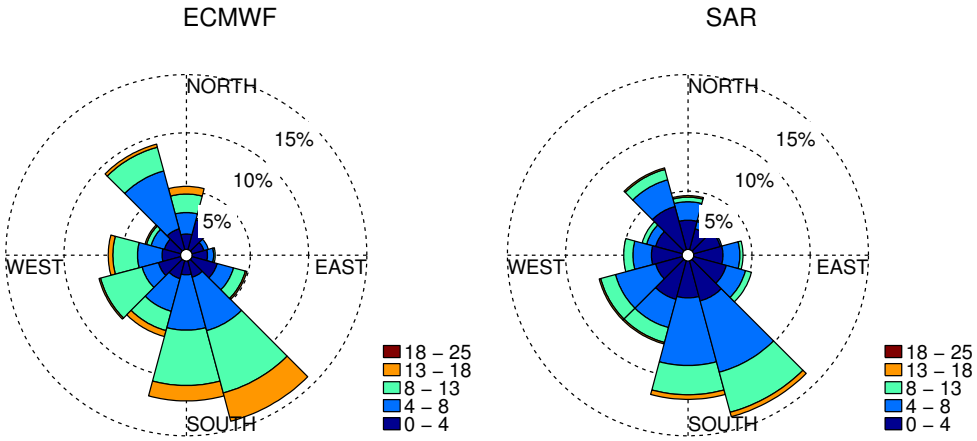


(b) *Point 3*

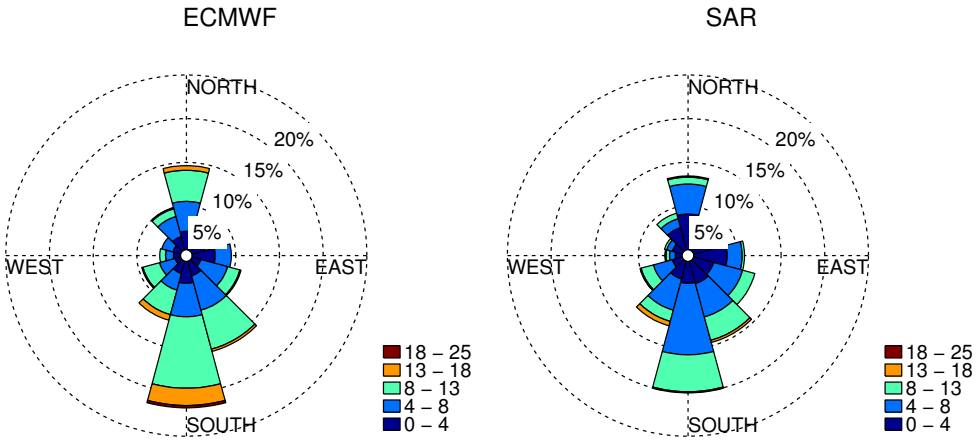


(c) *Point 4*

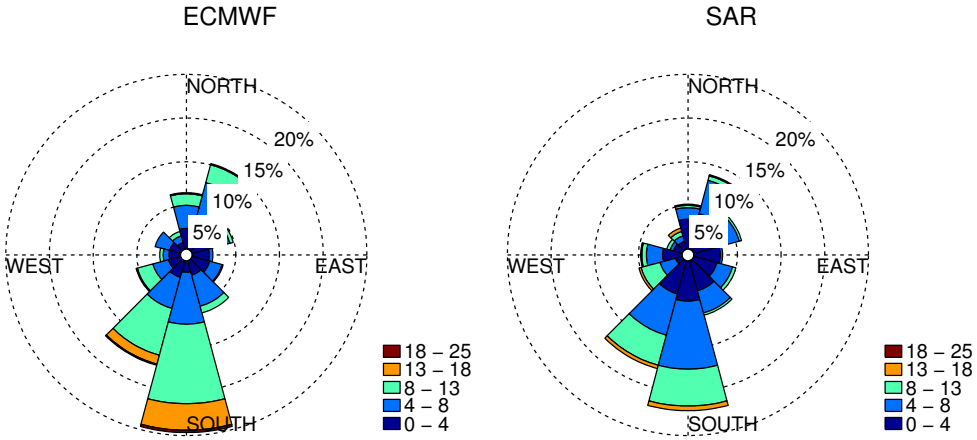
FIGURE 4.4 – *Zones fjords.*



(d) *Point 5*

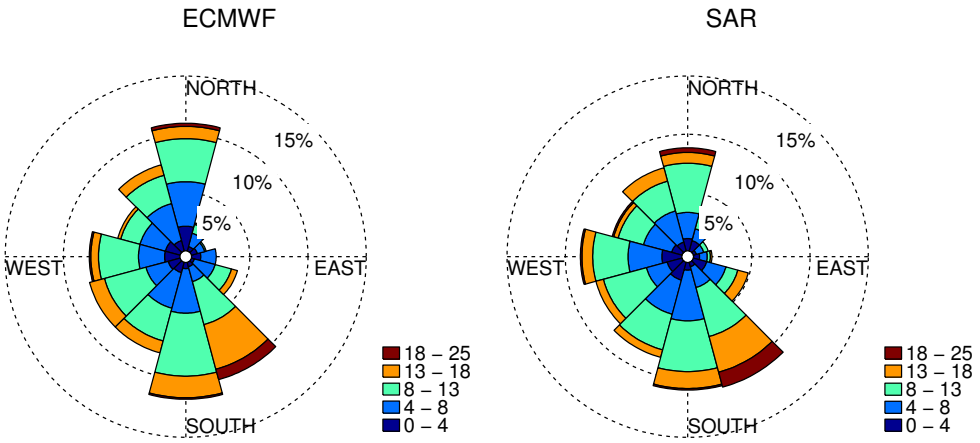


(e) *Point 7*

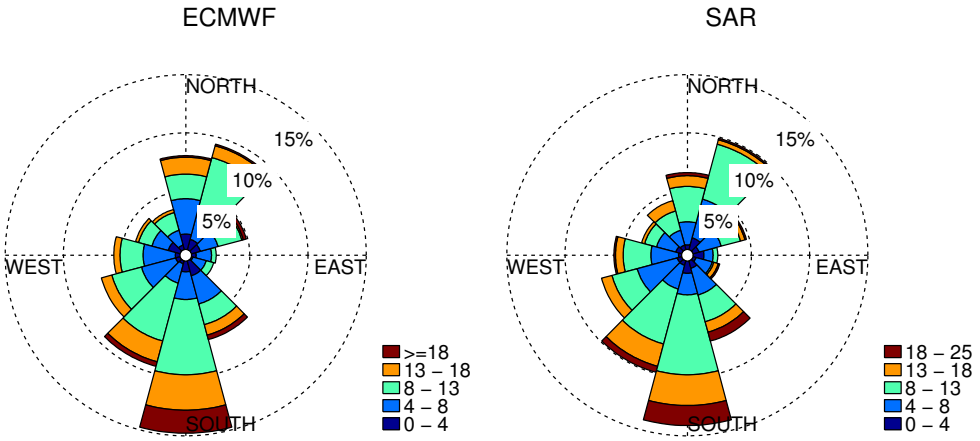


(f) *Point 8*

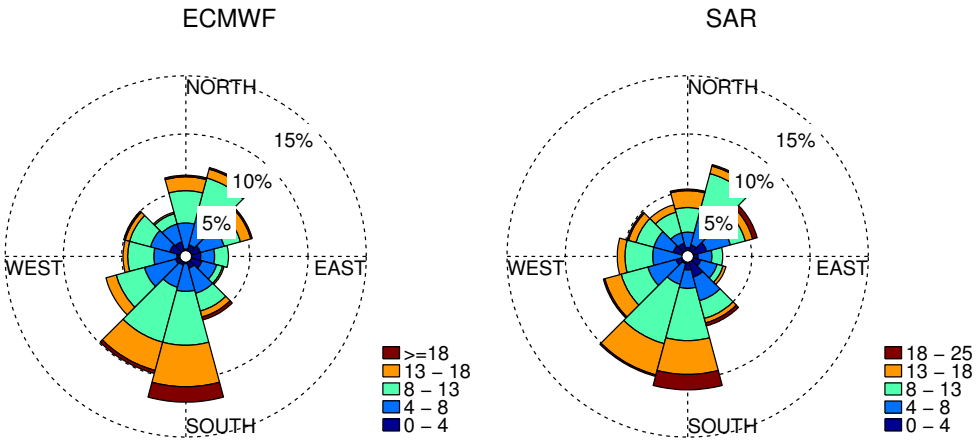
FIGURE 4.4 – *Zones côtières.*



(g) *Point 10*



(h) *Point 11*



(i) *Point 12*

FIGURE 4.4 – *Au large.*
Roses des vents de l'ECMWF (à gauche) et du SAR (à droite).

sont éloignés de la côte, moins les vents à HR sont impactés par la présence de continent, et plus les observations et les prévisions numériques restent proches.

En ce qui concerne les distributions des directions, les premiers 8 points sont très différents des points au large (9 – 12) et ils sont également très différents entre eux. Au large, les vents d'est restent rares et les vents de secteur sud et nord sont fréquents. Par contre, les observations montrent des vents d'est plus fréquents aux points 1, 2 et 3 qu'aux autres points. Toujours suivant les positions géographiques, les modifications des directions de vent diffèrent et prennent des amplifications différentes.

En conclusion, les roses des vents permettent de comparer les distributions des directions en même temps que les distributions des intensités aux différents points. Elles montrent que les différences entre les données SAR et les données ECMWF sont plus grandes dans les zones côtières qu'au large. Ces différences restent très locales et cela confirme la conclusion que chaque point doit être traité indépendamment pour les analyses et pour les émulations.

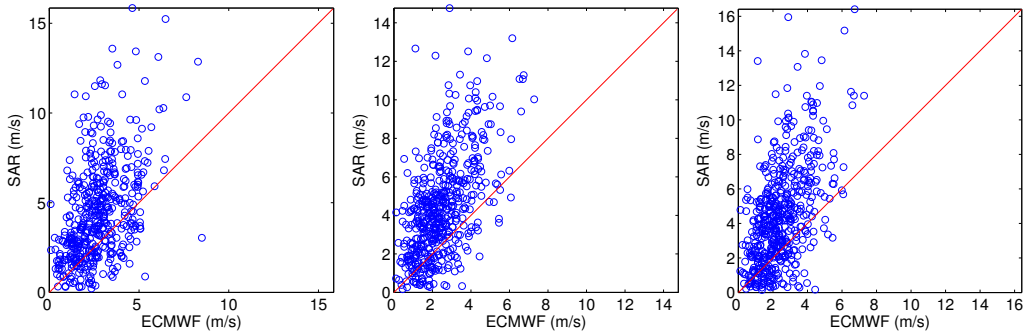
4.2.2 Diagrammes de dispersion

À la différence des roses de vent, un diagramme de dispersion du vent permet d'illustrer directement la dépendance entre les données ECMWF et SAR. Pour chaque point d'analyse, tous les couples de données ECMWF et SAR sont affichés sur une même figure. La figure 4.5 montre les diagrammes de dispersion du vent en intensité pour les trois zones différentes. 9 points parmi les 12 sélectionnés sont utilisés, et la ligne continue rouge représente l'identité. On observe une dépendance plus linéaire entre les données ECMWF et les données SAR aux points *offshore* (10, 11, 12) qu'aux points côtiers et *fjords*. On observe également que le vent SAR est systématiquement plus fort que le vent ECMWF aux points 1, 3, 4 et est systématiquement moins fort aux points 5, 7, 8.

Une analyse de la dispersion du vent ECMWF et SAR suivant la direction permet d'étudier la relation entre deux résolutions par secteur de vent. Pour chaque point, les vents sont classés en 8 secteurs selon la direction du vent ECMWF : les directions sont réparties sur 360° au compas avec un pas de 45°. Une fois les vents classés, on peut étudier la dispersion du vent ECMWF et SAR suivant la composante u, la composante v et l'intensité a.

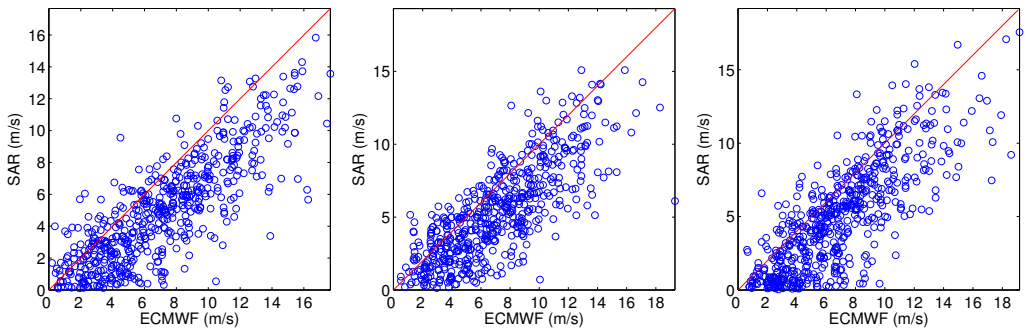
L'analyse de la dispersion du vent par direction est effectuée sur les 12 points. La figure 4.6 trace la dispersion du vent au point 1 à gauche, celle au point 8 au milieu et celle au point 12 à droite. L'axe des abscisses (resp. ordonnées) donne les valeurs des paramètres u, v et a en $\mathrm{m\,s^{-1}}$ pour les données ECMWF (resp. SAR). L'ensemble des observations montre que la dispersion dans les zones côtières et *fjord* est plus importante que celle *offshore*. Globalement, pour chaque direction, les vents ECMWF et SAR au large (point 12) restent très corrélés ; la dépendance du vent ECMWF et SAR dans la zone côtière est néanmoins non-linéaire.

Pour le point 1 dans le *fjord*, les vents ECMWF restent toujours très faibles quelque soit le niveau de vent SAR. Pour un vent du sud, de sud-est, et d'est (cf. Figure 4.6s), SAR observe des vents variant de $0\,\mathrm{m\,s^{-1}}$ jusqu'à $15\,\mathrm{m\,s^{-1}}$ en intensité lorsque les composantes u et v de l'ECMWF sont proches de zéro. Dans ce cas, on ne peut pas y établir de relation locale entre les vents à basse et à haute résolution.



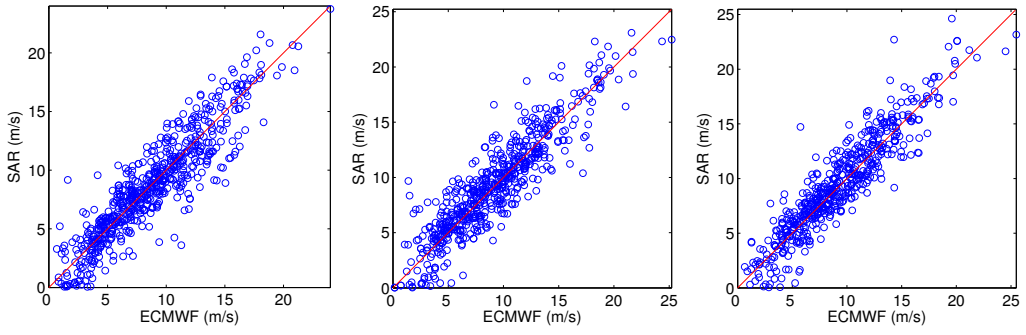
(a) *point 1* (b) *point 3* (c) *point 4*

FIGURE 4.5 – *Zones fjords.*



(d) *point 5* (e) *point 7* (f) *point 8*

FIGURE 4.5 – *Zones côtières.*



(g) *point 10* (h) *point 11* (i) *point 12*

FIGURE 4.5 – *Au large.*

Diagramme de dispersion de vent en intensité (en $\mathrm{m\,s^{-1}}$) pour les 12 points sélectionnés. L'axe des abscisses représente les vitesses des vents ECMWF et l'axe des ordonnées représente les vitesses de vent SAR.

On peut également remarquer pour le point 8 que la dispersion du vent ECMWF et SAR est plus proche de l'identité pour les secteurs des vents d'ouest, de sud-ouest et de nord-ouest. Pour les autres directions, SAR observe plutôt des vents plus faibles que ECMWF. On peut donc déduire que le point 8 est à l'abri pour tous les vents de terre.

Cette partie de l'étude nous apprend que la dispersion du vent entre ECMWF et SAR est plus grande dans la zone côtière et *fjord* qu'au large. L'analyse de la dispersion de vent par direction montre plus clairement certains effets locaux liés à la position géographique et à la direction du vent. Dans la zone côtière, la relation entre ECMWF et SAR reste toujours non-linéaire après la classification par secteurs.

La partie suivante effectue une analyse du comportement du vent du *fjord* vers le large pour mieux comprendre ses variations dans les différentes zones.

4.3 Comportement du vent du *fjord* vers le large

Cette partie analyse le comportement du vent en fonction de la distance du *fjord* vers l'océan. Le vent est plus faible en moyenne sur terre que sur mer, à cause des inégalités du frottement exercé sur le vent par le sol ; sur la côte, on observe normalement une vitesse intermédiaire [Mayençon (1982)].

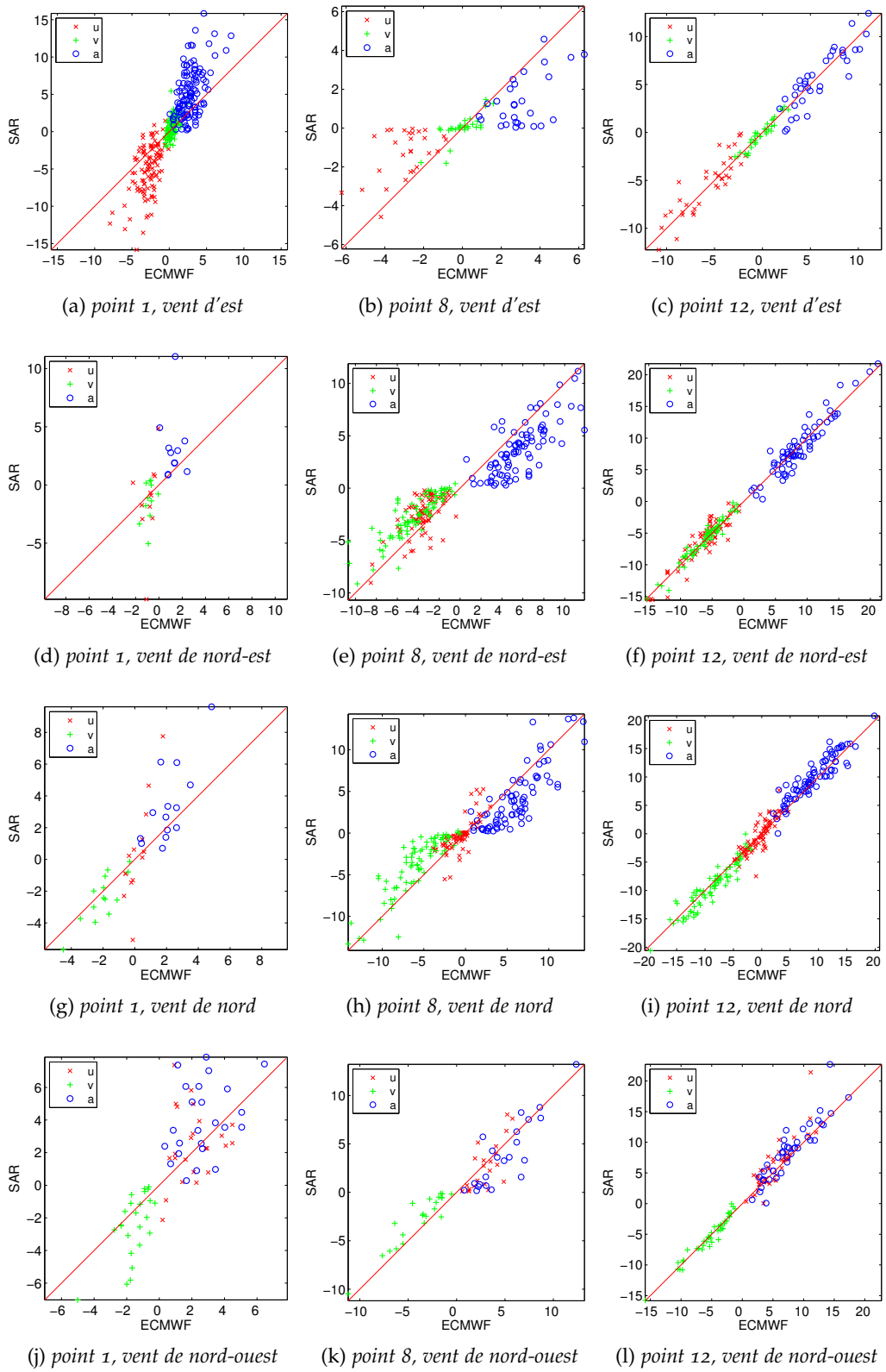
Le comportement du vent global est montré par la moyenne omnidirectionnelle de vitesse du vent à un point donné. Les points d'analyse se situent dans la direction perpendiculaire à la côte (cf. Figure 4.7a). Les points sont présents tous les 50 km pour ECMWF et tous les kilomètres pour SAR.

Sur la figure 4.7b, l'abscisse est la distance horizontale à partir du point de départ (N61.09°, E6.50°) dans le *fjord* tout à l'est de la ligne rouge. L'ordonnée correspond à la moyenne du vent en intensité ($m\,s^{-1}$). Le résultat montre que la moyenne diminue à partir du large vers la côte et qu'elle commence à décroître très rapidement à la distance de 250 km du point de départ dans le *fjord* qui correspond à une distance de 100 km à la côte. En même temps, la différence moyenne entre les données ECMWF et SAR augmente du large vers le *fjord*. Les vents ECMWF sont plus forts dans la zone côtière ensuite ils deviennent moins forts que SAR dans le *fjord*. Globalement, les vents moyens ECMWF et SAR sont proches dans la zone *offshore*.

Cette analyse montre que le vent moyen décroît à l'approche de la côte, contrairement à la différence entre ECMWF et SAR qui augmente à son approche. Dans les parties suivantes, les analyses du comportement du vent du *fjord* vers le large sont faites par direction et par intensité du vent.

4.3.1 Comportement du vent par direction

Pour estimer le rôle de la direction sur le comportement du vent en fonction de la distance, la moyenne des vents est tracée pour les mêmes points dans 4 directions — de nord, d'est, de sud et d'ouest. On observe la même tendance



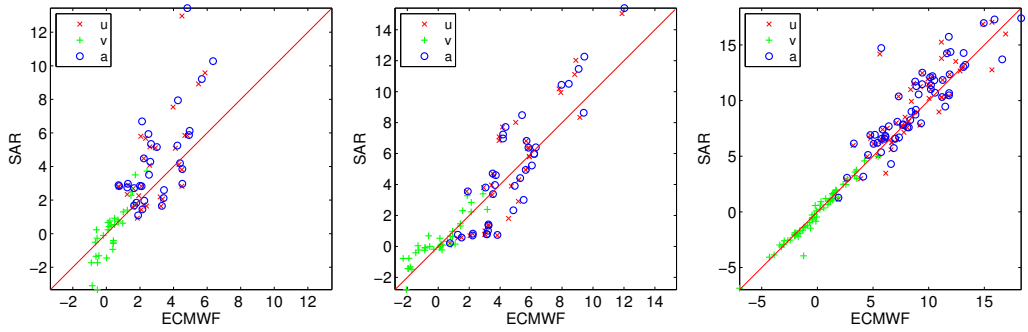
(a) *point 1, vent d'est* (b) *point 8, vent d'est* (c) *point 12, vent d'est*

(d) *point 1, vent de nord-est* (e) *point 8, vent de nord-est* (f) *point 12, vent de nord-est*

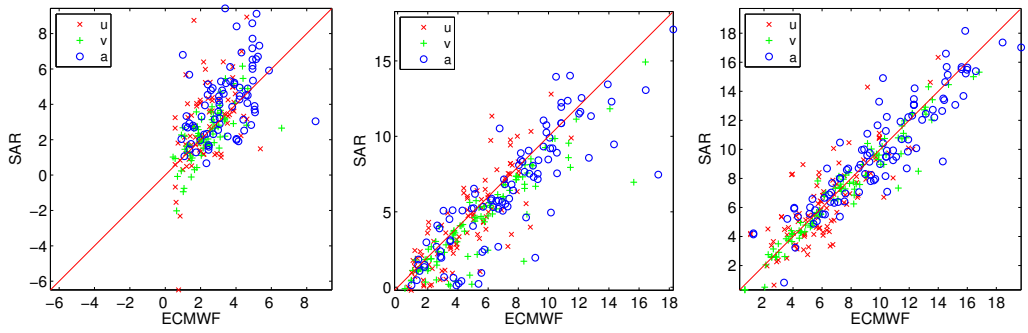
(g) *point 1, vent de nord* (h) *point 8, vent de nord* (i) *point 12, vent de nord*

(j) *point 1, vent de nord-ouest* (k) *point 8, vent de nord-ouest* (l) *point 12, vent de nord-ouest*

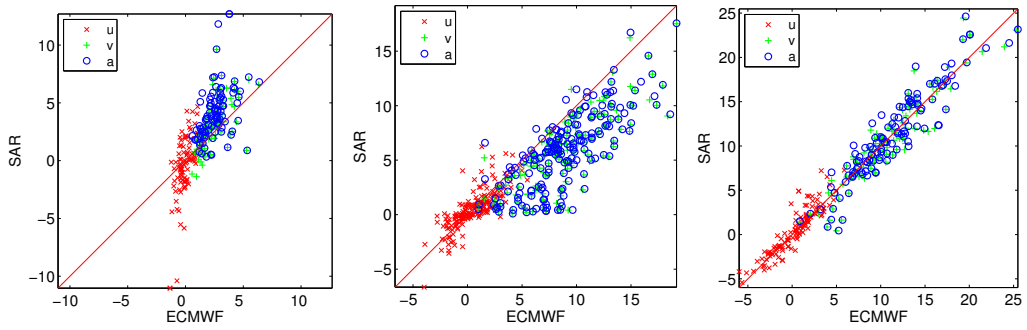
FIGURE 4.6 – *Partie I : comparaison du diagramme de dispersion entre les points 1, 8 et 12 par direction de vent. Les axes des abscisses et des ordonnées représentent les composantes u et v ainsi que l'intensité des vents (en* $\mathrm{m\,s^{-1}}$*) de l'ECMWF et du SAR respectivement.*



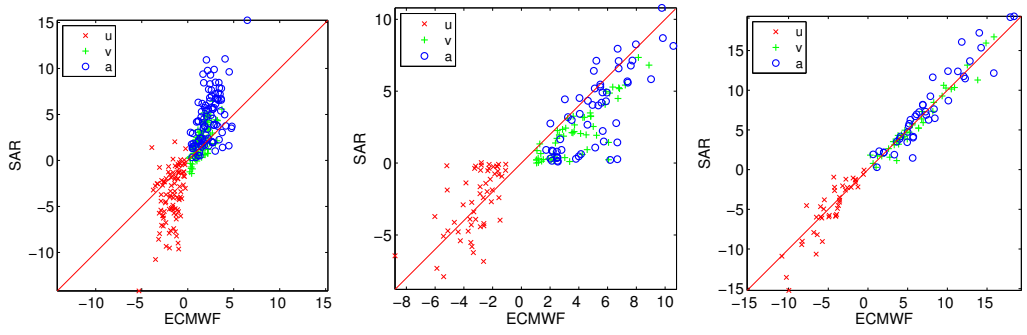
(m) *point 1, vent d'ouest* (n) *point8, vent d'ouest* (o) *point12, vent d'ouest*



(p) *point 1, vent de sud-ouest* (q) *point 8, vent de sud-ouest* (r) *point 12, vent de sud-ouest*

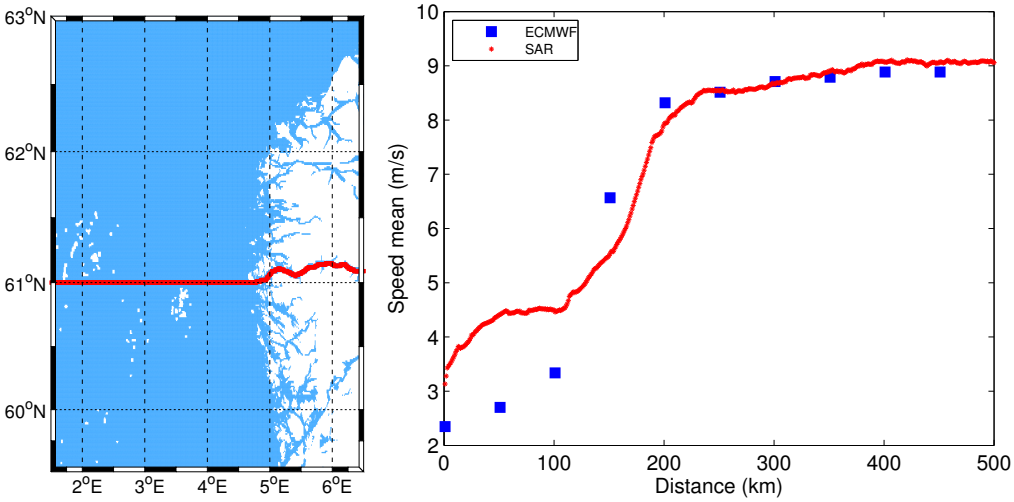


(s) *point 1, vent de sud* (t) *point 8, vent de sud* (u) *point 12, vent de sud*



(v) *point 1, vent de sud-est* (w) *point8, vent de sud-est* (x) *point12, vent de sud-est*

FIGURE 4.6 – *Partie II : comparaison du diagramme de dispersion entre les points 1, 8 et 12 par direction de vent. Les axes des abscisses et des ordonnées représentent les composantes u et v ainsi que l'intensité des vents (en* $\mathrm{m\,s^{-1}}$*) de l'ECMWF et du SAR respectivement.*



(a) *Plan perpendiculaire à la côte* (b) *Moyenne du vent ECMWF et SAR*

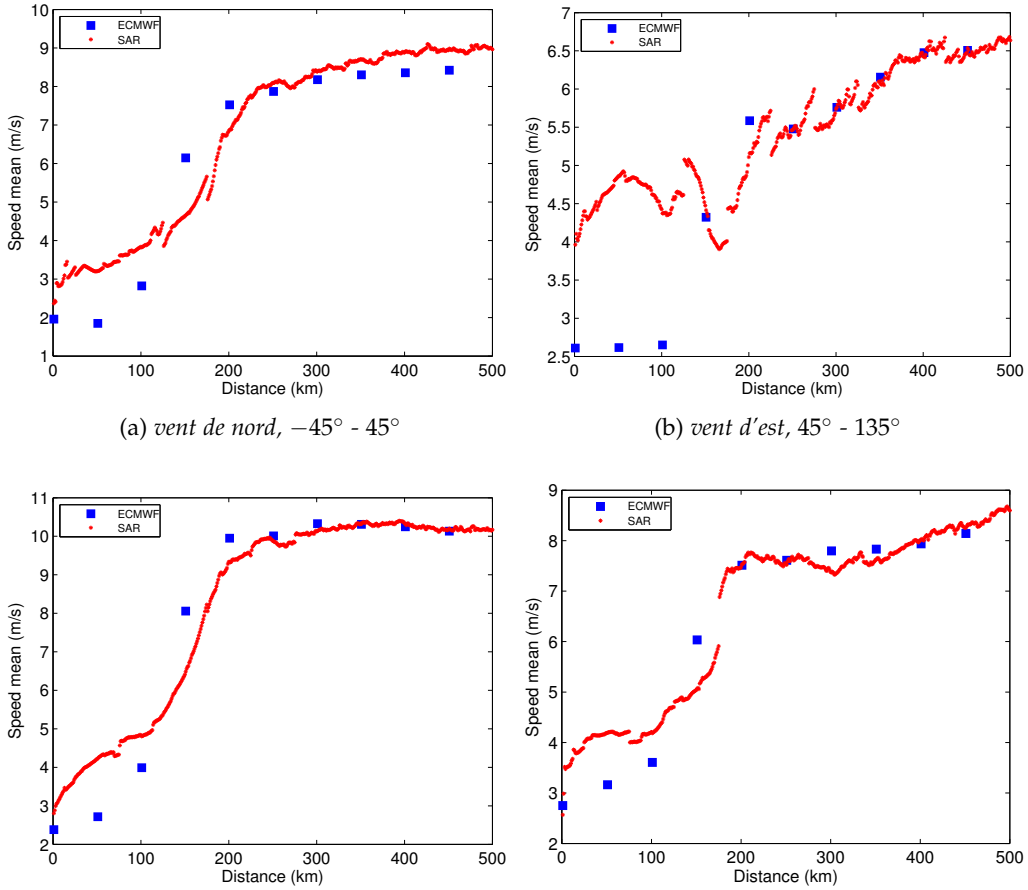
FIGURE 4.7 – *Zone du fjord vers le large ((a), ligne rouge) et comportement global de la moyenne du vent ECMWF et SAR en intensité en fonction de la distance à partir du point (N*61.09°, *E*6.50°*) dans le fjord (b).*

pour les 4 directions : la moyenne des vents diminue et la différence entre les vents ECMWF et SAR augmente à l'approchant de la côte.

Par contre, les données SAR montrent plus de variabilité pour un vent d'ouest et d'est par rapport à un vent de sud et d'est. Dans la direction d'ouest (cf. Figure 4.8d), la moyenne des vents diminue brutalement de $7\,\mathrm{m\,s^{-1}}$ à $6\,\mathrm{m\,s^{-1}}$. Les vents d'est (cf. Figure 4.8b) augmentent en plusieurs phases avec une fréquence d'environ 50 km. Chaque phase s'arrête à un maximum local, ce qui fait un effet d'*onde*.

L'effet d'onde observé dans la direction d'est est assez remarquable. Le vent d'est qui passe par dessus une chaîne de montagnes parallèle à la côte ouest de la Norvège, a créé une « onde de montagne ». Cela fait référence à une onde de gravité relativement stationnaire générée par l'action du vent sur le relief [Martin (2008)]. La figure 4.9 en donne un exemple concret. Le vent d'est observé par SAR le 24 janvier 2009 indique une fluctuation de vent. La fréquence de l'oscillation en différentes latitudes varie et celle entre latitude 59.50° et 60.50° est beaucoup plus grande.

La figure 4.10 compare la fluctuation d'un vent en moyenne (cf. Figure 4.8b) à un vent instantané pour les mêmes points que ceux indiqués sur la figure 4.7a. Le vent ECMWF du 24 janvier 2009 indique un vent un peu plus fort que la moyenne et les oscillations observées par SAR sont d'environ $8\,\mathrm{m\,s^{-1}}$. Dans cet exemple, la fréquence de l'oscillation du vent instantané est moins élevée que celle du vent en moyenne.



(a) *vent de nord*, −45° - 45°

(b) *vent d'est*, 45° - 135°

(c) *vent de sud*, 135° - 225°

(d) *vent d'ouest*, 225° - 315°

FIGURE 4.8 – *Comportement du vent ECMWF et SAR en fonction de la distance pour 4 directions, en latitude* 61°13′.

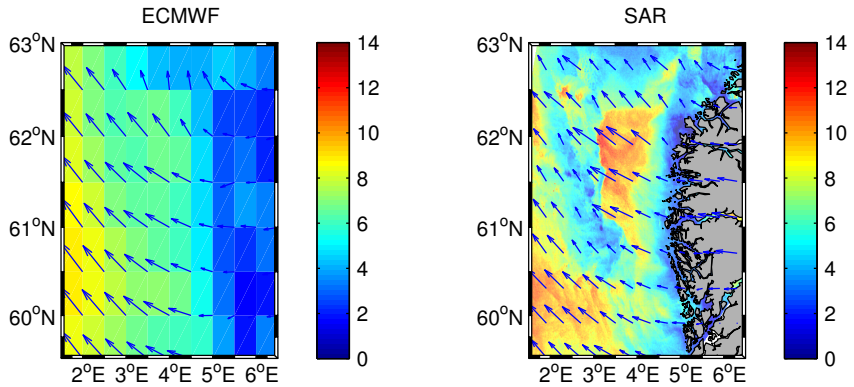


FIGURE 4.9 – *Donnée ECMWF et SAR du 24 janvier 2009, illustrant une « onde de montagne », principalement visible sur les données SAR.*

4.3.2 Comportement du vent par intensité

Le comportement du vent en fonction de la distance au *fjord* varie selon la direction. Dans cette partie, on analyse le comportement du vent par direction et suivant 3 différents niveaux d'intensité : la moyenne, la moyenne des vents plus forts que la moyenne et la moyenne des vents plus faibles que la moyenne.

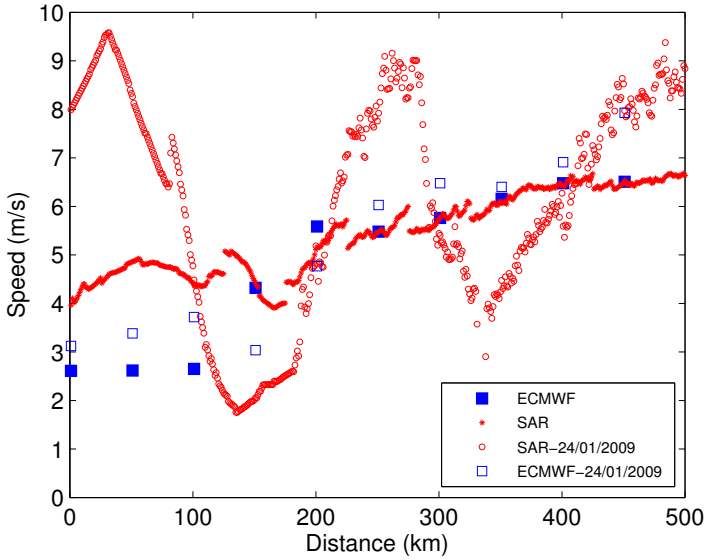


FIGURE 4.10 – *Comparaison de la fluctuation entre un vent en moyenne et un vent instantané le 24 janvier 2009.*

La figure 4.11 montre que la différence entre ECMWF et SAR dans les zones côtières est beaucoup plus grande pour les vents forts que pour les vents faibles. Le vent fort diminue plus vite à l'approche de la côte que le vent faible. Pour les vents d'est, la fluctuation des vents forts est beaucoup plus importante que pour les vents faibles.

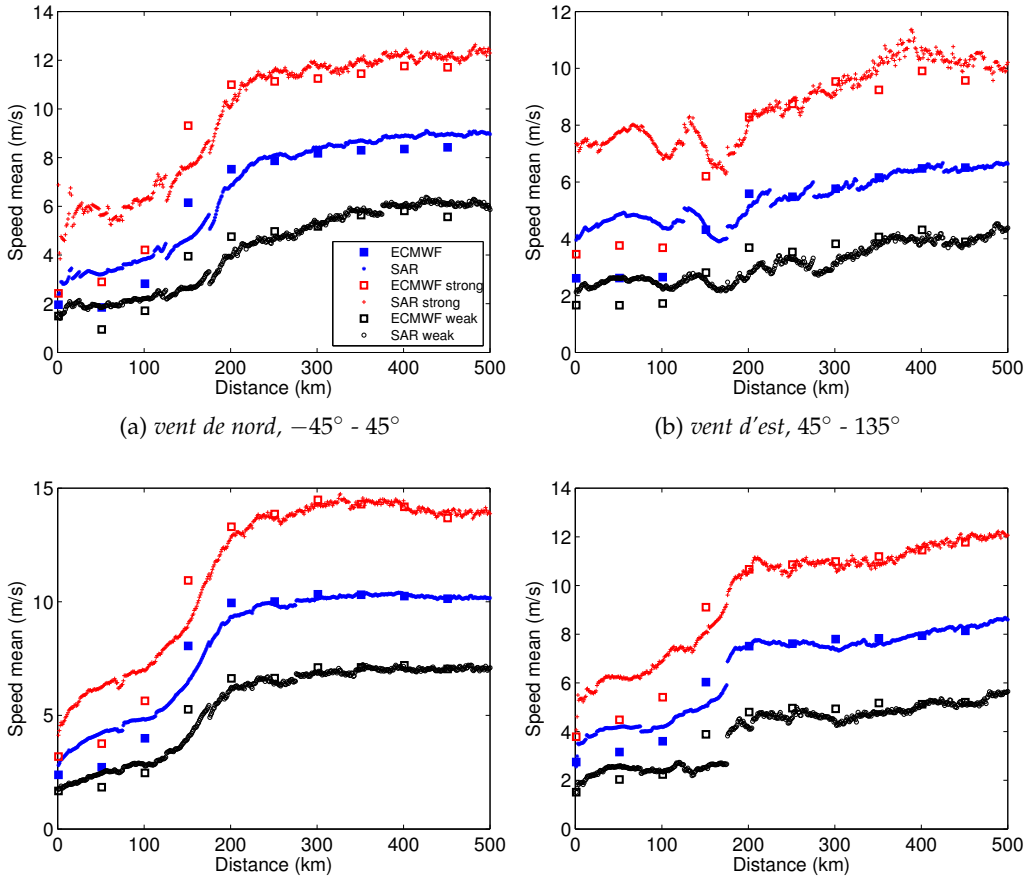
Ces résultats montrent que le comportement du vent en fonction de la distance au *fjord* varie également selon l'intensité du vent.

4.3.3 Synthèse

L'analyse du comportement du vent de la côte vers l'océan montre l'influence de la côte sur les vents à moyenne échelle et à micro-échelle. Le vent diminue rapidement à l'approche de la côte. Dans les zones côtières, la différence entre les données à BR et à HR est grande, tandis qu'elle reste assez faible *offshore*. L'analyse du comportement du vent par direction et par intensité permet de voir les différentes influences que la côte a sur le vent. Le fait que le vent souffle en moyenne plus fort sur la mer que sur la terre est valable pour toutes les directions et tous les niveaux de vent.

Néanmoins, les données SAR montrent mieux l'influence de la côte sur le vent. Le vent de mer diminue brutalement à partir d'environ 100 km de distance à la côte. Le vent qui vient du continent en passant par les côtes montagneuses crée des effets d'onde en surface de mer. La comparaison entre la fluctuation des vents en moyenne et d'un vent instantané pour la même latitude confirme l'existence d'ondes de montagne dans la zone d'étude, et que plus le vent est fort, plus l'oscillation en intensité est forte.

Tous ces phénomènes sont également observés sur les autres sections perpendiculaires au continent.



(a) *vent de nord*, −45° - 45°

(b) *vent d'est*, 45° - 135°

(c) *vent de sud*, 135° - 225°

(d) *vent d'ouest*, 225° - 315°

FIGURE 4.11 – *Comportement du vent par niveaux de densité. La moyenne des vents forts SAR est en croix rouges ; la moyenne des vents faibles SAR est en ronds noirs et la moyenne de tous les vents est en étoiles bleues. Les carrés représentent les moyennes pour l'ECMWF correspondant.*

4.4 Conclusion

Les analyses globales, locales ou par plan montrent toutes que la relation entre ECMWF et SAR est différente pour les différentes zones et aux différents points. Les différents effets locaux sont mieux vus par les données SAR, ce qui signifie que la résolution de SAR est plus adaptée pour analyser les dynamiques aux petites échelles.

Les analyses globales montrent que la basse et la haute résolution sont globalement plus différentes dans les zones côtières et dans les *fjords*. Des analyses plus locales comme les roses des vents et les dispersions de vent pour certains points sélectionnés permettent d'avoir en détail les différences de distributions statistiques. Les analyses de comportement du vent en fonction de la distance à partir du *fjord* vers l'océan montrent que le vent diminue rapidement et que la différence entre les basses et les hautes résolutions augmente du large à l'approche de la côte.

En résumé, pour l'émulation à HR, il est important de prendre en compte les conclusions suivantes :

1. Les caractéristiques de la distribution de la différence entre les basses et les hautes résolutions sont très locales, ce qui peut signifier que chaque position géographique fasse l'objet d'une étude indépendante et qu'il faille envisager un modèle d'émulation par point spécifique ;
2. La relation entre les hautes résolutions et les basses résolutions est assez linéaire au large et elle l'est beaucoup moins en s'approchant de la côte, ce qui veut dire qu'un type de méthode d'apprentissage basée sur une régression non-linéaire près de la côte est nécessaire alors qu'une régression linéaire peut être suffisante au large ;
3. Les analyses du comportement du vent du *fjord* vers le large par direction et par intensité montrent que la direction et l'intensité du vent ont une influence sur la variabilité à HR par rapport à la BR. Les approches comme l'injection de variabilité constante par direction de vent ne peut donc pas fonctionner [Ben Ticha (2007)].

5 Modèles d'émulation haute résolution

Cette partie présente de manière détaillée les méthodes d'émulation des champs de vent à HR. En général, les méthodes d'émulation considérées comportent deux phases :

1. Phase d'apprentissage : utiliser les données d'apprentissage pour obtenir la fonction de transfert BR/HR optimale et ses paramètres ;
2. Phase de prédiction : pour une nouvelle donnée BR, l'application de la fonction de transfert apprise fournit le champ HR émulé.

Pour la phase d'apprentissage, deux éléments clés doivent être distingués : 1) la définition des variables explicatives ; 2) la définition et la calibration du modèle de régression.

La définition des variables explicatives peut interagir avec la définition du modèle de régression ou bien se faire indépendamment de celle-ci. En parcourant l'espace des combinaisons possibles des variables d'entrée, la meilleure combinaison de variables est celle qui donne les erreurs de test du modèle minimales [Kohavi et John (1997)]. L'inconvénient majeur de ce type d'approche est que la recherche exhaustive dans l'espace des combinaisons est irréalisable pour un grand nombre de variables d'entrée (> 40) [Furnival et Wilson (1974), Guyon et al. (2002)]. Dans cette étude, nous privilégions une approche basée sur des critères statistiques qui permet d'éviter de tester toutes les combinaisons (§ 5.2).

Au niveau des méthodes d'apprentissage de la fonction de transfert, trois types de méthode de régression sont proposées : méthodes « analogues », méthode de Régression Linéaire Multiple (MLR) et méthode de régression non-linéaire avec Machine à Vecteurs de support pour la Régression (SVR) (§ 5.3). Comme évoqué au chapitre 2, ces différents modèles peuvent être formulés dans un cadre unifié à l'aide de fonction de noyau. Ils diffèrent par la paramétrisation retenue et la méthode de calibration associée. La méthode de régression non-linéaire avec SVR est en fait une méthode analogue générique et optimale. Pour une méthode analogue classique, les poids de chaque réponse et les paramètres de la fonction de distance sont tous fixés par l'utilisateur, tandis que les paramètres de la méthode SVR sont appris en résolvant un problème

d'optimisation. La méthode de régression non-linéaire avec SVR diffère de la méthode de MLR principalement par le type de noyau utilisé.

Même si un modèle général peut être formulé à partir des deux éléments-clés, c'est avant tout la méthode d'apprentissage de la fonction de transfert qui diffère. À chaque formulation correspond une méthode de calibration/apprentissage associée.

5.1 Algorithme général

La figure 5.1 illustre l'algorithme général pour construire un modèle d'émulation d'un champ HR à partir de données BR. Le modèle d'émulation HR revient à spécifier en chaque point (p,q) de la grille à HR une fonction de prédiction du champ HR au point (p,q). Nous l'appelons modèle par « point spécifique ». Les deux composantes d'un champ de vent sont émulées indépendamment. Prenons un exemple pour l'apprentissage de la composante u au point (p,q). Toutes les valeurs de la composante zonale du SAR $\{u_i\}$ disponibles dans le catalogue en ce point sont utilisées pour construire le vecteur des variables expliquées. L'étape de sélection des variables extrait la matrice des variables explicatives $\{x_i\}$ à partir des champs de vent à BR dans le catalogue. La phase d'apprentissage consiste à apprendre ces fonctions de transfert $f(p,q)$ en chaque point en utilisant un ensemble de paires de données d'apprentissage (x_i, y_i). Pour une nouvelle situation à BR et connaissant les modèles, sa situation à HR peut alors être prédite.

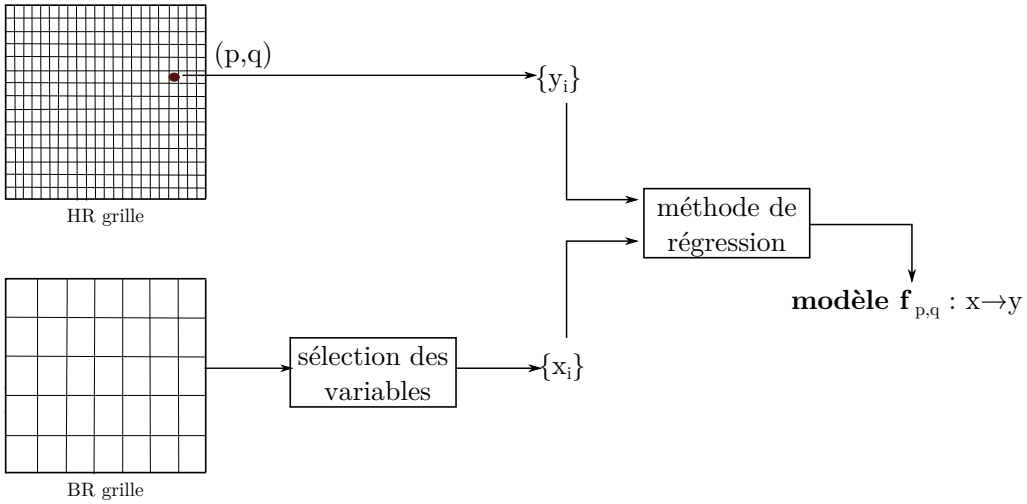


Figure 5.1 – *Principe de calibration d'un modèle d'émulation HR.*

5.2 Définition des variables explicatives

Pour qu'une méthode de régression fonctionne correctement, la recherche des variables explicatives pertinentes est fondamentale. Elle a principalement pour objectif de diminuer la dimension des variables d'entrée, de supprimer les

variables non dépendantes qui se comportent comme des bruits, d'avoir une meilleure compréhension du modèle et finalement d'améliorer la performance et la robustesse de l'apprentissage [Guyon et Elisseeff (2003)]. La définition des variables explicatives peut être associée à trois objectifs [Besse et Laurent (2012)] :

Descriptif Il vise à rechercher de façon exploratoire les liens entre la variable à prédire et les variables potentiellement explicatives qui peuvent être nombreuses afin, par exemple, d'en sélectionner un sous-ensemble. À cette stratégie, à laquelle peuvent contribuer des Analyse en Composantes Principales (ACP), correspond des algorithmes de recherche moins performants mais économiques en temps de calcul si le nombre de variable est grand.

Explicatif Une connaissance *a priori* du domaine concerné par exemple dont des résultats théoriques peuvent vouloir être confirmés, infirmés ou précisés par l'estimation des paramètres.

Prédictif L'accent est mis sur la qualité de définition des variables explicatives, afin de minimiser un critère d'erreur, par exemple l'erreur quadratique moyenne.

5.2.1 Types de variables explicatives

Le nombre de variables et la combinaison de sous-ensembles des variables explicatives influencent tous les deux la capacité prédictive d'une méthode de régression. La figure 5.2 extraite de [Hastie et al. (2009)] illustre les erreurs de test des modèles pour prédire l'antigène spécifique de la prostate par des variables explicatives comme le poids, l'âge, le volume du cancer, *etc*. en fonction du nombre de variables et en fonction des différentes combinaisons pour chaque nombre de variables donné. Pour l'exemple choisi dans Hastie et al. (2009), l'utilisation de 8 variables donne le meilleur résultat.

Pour ne pas avoir à tester toutes les combinaisons possibles, une approche est de commencer par classer les variables explicatives en fonction de certains critères de dépendance, par exemple les coefficients de corrélation et l'information mutuelle [Guyon et Elisseeff (2003)]. Une fois les variables classées, on évalue le nombre de variables optimal qui correspond aux K premières variables. Alors que les critères de corrélation reposent sur une hypothèse de dépendance linéaire. L'information mutuelle permet de prendre en compte les couples de variables ayant une dépendance non-linéaire. L'utilisation de l'entropie conditionnelle, aussi appelée « information non-locale », est une alternative très proche (cf. § 5.2.4).

Trois types d'information (ou variables explicatives) sont envisagés : informations globales, informations locales et informations non-locales. Les informations globales utilisent les K_G premières composantes principales de l'-Analyse en Composantes Principales (ACP) qui fournissent une représentation

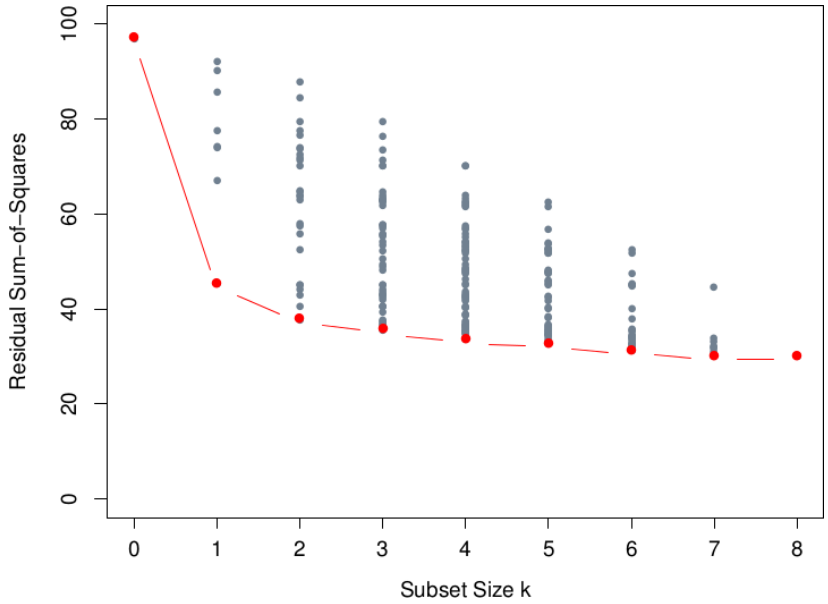


FIGURE 5.2 – *Influence du nombre de variables sur la performance du modèle appris [Hastie et al. (2009)]. L'axe des abscisses indique le nombre de variables utilisées (maximum 8) et l'axe des ordonnées indique les erreurs de test du modèle. Les points indiquent les différentes combinaisons des variables. Les points rouges indiquent la meilleur combinaison pour chaque taille de sous-ensemble.*

globale des variabilités spatiales du champ à BR ; les informations locales utilisent les informations des points les plus proches du point émulé et les informations non-locales sélectionnent K points parmi tous les points à BR sur la base d'un critère d'entropie conditionnelle. Dans les deux premières méthodes, il ne s'agit pas de sélection de variables. Ce sont des méthodes assez simples et intuitives pour exploiter les champs à BR.

5.2.2 Informations globales

L'analyse Analyse en Composantes Principales (ACP), aussi appelée Fonctions Orthogonales Empiriques (EOF) dans le domaine des sciences de l'environnement, permet de représenter un espace caractéristique initial par toutes les variables explicatives possibles pour représenter l'espace d'entrée dans un autre espace de dimension plus petite [Goubanova et al. (2010)]. Le principe est d'utiliser une transformation orthogonale pour convertir un ensemble de variables éventuellement corrélées en un ensemble de variables non corrélées. Les nouvelles variables sont appelées « composantes principales » [Bretherton et al. (1992), Björnsson et Venegas (1997), Jolliffe (2005)].

Pour n échantillons, chacun décrit par d variables explicatives, la méthode d'EOF consiste à identifier les modes orthogonaux de décomposition de la matrice de covariance empirique. Les M premiers modes sont choisis comme base de projection suivant deux critères : ils représentent un certain pourcentage important des variances expliquées ; ils s'ajustent principalement à la variabilité au large puisque les BR en zones côtières ne sont pas toujours représentatives.

La matrice d'entrée est projetée sur cette base pour obtenir les coefficients de projection (ou composantes principales). Ce sont ces coefficients qui sont ensuite utilisés comme variables explicatives pour l'apprentissage. Dans la phase de prédiction, la nouvelle situation à BR est d'abord projetée sur les M premiers modes retenus, puis le nouveau vecteur de coefficient est utilisé comme variable d'entrée du modèle d'émulation HR.

5.2.3 Informations locales

Pour un point (p, q) sur la grille à HR, le point correspondant sur la grille à BR est (p', q'). Les informations locales utilisent une fenêtre carrée centrée sur (p', q'). Le vecteur de variables explicatives est construit par les vecteurs vent sur les points à l'intérieur de la fenêtre. Le nombre de variables explicatives est donné par $(2 \times Wz + 1)^2$ où Wz est le nombre de pixels séparant le point (p', q') du bord de la fenêtre.

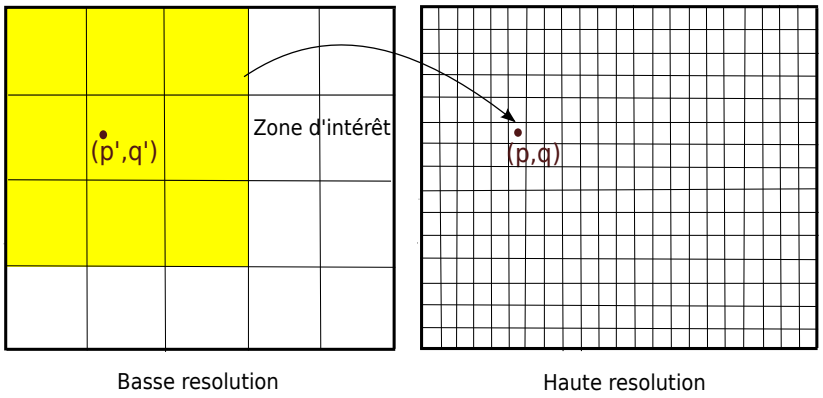


FIGURE 5.3 – *Informations locales. La grille de gauche (resp. de droite) représente les données à BR (resp. à HR). La zone en jaune indique les points auxquels les vecteurs vent sont utilisés comme variables explicatives.*

La taille de la fenêtre locale est un paramètre qui peut être optimisé dans la phase d'apprentissage. Il peut également être souligné que pour des tailles de fenêtre couvrant l'ensemble de la zone d'étude on se ramène aux informations globales considérées ci-dessus.

5.2.4 Informations non-locales

Même si l'hypothèse de relations significatives entre les informations locales à BR et l'information à HR semble pertinente, il peut être attendu dans certains cas, comme les zones d'abris par exemple, que des informations non-locales soient également pertinentes pour émuler l'information à HR. Nous introduisons donc une méthode de sélection de variables non-locales. Ce principe est illustré par la figure 5.4.

Nous introduisons une méthode analytique, l'entropie conditionnelle, pour obtenir les points les plus pertinents pour l'apprentissage. Le principe est

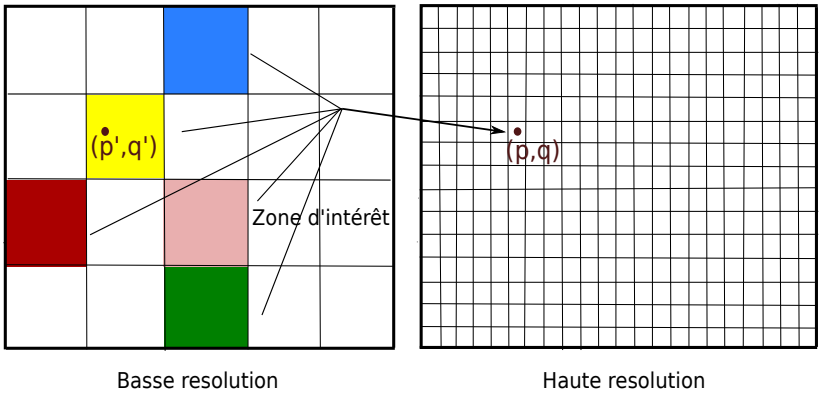


FIGURE 5.4 – *Informations non-locales. La grille de gauche (resp. de droite) représente les données à BR (resp. à HR). Les vecteurs vent aux points en différentes couleurs sont utilisés comme variables explicatives.*

d'utiliser les points sur la grille à BR qui ont les entropies conditionnelles les plus petites pour estimer les informations à HR à un point donné.

5.2.4.1 Définition et propriété d'entropie conditionnelle

En théorie de l'information, le concept d'entropie est introduit par Shannon (1948) pour mesurer la quantité d'information contenue ou délivrée par une source d'information. L'entropie conditionnelle quantifie la quantité d'information nécessaire pour décrire le résultat d'une variable aléatoire Y étant donnée la valeur d'une autre variable aléatoire X connue. Elle permet également de mesurer l'incertitude sur la valeur d'une variable sachant celle d'une autre. Plus l'entropie conditionnelle est élevée, plus l'incertitude est élevée.

Dans cette étude, pour un point (p,q) sur la grille à HR, on calcule l'entropie conditionnelle de sa situation à HR sachant les informations à BR au point (p',q') correspondant à un point quelconque sur la grille à BR. Les points les plus critiques sont ceux qui ont les entropies conditionnelles les moins élevées.

La partie suivante détaille quelques concepts sur l'entropie.

Définition 5.1 *Soit X une variable aléatoire à valeur dans $\{x_1, \cdots, x_n\}$, n étant le nombre d'événement. L'entropie de X est définie par :*

$$H(X) \triangleq -\sum_{i=1}^{n} P(X = x_i) \log P(X = x_i)$$

Propriété 5.1
- $0 < H(X) < log(n)$
- *L'entropie $H(X)$ est minimale, si $\exists i$ tel que $P(X = x_i) = 1$;*
- *L'entropie $H(X)$ est maximale, si chaque cas de X est équiprobable.*

Exemple 5.1 *Soit X une variable aléatoire binaire, avec $P(X = x_1) = p$ et $P(X = x_2) = 1 - p$, l'entropie de X est donnée par $H(p) = -p\log(p) - (1-p)\log(1-p)$. La figure 5.5 donne la distribution de $H(p)$ en fonction de p sur la base de $\log_2$. On remarque que l'entropie est maximale quand X est équiprobable et l'entropie est minimale quand X est complètement dans le cas de x_1 ou x_2.*

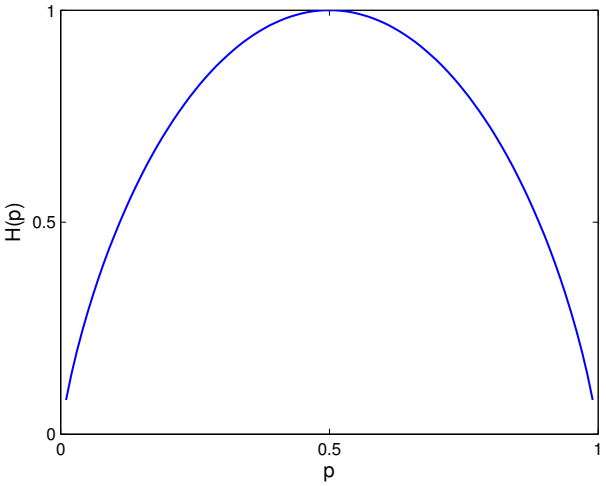


FIGURE 5.5 – *Entropie d'une variable aléatoire binaire.*

Définition 5.2 *Soit X et Y des variables aléatoires à valeur dans $\{x_1, \cdots, x_n\}$ et $\{y_1, \cdots, y_n\}$. L'entropie conditionnelle de Y sachant X est donnée par :*

$$H(Y|X = x_i) \triangleq -\sum_{j=1}^{m} P(Y = y_j|X = x_i) \log P(Y = y_j|X = x_i)$$

où $P(Y = y_j|X = x_i)$ est la probabilité conditionnelle d'obtenir y_j sachant x_i et $H(Y|X = x_i)$ est l'incertitude moyenne sur le résultat de Y sachant que celui de X est x_i.

Définition 5.3 *L'entropie conditionnelle moyenne de Y sachant X est donnée par :*

$$\begin{aligned} H(Y|X) &\triangleq \sum_{i=1}^{n} P(X = x_i)\, H(Y|X = x_i) \\ &\triangleq -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X = x_i) P(Y = y_j|X = x_i) \log P(Y = y_j|X = x_i) \end{aligned}$$

Définition 5.4 *L'information mutuelle entre Y et X est donnée par :*

$$I(X, Y) = H(Y) - H(Y|X)$$

Pour Y donné, l'entropie de Y, $H(Y)$, est fixée. Comparer l'information mutuelle entre Y et X revient à comparer leur entropie conditionnelle $H(Y|X)$. Plus l'information mutuelle est élevée, plus l'entropie conditionnelle est petite.

5.2.4.2 Entropie conditionnelle moyenne

Si Y correspond à l'intensité du vent à HR au point (p, q), X^I et X^D correspondent respectivement à l'intensité et la direction du vent à BR au point (p', q'). L'intensité du vent est représentée par n classes, et la direction du vent est composée de m classes. L'entropie conditionnelle moyenne est donnée par :

$$H(Y|X^I, X^D) = -\sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{k=1}^{m} p(Y = y_j, X^I = x_i, X^D = x_k) \cdot \log p(Y = y_j|X^I = x_i, X^D = x_k) \quad (5.1)$$

La difficulté de cette approche est le critère de division des classes de vent pour calculer les entropies conditionnelles. Ici, la définition des classes fournit seulement une approximation discrète de la loi continue par son évaluation empirique.

Tous les points sur la grille à BR sont classés en fonction de la croissance de leurs entropies conditionnelles. Si on évalue une taille de $2K$ sous-ensembles des variables explicatives, les vecteurs vent aux premiers K points classés sont utilisés.

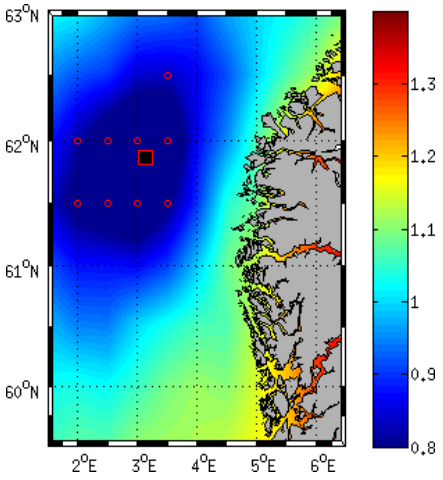
Pour illustrer l'estimation de l'entropie, on distingue trois types de points (p, q) : un point au large, un point côtier et un point *fjord*. Pour chaque type de point, les entropies sont calculées pour tous les points de la grille à BR. Les 3 différents niveaux de vent — vent faible, vent moyen et vent fort — correspondent à $0\,\mathrm{m\,s^{-1}}$ – $4\,\mathrm{m\,s^{-1}}$, $4\,\mathrm{m\,s^{-1}}$ – $9\,\mathrm{m\,s^{-1}}$, $> 9\,\mathrm{m\,s^{-1}}$ et les 4 classes {nord, east, sud, ouest} sont utilisées pour les directions des vents. Comme Y est composé de trois cas, l'entropie maximale vaut $\log_2(3)$ (≈ 1.59) en utilisant la base de $\log_2$, lorsque Y et X sont indépendants.

La figure 5.6 illustre les entropies aux 3 points différents. Le carré rouge face noir indique le point d'analyse sur la grille à HR et les cercles rouges indiquent les 9 points sur la grille à BR qui ont les entropies conditionnelles les moins élevées pour le point d'analyse. Par la suite, on appelle ces points « les points critiques ». Pour un point au large (cf. Figure 5.6a), les points critiques se situent autour du point (p', q') et un peu plus vers l'ouest. La situation pour un point *fjord* (cf. Figure 5.6c) est différente : les points critiques sont moins locaux, près de la côte. Pour un point côtier (cf. Figure 5.6b), les points critiques se situent à l'ouest du point d'analyse. L'entropie conditionnelle minimale pour les 3 points dans l'ordre au large, côtier, *fjord* sont de 0.80, 0.95 et 1.08. Ce résultat reste cohérent avec les analyses des données : l'incertitude pour prédire la HR est plus petite au large qu'en zones côtières et les points critiques pour un point au large sont plus locaux.

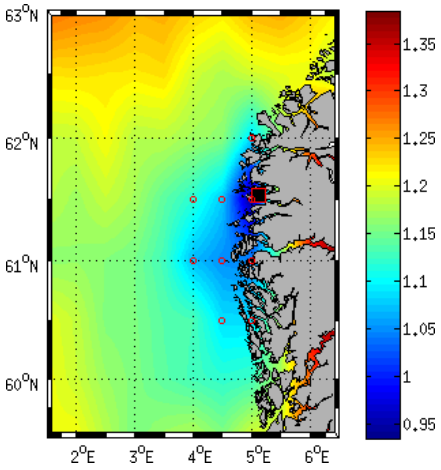
5.2.4.3 Influence de la direction du vent

L'analyse de l'entropie conditionnelle par direction du vent permet de mettre en évidence l'influence de la direction sur la distribution des points critiques. Dans l'équation 5.1, l'entropie conditionnelle moyenne peut être vue comme la moyenne des entropies en fonction des différentes direction des vents X_q^D :

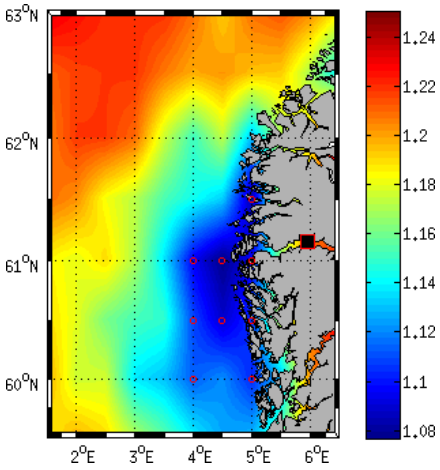
$$H(Y|X^I, X^D = x_k) = \sum_{k=1}^{m} p(X^D = x_k) \sum_{i=1}^{n} p(X^I = x_i|X^D = x_k) H(Y|X^I = x_i, X^D = x_k) \quad (5.2)$$



(a) *Point (N61.86°, E3.13°)*



(b) *Point (N61.53°, E5.11°)*



(c) *Point (N61.15°, E5.95°)*

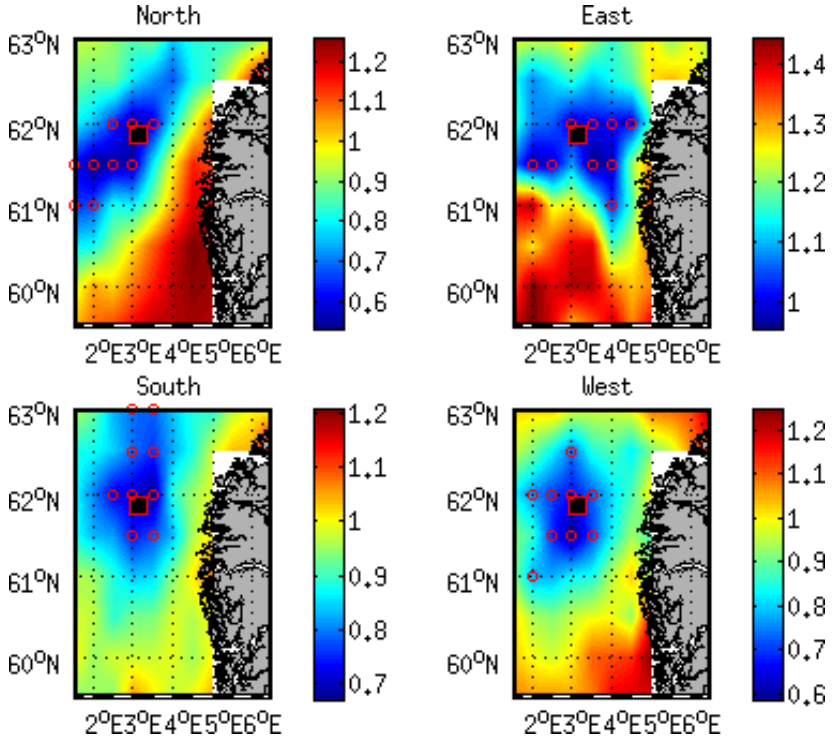
FIGURE 5.6 – *Entropie pour un point au large (a), un point côtier (b) et un point dans le fjord (c). Le carré rouge face noir indique le point (p, q) sur la grille à HR. Les cartes de l'entropie sont calculées pour chaque point à BR. Les illustrations sont les interpolations des entropies conditionnelles moyennes sur la grille à HR. Les cercles rouge indiquent les 9 premiers points ayant les entropies conditionnelles les plus petites.*

Pour une direction du vent donnée $X^D = x_k$, l'estimation des entropies conditionnelles est donnée par :

$$H(Y|X^I, X^D = x_k) = \sum_{i=1}^{n} p(X^I = x_i|X^D = x_k) H(Y|X^I = x_i, X^D = x_k) \qquad (5.3)$$

La figure 5.7 illustre les entropies conditionnelles pour les 3 mêmes points en fonction de la direction du vent. L'emplacement des points critiques changent en fonction de la direction du vent par rapport à celui de l'entropie moyenne. Pour le point au large et pour les vents de sud et de nord, les positions des points critiques ont tendance à se situer sous le vent.

Pour les points côtiers et *fjord*, l'emplacement des points critiques pour les vents d'est s'éloigne beaucoup du point d'analyse et se situe vers le large.



(a) *Point (N61.86°, E3.13°), au large.*

FIGURE 5.7 – *Partie I : entropie conditionnelle en fonction des 4 directions pour le point au large. Le carré rouge face noir indique le point (p,q) sur la grille à HR. Les cartes de l'entropie sont calculées pour chaque point à BR. Les illustrations sont les interpolations des entropies conditionnelles moyennes sur la grille à HR. Les cercles rouge indiquent les 9 premiers points ayant les entropies conditionnelles les plus petites.*

5.2.5 Synthèse

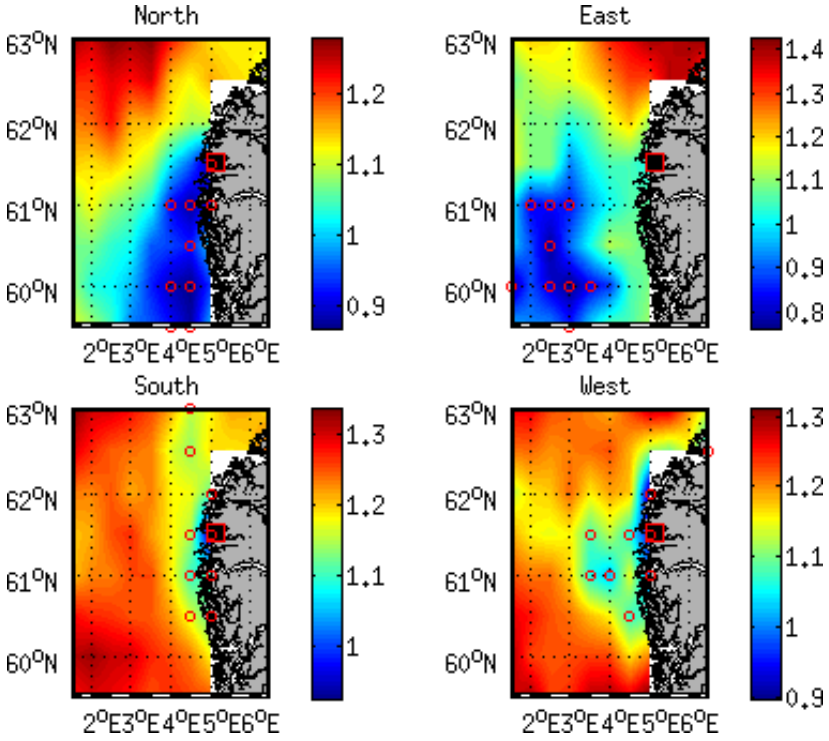
Cette partie propose une méthodologie pour utiliser les informations à BR pour obtenir les variables explicatives permettant d'émuler une information de vitesse à HR. Trois types de variables explicatives sont envisagés :

Informations globales utilisent les coefficients de projection sur les K_G premières EOFs pour représenter la situation globale de toute la zone d'étude ;

Informations locales utilisent tout simplement les informations autour du point émulé ;

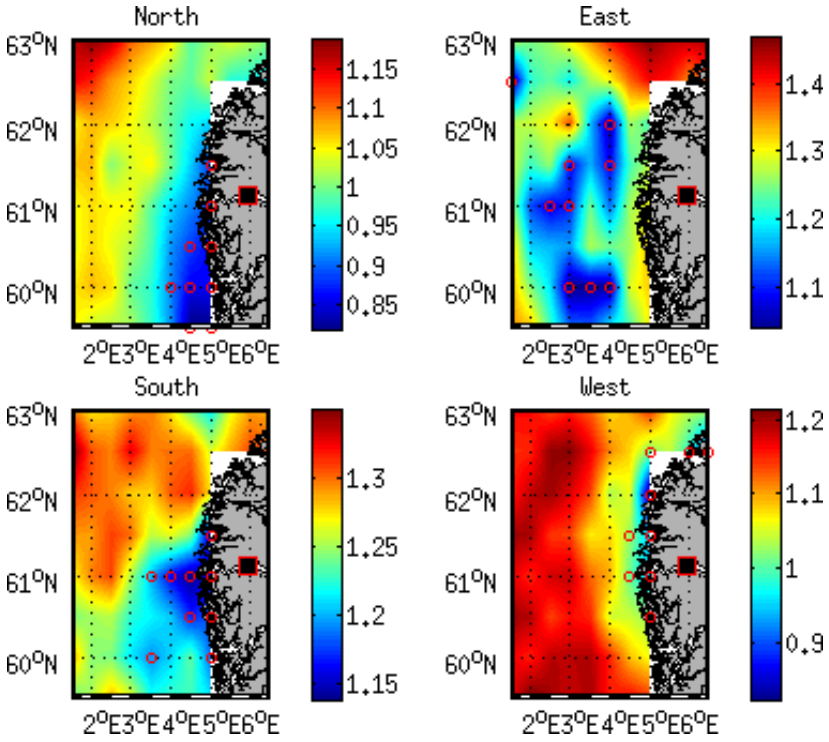
Informations non-locales reposent sur les analyses de dépendances en calculant les entropies conditionnelles pour sélectionner les points les plus pertinents pour l'émulation en un point donné.

Pour les informations non-locales, l'emplacement des points critiques est plus local pour un point d'analyse au large qu'en zones côtières et *fjords*. Les influences de la direction des vents sur l'emplacement des points critiques ont été identifiées : les points critiques ont tendance à se situer sous le vent, surtout pour un vent de sud et de nord.



(b) *Point (N61.53°, E5.11°), côtier.*

FIGURE 5.7 – *Partie II : entropie conditionnelle en fonction des 4 directions pour le point côtier.*



(c) *Point (N61.15°, E5.95°), fjord.*

FIGURE 5.7 – *Partie III : entropie conditionnelle en fonction des 4 directions pour le point dans le fjord.*

Pour les informations locales et non-locales, la taille optimale des sous-ensembles est à évaluer en associant le type d'information à différentes méthodes d'apprentissage. Ces évaluations sont faites dans le chapitre 6.

5.3 Méthodes de régression

Étant données les variables explicatives définies dans la section 5.2, trois méthodes de régression sont proposées pour obtenir la fonction de transfert : méthode analogue, méthode de Régression Linéaire Multiple (MLR) et méthode de régression non-linéaire avec Machine à Vecteurs de support pour la Régression (SVR). Ces méthodes sont implémentées et leur performances sont comparées en combinant les types de variables explicatives dans le chapitre 6.

Étant donnée une tâche spécifique à résoudre et un ensemble de fonctions $\mathcal{F}$, les méthodes de régression utilisent les données d'apprentissage pour trouver l'optimal $f^* \in \mathcal{F}$ en minimisant la différence entre la prédiction $f(X)$ et l'observation Y. En général, une fonction de coût $L : \mathcal{F} \to \mathbb{R}$ est définie, de sorte que pour la solution optimale f^*, $L(f^*) \leq L(f)$, $\forall f \in \mathcal{F}$. La fonction de coût est un concept très important pour les algorithmes d'apprentissage, car elle permet de mesurer l'écart d'une solution particulière par rapport à la solution optimale.

5.3.1 Méthode de régression linéaire

Étant donné le vecteur des variables explicatives $X = (X^1, X^2, \cdots, X^d)$ et la variable à prédire Y, un modèle de la régression linéaire a la forme suivante :

$$f(X) = \beta_0 + \sum_{j=1}^{d} X^j \beta_j \tag{5.4}$$

où $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_d)^T$ est le vecteur de coefficients de régression, T désigne la transposée, et β_0 est le terme du biais. Quand $d = 1$, il s'agit d'une régression linéaire simple. Pour toutes les dimensions supérieures, nous l'appelons Régression Linéaire Multiple (MLR). Pour pouvoir estimer les coefficients de régression, il faut que le nombre d'échantillons de X soit supérieur au nombre d de variables explicatives.

Une série des données d'apprentissage $(x_1, y_1), \cdots, (x_n, y_n)$ est utilisée pour l'estimation des coefficients de régression, où n est le nombre d'échantillons. Chaque x_i est une réalisation du vecteur des variables explicatives et y_i est une réalisation de la variable à expliquer. La technique la plus utilisée pour l'estimation des coefficients de régression est la méthode des moindres carrés (Ordinary Least Square (OLS)), où les coefficients β sont obtenus par minimisation de la fonction de coût L qui est définie par la somme des carrés du résidu :

$$L(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{5.5}$$

Notons $\mathbf{y}$ la colonne de toute la réalisation de la variable expliquée et $\mathbf{X}$ la matrice des réalisations de toutes les variables explicatives à la dimension $n \times (d+1)$. L'équation ci-dessus devient sous forme matricielle :

$$L = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \tag{5.6}$$

En dérivant la fonction de coût L suivant β, on obtient :

$$\frac{\partial L}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \tag{5.7}$$

$$\frac{\partial^2 L}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X} \tag{5.8}$$

En supposant $\mathbf{X}$ de rang dimension $d+1$, $\mathbf{X}^T\mathbf{X}$ est défini positif. La solution unique au problème s'obtient en évaluant la dérivée d'ordre 1 en zéro. Ainsi on obtient les estimations des coefficients de régression :

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{5.9}$$

Une fois les coefficients de régression estimés, pour une nouvelle observation de x, sa réponse y peut être prédite par l'équation suivante :

$$\hat{y} = \sum_{j=0}^{d} x^j \hat{\beta}_j \tag{5.10}$$

L'équation ci-dessus peut également être réécrite sous la forme d'une somme sur l'ensemble des échantillons de la base d'apprentissage :

$$\hat{y} = \sum_{i=1}^{n} \hat{c}_i \langle x, x_i \rangle \tag{5.11}$$

où $\langle x, x_i \rangle$ est le produit scalaire entre x et x_i, avec l'estimation des coefficients :

$$\hat{c} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}$$

On retrouve donc la forme d'une fonction de noyau correspondant au produit scalaire entre x et les échantillons de x_i dans la base d'apprentissage.

La méthode de MLR est basée sur plusieurs hypothèses. Elle suppose que les erreurs ont une espérance nulle et de covariance sphérique. Quand ces conditions sont satisfaites, les estimateurs de la MLR sont optimaux dans le sens où ils sont non biaisés, efficaces et cohérents.

La figure 5.8 donne un exemple du plan OLS avec deux variables explicatives. Le meilleur plan d'une régression linéaire est celui qui minimise la somme des carrées du résidu entre les échantillons de la variable expliquée Y et ses estimations par le modèle.

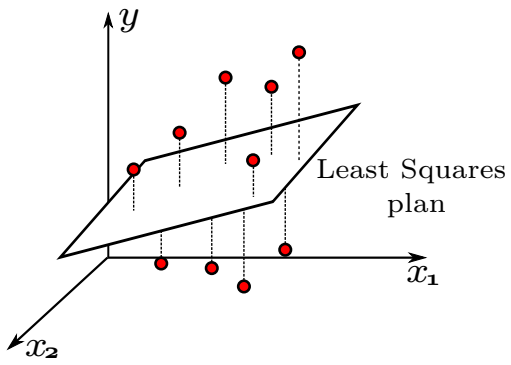


FIGURE 5.8 – *Exemple du plan OLS de la MLR avec deux variables explicatives* $(x_1\ x_2)$.

5.3.2 Méthodes analogues

La méthode analogue est définie comme une méthode qui ré-échantillonne les réponses dans le passé en fonction de la similarité entre la nouvelle entrée et les données historiques. La fonction de transfert pour une méthode analogue s'écrit comme suit :

$$f(x) = \sum_{i=1}^{n} w_i\, y_i\, g(x, x_i) \tag{5.12}$$

où w_i est un coefficient. $g(x, x_i)$ est la fonction de similarité entre la nouvelle entrée x et l'échantillon x_i, $x_i \in \mathcal{X}$, où $\mathcal{X}$ est l'ensemble des échantillons historiques dans la base d'apprentissage. Si $g(x, x_i)$ est non linéaire, la méthode analogue est une méthode de régression non linéaire. Typiquement, la fonction $g(x, x_i)$ est une mesure de similarité en fonction de la distance entre x et x_i. Voici quelques exemples de fonction de similarité :

- le plus proche voisin : $g(x, x_j) = 1, si\ \forall k \neq j,\ d(x, x_j) < d(x, x_k)$; sinon, $g(x, x_j) = 0$;
- une fonction exponentielle :

$$g(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{5.13}$$

- l'inverse de p-distance : $g(x, x_i) = \frac{1}{\sqrt[p]{\sum_j |x_j - x_{i,j}|^p}}$.

Dans le cas du plus proche voisin, la réponse y de la nouvelle entrée x vient directement de celle du plus proche voisin. C'est une approche où il n'y a pas besoin de paramètre, donc une méthode de modèle-libre. Les autres exemples assignent la valeur à prédire par pondération en fonction de la distance, de sorte que les plus proches voisins contribuent toujours plus que ceux plus éloignés.

Les coefficients $\{w_i\}$ peuvent être constants ou estimés par un critère d'optimisation. Les méthodes analogues consistent à fixer *a priori* les coefficients de régression. Pour une pondération optimale, Zorita et von Storch (1999) considèrent un critère des moindres carrés pour optimiser les coefficients $\{w_i\}$:

$$L = \sum_j \sum_i \frac{(y_{i,j} - f(x_{i,j}))^2}{\sigma_j^2} \tag{5.14}$$

où i est l'instant d'échantillonnage et j est le numéro de la station des observations. $y_{i,j}$ est la variable observée et $f(x_{i,j})$ est l'estimation de $y_{i,j}$. Finalement, à cause de la difficulté de résolution du problème d'optimisation associé, Zorita et von Storch (1999) choisissent une solution plus simple : le plus proche voisin. L'approche non retenue ressemble à la formulation de SVR et ne s'en différencie que par le critère d'optimisation considéré. Finalement, nous montrons en section 5.3.3 que dans le cas d'un problème non-linéaire SVR permet d'obtenir une pondération optimale w_i.

La méthode analogue est facile à implémenter techniquement et à interpréter physiquement. Pour les mêmes situations à BR, les mêmes types de comportement à HR sont reproduits. Si on a une grande quantité de données permettant de représenter toutes les situations possibles, la méthode analogue peut être un très bon choix pour l'émulation à HR. Mais en réalité, cela représente une grande contrainte et ce choix n'est pas toujours envisageable.

5.3.3 Machine à Vecteurs de support pour la Régression (SVR)

Les Machine à Vecteurs de Support (SVM) sont des modèles très populaires comme méthode d'apprentissage automatique pour la classification, la régression ou d'autres tâches d'apprentissages [Schölkopf et Smola (2001)]. Ils sont introduits par Vapnik et Chervonenkis [Vapnik et Chervonenkis (1971)]. Leur théorie est connue sous le nom de *VC théorie*. La méthode initialement développée pour résoudre les problèmes de classification est très adaptée aux problèmes de régression [Smola et Schölkopf (2004), Basak et al. (2007)].

Comparée aux autres types de méthode de régression, la méthode SVR a deux astuces particulières :

- un hyperplan est défini comme solution d'un problème d'optimisation sous contraintes dont la fonction objectif ne s'exprime qu'à l'aide de produits scalaires entre vecteurs et dans lequel le nombre de vecteurs supports contrôle la complexité du modèle ;
- le passage à la recherche de surface non-linéaire est obtenu par l'introduction d'une fonction noyau (*kernel*) induisant implicitement une transformation non linéaire des données vers un espace intermédiaire (*feature space*) de plus grande dimension. Le noyau spécifie le produit scalaire dans l'espace transformé.

Pour comprendre le principe de la méthode SVR, on considère d'abord le cas d'une régression linéaire. Pour être cohérent avec les notations courantes de la méthode SVR [Schölkopf et Smola (2001)], les notations utilisées pour les modèles linéaires dans la partie 5.3.1 ne sont pas reprises. Ici un hyperplan $f(x, \omega)$ d'une régression linéaire est défini par :

$$f(x) = \langle \omega, x \rangle + b \tag{5.15}$$

où $\omega \in \mathbb{R}^d$ et $b \in \mathbb{R}^1$ sont les paramètres qui définissent l'hyperplan (f), avec d

le nombre de variables explicatives. $\langle :, : \rangle$ définit le produit scalaire entre deux vecteurs.

Dans un cas linéaire, l'objectif est de trouver la fonction f la plus « plate » possible, ce qui revient à chercher ω le plus petit possible en norme. La façon la plus simple est de minimiser la norme euclidienne au carré $\|\omega\|^2$. En même temps, on n'accepte aucun biais d'estimation supérieur à ε. Le problème d'optimisation convexe s'écrit donc :

$$\text{minimiser} \quad \frac{1}{2}\|\omega\|^2 \tag{5.16}$$

sous les contraintes :

$$|y_i - \langle \omega, x_i \rangle - b| \leq \varepsilon \tag{5.17}$$

où ε est le biais maximum autorisé. Cette formulation introduit la notion de marge qui correspond à la tolérance acceptée sur l'erreur de prédiction. Ceci établit le lien mais également la différence entre les critères classiques de minimisation d'erreur quadratique et la méthode SVR. C'est cette formulation du problème sous les contraintes à « une marge près » qui permet de définir un hyperplan à l'aide des vecteurs de supports. Chaque donnée (x_i, y_i) est un vecteur support potentiel.

L'hypothèse précédente admet qu'il existe un hyperplan le plus « plat » possible qui approxime tous les couples de données $\{(x_i, y_i)\}$ avec la même précision ε. Or ce problème d'optimisation n'a pas toujours de solution. La variable de relâchement ξ_i spécifique à chaque donnée est introduite pour autoriser certaines erreurs. Ce concept est analogue à la marge douce (*Soft Margin*) de la fonction de perte [Bennett et Mangasarian (1992)] (cf. Figure 5.9) :

$$|\xi_i| = \begin{cases} 0 & \text{si } |y_i - f(x_i)| \leq \varepsilon \\ |y_i - f(x_i)| - \varepsilon & \text{sinon} \end{cases} \tag{5.18}$$

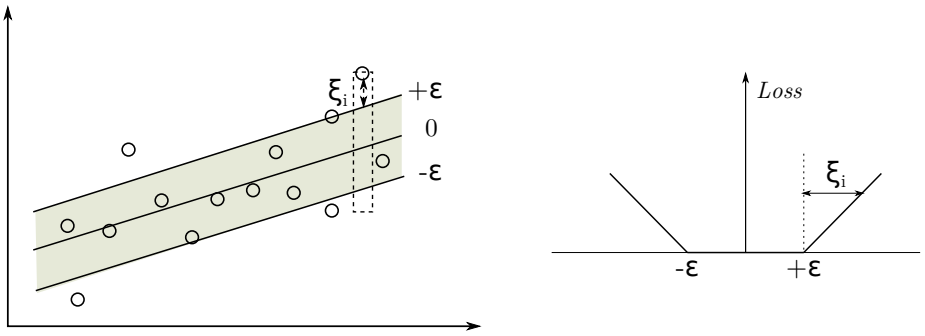


FIGURE 5.9 – *Marge douce pour un problème linéaire de la méthode SVR [Smola et Schölkopf (2004)]. ε est l'erreur permise pour toutes les données. ξ est la marge.*

Dans cette formulation, on ne pénalise pas les erreurs tant que celles-ci restent inférieures à un seuil ε. Précisément, l'hyperplan de la régression est donc construit par minimisation de :

$$\frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{5.19}$$

sous les contraintes :

$$\begin{cases} y_i - \langle \omega, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \varepsilon, \xi_i, \xi_i^* \geq 0 \end{cases} \qquad \forall i \in (1, n) \tag{5.20}$$

Le paramètre C est une constante, que l'on appelle « coefficient de régularisation » qui pondère l'influence du terme de régularisation par rapport à celle du terme d'erreur, dans l'expression à minimiser [Guermeur et Paugam-Moisy (1999)]. ξ_i et ξ_i^* sont les écarts des données en dessous et au dessus de ε.

En pratique, ce problème d'optimisation sous contrainte est résolu en minimisant une formulation lagrangienne [Kecman (2001)] :

$$\begin{aligned} L := {} & \tfrac{1}{2}\|\omega\|^2 + C \textstyle\sum_{i=1}^n (\xi_i + \xi_i^*) - \sum_{i=1}^n (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \textstyle\sum_{i=1}^n \alpha_i (\varepsilon + \xi_i - y_i + \langle \omega, x_i \rangle + b) - \sum_{i=1}^n \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle \omega, x_i \rangle - b) \end{aligned} \tag{5.21}$$

Où η, η^*, α, α^* sont les multiplicateurs de Lagrange, qui sont tous positifs. En annulant les dérivées partielles du lagrangien sur les paramètres primitifs $(\omega, b, \xi_i, \xi_i^*)$, on obtient :

$$\begin{aligned} \partial_b L &= \textstyle\sum_{i=1}^n (\alpha_i^* - \alpha_i) = 0 \\ \partial_w L &= \omega - \textstyle\sum_{i=1}^n (\alpha_i^* - \alpha_i) x_i = 0 \\ \partial_{\eta_i^{(*)}} L &= C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \end{aligned} \tag{5.22}$$

Le problème d'optimisation de l'équation 5.21 revient à maximiser :

$$\begin{aligned} & -\tfrac{1}{2} \textstyle\sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) x_i x_j - \varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) + \sum_i^n y_i (\alpha_i - \alpha_i^*) \\ & \text{sous les contraintes :} \\ & \qquad \textstyle\sum_i^n (\alpha_i - \alpha_i^*) = 0 \quad \text{et} \quad \alpha_i, \alpha_i^* \in [0, C] \end{aligned} \tag{5.23}$$

Après avoir résolu le problème de maximisation, on obtient les multiplicateurs de Lagrange α_i et α_i^* et le vecteur ω :

$$\omega = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) x_i \tag{5.24}$$

ce qui signifie que ω est complètement décrit par une combinaison des échantillons de x. La complexité du problème est donc indépendante de l'espace d'entrée $\mathbb{R}^d$ et elle dépend seulement du nombre de vecteurs supports actifs ($(\alpha_i^* - \alpha_i) \neq 0$).

En remplaçant ω dans l'équation (5.15) par l'équation (5.24), l'hyperplan optimal est tel que :

$$f(x, \omega) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b \tag{5.25}$$

L'hyperplan estimé est représenté par le produit scalaire entre x et les échantillons de x_i dans la base d'apprentissage. Grâce à l'autorisation du biais ε,

les erreurs des estimations de certains échantillons sont considérées à zéro, et par conséquent seule une partie des données a des coefficients $(\alpha_i - \alpha_i^*)$ non nuls. Cela signifie que la fonction de régression ne dépend que d'une sous-partie des données d'apprentissage, appelées vecteurs supports (Vectors de Support (SVs)). Ce sont les points les plus proches de l'hyperplan optimal [Guermeur et Paugam-Moisy (1999)].

Dans le cas d'un problème non-linéaire, la transformation par une fonction non-linéaire $\Phi(x)$ dans un autre espace où il existe une solution linéaire au problème est utilisée à la place de x. La régression non-linéaire est donc définie par :

$$f(x) = \omega^t \Phi(x) + b \tag{5.26}$$

À nouveau, l'hyperplan optimal est obtenu en minimisant l'équation (5.19) sous les contraintes (5.20), en remplaçant l'équation (5.15) par l'équation (5.26). On obtient alors :

$$\begin{cases} \omega = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \Phi(x_i) \\ f(x, \omega) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle \Phi(x_i), \Phi(x) \rangle + b \end{cases} \tag{5.27}$$

La fonction de noyau est équivalente à :

$$\mathrm{K}(x_i, x) = \langle \Phi(x_i), \Phi(x) \rangle \tag{5.28}$$

L'hyperplan optimal en fonction de noyau devient :

$$f(x, \omega) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \mathrm{K}(x_i, x) + b \tag{5.29}$$

L'estimation du biais b par la condition de Karush-Khun-Tucker est expliquée en détail dans le chapitre 9 de Schölkopf et Smola (2001).

Il existe différents types de fonctions de noyau. Pour SVR, il suffit qu'elle soit un noyau de « Mercer » qui vérifie $K(x, x_i)$ continu, symétrique et défini positif [Schölkopf et Smola (2001)]. Par exemple :

- noyau linéaire $K(x, x_i) = \langle x, x_i \rangle$;
- noyau polynomial $K(x, x_i) = (x \cdot x_i^T + 1)^d$;
- noyau gaussien

$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2) \tag{5.30}$$

- d'autres noyaux comme la tangente hyperbolique.

La fonction de noyau la plus adaptée au problème peut être déterminée en fonction des erreurs de test par validation croisée.

Au niveau de la complexité algorithmique, la méthode SVR est plus pénalisée par le nombre de vecteurs supports que par le nombre de variables explicatives, du fait que la fonction objectif ne s'exprime qu'à l'aide de vecteurs supports. Néanmoins, des versions performantes des algorithmes d'optimisation permettent de prendre en compte des bases de données volumineuses dans des temps de calcul acceptables [Besse et Laurent (2012)].

5.3.3.1 Interprétation

En reprenant l'expression 5.29 avec un noyau gaussien (*radial basis function*) $K(x, x_i) = \exp(-\gamma\|x - x_i\|^2)$, le plan optimal de 5.29 devient :

$$f(x, \omega) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\exp(-\gamma\|x - x_i\|^2) + b \tag{5.31}$$

$\|x - x_i\|^2$ peut être vu comme la distance euclidienne entre deux vecteurs. La valeur d'un noyau gaussien (Radial Basis Function) varie entre 0 et 1. La fonction de noyau est en effet une mesure de similarité. Plus le vecteurs support est proche de la nouvelle entrée x, plus il est pondéré par un poids important. Si on compare l'équation 5.31 avec l'équation 5.12 quand Φ est une fonction exponentielle, on s'aperçoit qu'elles ont la même forme. La méthode SVR est en effet une méthode analogue générique et optimale.

5.3.3.2 Choix de hyper-paramètres

La qualité d'un modèle SVR dépend du choix de ses paramètres comme le coefficient de régularisation C, la précision ε et des paramètres de la fonction de noyau choisie [Cherkassky et Ma (2002)]. Ces paramètres sont appelés hyper-paramètres. Ils doivent être choisis par l'utilisateur, en général par une recherche exhaustive dans l'espace des paramètres.

Un modèle SVR dépend de tous ces hyper-paramètres. L'apprentissage de la meilleure combinaison d'hyper-paramètres ne peut pas se faire indépendamment : il est possible de parcourir une grille exhaustive de toutes les combinaisons possibles, mais cela est peu réaliste au niveau du calcul. Cherkassky et Ma (2004) proposent une approche analytique pour fixer une approximation de C et ε directement à partir des données d'apprentissage.

Le paramètre C détermine le compromis entre la complexité du modèle (la régularité) et le degré de tolérance des erreurs (la marge d'erreur). Dans la démonstration pour arriver à la solution générale de la méthode SVR (cf. Équation (5.23)), Cherkassky et Ma (2004) montrent que :

$$\alpha_i, \alpha_i^* \in [0, C]$$

et que la solution générale est donnée par l'équation (5.29) :

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathrm{K}(x_i, x)$$

On s'aperçoit ainsi que le paramètre C a un lien avec la plage de valeurs de la sortie y. Une approximation de la solution est donnée par [Cherkassky et Ma (2002)] :

$$C = \max(|\overline{y} + 3\sigma|, |\overline{y} - 3\sigma|) \tag{5.32}$$

Le paramètre ε dépend du bruit d'entrée [Cherkassky et Mulier (2007)]. La solution suivante est détaillé dans Cherkassky et Ma (2004) :

$$\varepsilon = 3\sigma\sqrt{\ln n / n} \tag{5.33}$$

où σ est l'écart-type du bruit d'entrée et n est le nombre de données d'apprentissage.

Pour fixer les paramètres d'une fonction de noyau — prenons l'exemple d'un noyau gaussien — le choix du paramètre γ est également très important. $\gamma \to \infty$ ou $\gamma \to 0$ conduisent tous les deux à une mauvaise performance de généralisation [Wang et al. (2003)]. Quand $\gamma \to \infty$, toutes les données d'apprentissages sont considérées comme les vecteurs supports, ce qui provoque un problème de sur-apprentissage. Quand $\gamma \to 0$, toutes les données sont vues comme un seul point, ce qui fait qu'aucune nouvelle donnée ne sera reconnue.

5.3.3.3 Données d'apprentissage et données de test

Dans la phase d'apprentissage, les données peuvent être divisées en deux parties : une partie pour l'apprentissage du modèle, et une partie pour évaluer le modèle. Notons les données d'apprentissage $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ et les données de test $\{(x'_1, y'_1), (x'_2, y'_2), \cdots, (x'_m, y'_m)\}$, avec $\{x_1, x_2, \cdots, x_n\} \cap \{x'_1, x'_2, \cdots, x'_m\} = \emptyset$. L'erreur d'apprentissage, appelée aussi « risque empirique », est définie comme :

$$R_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{i=n} \|f(x_i) - y_i\| \tag{5.34}$$

et l'erreur de test, appelée « risque », est définie comme :

$$R(f) = \frac{1}{n} \sum_{i=1}^{i=m} \|f(x'_i) - y'_i\| \tag{5.35}$$

À partir de ces définitions, on peut voir que la minimisation de l'erreur d'apprentissage n'implique pas une erreur de test faible. Cette technique combinée à une validation croisée (cf. § 6.2.1) est utilisée pour choisir la meilleur combinaison des hyper-paramètres.

5.3.3.4 Exemple de régression non-linéaire avec SVR

Pour comprendre comment fonctionne la méthode SVR, on peut utiliser un exemple artificiel. On crée une série d'observations de (x, y) (cf. Figure 5.10), générée par la fonction $y = 0.8sin(2x) + 2 + br$, $br \in [-0.4, 0.4]$ étant un bruit suivant une loi uniforme. La relation entre x et y que l'on cherche à estimer est donc : $y = 0.8sin(2x) + 2$.

Avec la méthode SVR, on doit choisir une fonction de noyau et rechercher la meilleur combinaison des paramètre dans l'espace des hyper-paramètres. Ici un noyau gaussien (Équation 5.30) est utilisé, avec son paramètre γ. Une grille composée de 3 hyper-paramètres est définie par le paramètre de noyau $\gamma = \{0.0005, 0.001, 0.05, 0.1, 0.2, 0.5, 1, 2, 3, 4\}$, la pénalité $C = \{1, 2, 5, 10, 20, 50, 100, 500\}$ et la ε-sensibilité entre 0.1 et 1.5 avec un pas de 0.1. Pour fixer la meilleur combinaison des 3 hyper-paramètres, une validation croisée (§ 6.2.1) en 3 parties est utilisée.

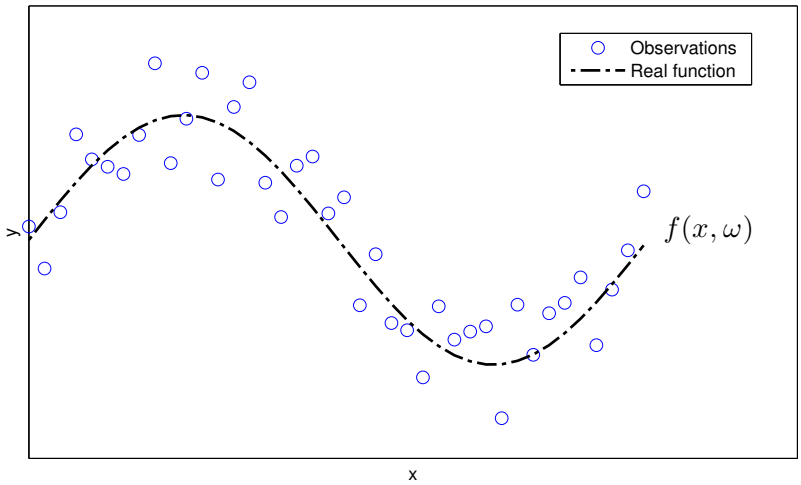


FIGURE 5.10 – *Une série d'observations $x - y$ (cercles) et la fonction de transfert à estimer (ligne pointillée).*

Finalement, La combinaison $(\gamma,\ C,\ \varepsilon)$ égale à $(0.1,\ 500,\ 0.3)$ donne le meilleur résultat en terme d'erreur quadratique moyenne de la validation croisée. Selon la valeur de ε, on peut tracer les courbes des marges $f(x,\omega) \pm \varepsilon$. Pour la combinaison $(0.1,\ 500,\ 0.3)$, les vecteurs de supports qui définissent la régression optimale sont illustrés par les points noirs cerclés de rouge sur la figure 5.11 et le modèle estimé (cf. Équation (5.31)) correspond à la courbe fixée par les points bleus en diamants sur la figure 5.12.

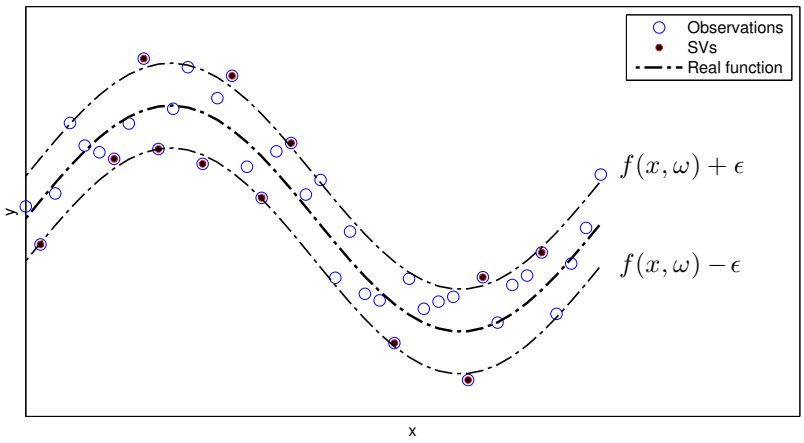


FIGURE 5.11 – *Vecteurs de support (en points noirs cerclés rouge) pour la meilleur combinaison de γ, C et ε.*

5.3.4 Régression multi-modèles

Dans la littérature, on utilise souvent des approches multi-modèles pour l'émulation de vent, reposant sur une première étape de classification de la situation à BR, en associant un modèle de régression à chaque classe à HR [Walmsley et al. (2001), Ben Ticha (2007), Minvielle (2009)]. Les méthodes de classification font partie de l'ensemble des méthodes statistiques descriptives

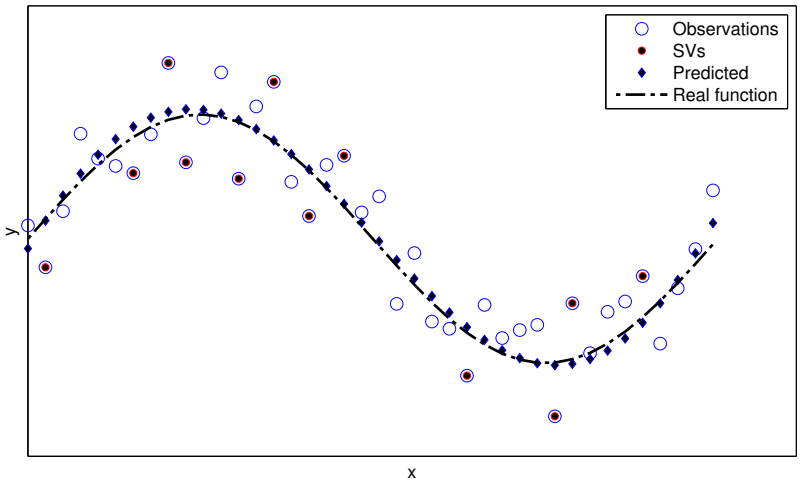


FIGURE 5.12 – *Hyperplan estimé (points bleus en diamants) pour la meilleur combinaison de γ, C et ε et ses vecteurs de support (en points noirs cerclés rouge).*

multidimensionnelles dont le but est d'expliciter la structure d'un ensemble de données. Les algorithmes de classification répartissent les individus en groupes homogènes selon un ensemble de variables disponibles. Elles aident à extraire les informations cachées, et elles peuvent simplifier la complexité d'un modèle d'apprentissage.

Concernant l'identification des classes, Minvielle (2009) utilise une classification non-supervisée, *k-means*, pour regrouper les situations similaires dans la même classe. Walmsley et al. (2001) montre que la classification par intensité du vent apporte très peu d'amélioration sur la performance prédictive par rapport aux modèles de régression linéaire seuls. Nous cherchons d'autres types de critère de classification. En 4.3 et en 5.2.4, on a montré que les comportements de vent diffèrent en fonction de la direction du vent et que la direction du vent impacte également l'emplacement des points critiques pour les informations non-locales. On propose donc une approche de classification basée sur l'analyse de la direction de vent de la situation à BR.

La figure 5.13 illustre le schéma d'une régression multi-modèles basée sur une classification au préalable. Pour chaque point émulé (p, q), l'ensemble des données d'apprentissage est réparti dans G classes en fonction de la direction du vent au point (p', q') de la grille à BR. Chaque classe contient donc un jeu de données qui permet d'apprendre un modèle de régression en utilisant les méthodes de régression proposées précédemment (cf. § 5.3) pour chaque classe. Un modèle par point spécifique devient maintenant un ensemble de sous-modèles $\{f^k\}$ par point spécifique.

Pour éviter le problème marginal, nous considérons la division en classes avec recouvrement. Donnons un exemple : pour une division en 4 classes, au lieu d'utiliser $[-45°, 45°]$, $[45°, 135°]$, $[135°, 225°]$ et $[225°, 315°]$, la division de classes avec une marge de $\pm 45°$ est la suivante : $[-90°, 90°]$, $[0°, 180°]$, $[90°, 270°]$ et $[180°, 360°]$.

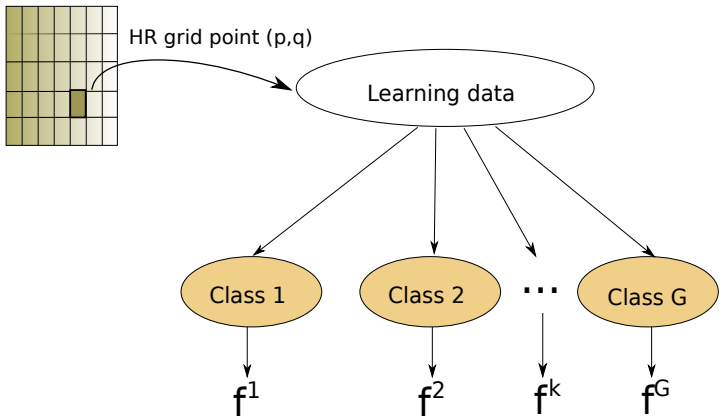


FIGURE 5.13 – *Schéma pour un modèle par point spécifique basé sur la classification. G est le nombre de classes et f^k est le modèle appris pour la k^e classe.*

Une fois l'ensemble de sous-modèles $\{f^k\}$ appris pour chaque point dans la zone émulée, la prédiction à HR y_{n+1} pour une nouvelle situation à BR x_{n+1} donnée se fait de la façon suivante : pour un point (p,q), l'estimation de la réponse $y_{n+1}(p,q)$ est une somme pondérée des $\{y^i(p,q)\}$ prédits par chaque sous-modèle en fonction de la distance entre la nouvelle direction du vent au point (p',q') et celle aux centroïdes de chaque classe, qui s'écrit :

$$y_{n+1}(p,q) = \sum_{i=1}^{G} w(d(\theta_{n+1}(p',q'), \theta_o^i(p,q)) f^k(x_{n+1}) \tag{5.36}$$

où $\theta_{n+1}(p',q')$ est la nouvelle direction du vent au point (p',q'), $\theta_o^i(p,q)$ est le centroïde de la i^e classe de vent au point (p,q), d est la distance au carré entre la nouvelle direction et le centroïde et w est le poids de pondération en fonction de la distance d au carré.

5.4 Conclusion

Ce chapitre décrit la méthodologie utilisée pour proposer un modèle d'émulation de champ HR. De manière générale, les modèles considérés sont formulés comme des modèles de régression du champ HR à partir de variables explicatives issues du champ BR. Deux éléments clés sont distingués :

Définition des variables explicatives il existe trois façons de définir les variables explicatives à partir du champ à BR : les informations globales utilisent les M premières composantes principales pour représenter la situation globale de toute la zone d'étude ; les informations locales utilisent les variables aux points locaux à l'intérieur d'une fenêtre carrée centrée sur le point émulé ; les informations non-locales utilisent les champs à BR aux K premiers points parmi tous les points sur la grille à BR qui ont les entropies conditionnelles les moins élevées sur la situation à HR au point émulé.

Définition du modèle de régression Compte tenu des approches de régression les plus classiques et les plus utilisées, les méthodes analogues et

la méthode MLR sont introduites puis comparées avec une approche plus récente, plus complexe et plus robuste : la méthode de régression non-linéaire avec SVR. Les trois approches peuvent être formulées de la même façon à l'aide d'une fonction de noyau. La méthode de régression non-linéaire avec SVR diffère de la méthode MLR par le type de noyau utilisé et par le choix des données de référence, et elle diffère des méthodes analogues classiques par la calibration de ces coefficients de régression. Elle est préférée aux deux autres types de régression en raison de sa propriété d'optimisation du choix des données de référence et des coefficients de régression, et de sa meilleure capacité de généralisation.

Le chapitre 6 met en place et évalue les modèles constitués par un type d'information et une méthode de régression. Le ou les meilleurs modèles sont proposés à la fin des évaluations. La méthode de régression multi-modèles est également comparée avec la méthode de régression seule.

Évaluation expérimentale

6

Ce chapitre présente l'évaluation expérimentale des modèles présenté au chapitre 5 sur la base de données réelles décrites au chapitre 3. Chaque modèle d'émulation se compose de deux éléments : un ensemble de variables explicatives et une méthode de régression. Quatre méthodes de régression sont comparées : la méthode du plus proche voisin (NN), méthode analogue par somme pondérée (AN), Régression Linéaire Multiple (MLR) et Machine à Vecteurs de support pour la Régression (SVR). Chaque méthode de régression est évaluée en l'associant avec une des trois types de variables : informations globales, informations locales et informations non-locales. La mise en œuvre est précisée pour chacune de ces méthodes.

Une procédure de validation croisée est utilisée pour calibrer les modèles d'émulation et en évaluer les performances. L'évaluation des performances sur une série de points caractéristiques sur la grille à HR permet de comparer qualitativement et quantitativement les différents modèles pour différentes situations caractéristiques (zones *offshores*, zones côtières et zones *fjords*). Finalement, le meilleur modèle est utilisé pour évaluer l'émulation sur tout le catalogue pour toute la zone d'étude. La plupart des évaluations peut être comparée avec les analyses spatiales conjointes définies au chapitre 4, ce qui permet de montrer l'amélioration de l'émulation à HR par rapport aux données ECMWF à BR.

6.1 Mise en œuvre expérimentale

Cette partie consiste à présenter la mise en œuvre des différents modèles considérés en termes de variables explicatives considérées et de méthodes de régression.

6.1.1 Types d'information

Rappelons que le point émulé de la grille à HR est noté (p, q), et le point correspondant de la grille à BR est noté (p', q'). Trois types de variables explicatives sont mis en œuvre de la façon suivante :

Informations globales La décomposition en Fonctions Orthogonales Empiriques (EOF) se fait sur toute la zone d'étude à BR. La taille d'un champ de vent ECMWF est de 11 pixels $\times$ 8 pixels, ce qui fait un total de

88 points de la grille à BR. Les composantes zonale et méridionale d'un vecteur de vent à BR sont toutes les deux utilisées pour la prédiction, ce qui donne un vecteur de 176 variables explicatives. Pour n données utilisées, la décomposition se fait sur une matrice de dimension $n \times 176$. Finalement, les 8 premiers modes de la décomposition, qui représentent environ 96% de variance expliquée et qui expliquent principalement la variabilité au large, sont conservés. La matrice des coefficients des projections sur les 8 modes EOFs est utilisée comme matrice de variables explicatives.

Informations locales Les informations sur une fenêtre carrée centrée sur (p', q') sont utilisées comme variables explicatives (cf. § 5.2.3). On fait varier la taille de la fenêtre qui correspond au nombre de pixels séparant le point (p', q') du bord de la fenêtre pour évaluer son influence. Le nombre de points à l'intérieur du carré est toujours égal au carré de la taille de fenêtre : $(2 \times Wz + 1)^2$. Les évaluations sont réalisées avec des tailles de fenêtre variant de 0 à 4 soit un nombre de points $\{1\ 9\ 25\ 49\ 81\}$.

Informations non-locales Pour la sélection de variables, les points sont choisis en fonction du résultat du calcul d'entropie conditionnelle en (p, q) (§ 5.2.4). Les points de la grille à BR sont classés par entropie conditionnelle moyenne croissante. Pour évaluer l'influence du nombre des points sélectionnés pour l'émulation, les K premiers points classés sont utilisés pour définir le vecteur de variables explicatives. Pour être cohérent avec la méthode des informations locales, le nombre de points est pris parmi les valeurs suivantes $K = \{1\ 9\ 25\ 49\ 81\}$.

Pour chaque champ de vent à BR, le vecteur de variables explicatives est défini par un des ces trois types d'information. Toutes les réalisations des variables explicatives constituent la matrice d'entrée $\mathbf{X}$. Une fois la matrice d'entrée construite, chaque colonne de $\mathbf{X}$ est normalisée. La normalisation est surtout importante pour la méthode SVR. Elle permet à chaque émulation de partager la même grille hyper-paramètres (cf. § 5.3.3.2). Par exemple, les coefficients de décomposition EOF n'ont pas la même échelle que les données de vent. Ici, la normalisation par rapport aux valeurs minimales et maximales est utilisée : chaque colonne est normalisée entre -1 et 1 suivant l'équation (2.2) dans la partie 2.1.3.

6.1.2 Méthodes de régression

Les quatre méthodes de régression sont mises en œuvre de la façon suivante :

Le plus proche voisin (NN) choisit simplement la première situation pour laquelle les deux conditions suivantes sont vérifiées : elle est la plus proche de la nouvelle situation d'entrée, et la HR correspondante en (p, q) est disponible. Cela revient à calculer la distance entre le vecteur de

variables explicatives extraites de la nouvelle situation à BR et la matrice d'entrée. La HR du j^e vecteur le plus proche de la nouvelle situation est affectée à la prédiction en (p, q).

Analogue par somme pondérée (AN) utilise les premières situations analogues les plus proches qui ont une distance moyenne plus petite que $4\,\mathrm{m\,s^{-1}}$ et pour lesquelles les hautes résolutions en (p, q) sont disponibles. Le poids de chaque situation analogue est calculé par une fonction exponentielle en fonction de la distance entre le nouveau vecteur et ces situations analogues (cf. Chapitre 5, l'équation 5.13). La prédiction en (p, q) est égale à la somme des hautes résolutions des situations analogues attribuées proportionnellement selon leur poids.

Régression Linéaire Multiple (MLR) est implémentée en utilisant la fonction `regress` du logiciel Matlab, en ajoutant une colonne de constante 1 qui correspond à un biais constant dans la matrice de variables expliquées (cf. Chapitre 5, l'équation 5.6). Le résultat de la sortie de la fonction `regress` correspond aux coefficients de la régression. La prédiction de la nouvelle situation en (p, q) est égale au produit scalaire entre les coefficients appris et le nouveau vecteur de variables explicatives.

Machine à Vecteurs de support pour la Régression (SVR) est réalisée à l'aide de la bibliothèque fournie par Chang et Lin (2011). Dans ce paquet, plusieurs types de techniques SVR sont implémentées. Nous utilisons le type « epsilon-SVR ». La fonction de noyau choisie est de type gaussien $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$ avec un paramètre scalaire γ. Il y a donc trois hyper-paramètres à fixer : γ, le paramètre de la fonction de noyau, C, le paramètre de régularisation entre la complexité du modèle et les erreurs, et la marge d'erreur ϵ, la marge d'erreur. Dans la section 5.3.3.2, les techniques pour obtenir les approximations des meilleurs hyper-paramètres ont été introduites. En prenant en compte les valeurs données par ces techniques et une vérification manuelle des performances de la grille globale, une sous-grille est sélectionnée avec $\gamma = \{0.001\ 0.01\ 0.04\ 0.1\ 1\}$, $C = \{14\ 30\ 60\ 90\ 150\}$ et $\epsilon = \{0.3\ 0.6\ 0.9\ 1.2\ 1.5\}$. La combinaison optimale d'hyper-paramètres est obtenue par validation croisée en 3 parties (cf. Section 6.2.1).

6.2 Protocole expérimental

La procédure de la validation croisée permet d'obtenir l'émulation de tout le catalogue. Par contre, les traitements en parallèle permettent de réaliser toute les expériences en un temps de calcul raisonnable. Les résultats obtenus pour les différents modèles sont évalués et comparés en suivant des paramètres statistiques différents.

Pour faciliter les comparaisons, une séries des points vis-à-vis de différents critères statistiques représentatifs des trois zones différentes (large, côtière,

fjord) sont utilisées pour notre analyse. Une évaluation plus globale est faite pour le modèle optimal choisi.

6.2.1 Validation croisée

Dans un objectif de validation, la technique de validation croisée est utilisée pour obtenir les émulations pour toute la base de données pour chaque modèle. Le principe de la validation croisée est de séparer les données en deux sous-ensembles : « ensemble d'apprentissage » et « ensemble de validation ». La validation croisée désigne le processus qui permet de tester la précision prédictive d'un modèle dans un échantillon test, l'ensemble de validation, par rapport à la précision prédictive de l'échantillon d'apprentissage à partir duquel le modèle a été développé. Cette dernière technique est beaucoup utilisée dans l'apprentissage statistique.

Il existe plusieurs techniques pour réaliser des validations croisées en construisant des échantillons de validation et des échantillons d'apprentissage indépendants. Le choix se fait par rapport au nombre de données d'apprentissage nécessaire. Dans cette étude, la technique de validation croisée en *N-folds* est utilisée. Les données sont divisées en N parties égales disjointes ; chaque fois, une seule partie est utilisée comme jeu de test et le reste ($N-1$ parties) est réservé pour l'apprentissage. L'opération se répète ainsi N fois pour qu'en fin de compte chaque partie soit utilisée exactement une fois comme ensemble de validation. La moyenne des erreurs quadratiques et les mesures par autre type de paramètres statistiques (§ 6.2.3) de N opérations sont enfin calculées pour estimer la performance de prédiction.

La figure 6.1 illustre le schéma de la division en N partie. Une partie des couples de données $\{x_j, y_j\}$ est mise de côté (en vert) pour la validation. Le modèle est développé à partir du reste des données et les prédictions pour les données $\{x_j\}$ sont réalisées avec le modèle appris f. Dans l'opération suivante, par exemple, $\{x_i, y_i\}$ est choisi comme données de validation. Les opérations sont répétées jusqu'à l'évaluation des performances de prédiction pour toutes les données disponibles.

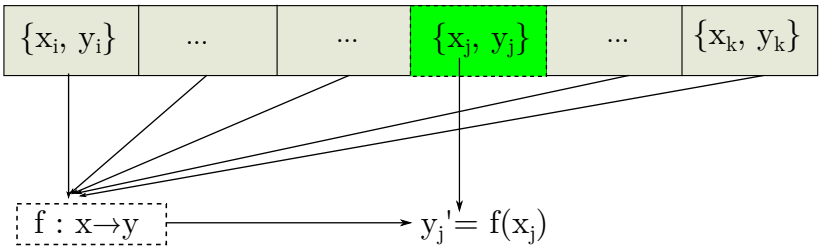


FIGURE 6.1 – *Schéma de la validation croisée avec la division en N parties.* $\{x_j, y_j\}$ *est l'ensemble de validation ; le reste des ensembles est utilisé pour l'apprentissage du modèle* f. *Avec le modèle appris* f, *les prédictions* $\{y'_j\}$ *pour le* j^e *ensemble de validation sont calculées et* $\{y_j\}$ *sont utilisés comme références.*

Cette technique de validation croisée est également utilisée dans la phase

d'apprentissage pour calibrer les hyper-paramètres des modèles. Les hyper-paramètres sont associés à chaque modèle. Les données réservées pour chaque apprentissage sont divisées elle-même en deux parties : une partie pour l'apprentissage du modèle et l'autre partie pour tester le performance du modèle. La combinaison des hyper-paramètres qui a l'erreur de test en moyenne la plus petite est choisie comme la combinaison optimale (cf. 5.3.3.2).

6.2.2 Parallélisme

Les trois premières méthodes d'apprentissage sont peu gourmandes en ressource de calcul : moins de 0.5 seconde par point à HR pour apprendre le modèle. Par contre, la méthode SVR a besoin de beaucoup plus de temps pour établir un modèle : environ 10 secondes par point à HR pour une grille de 125 combinaisons de hyper-paramètres. Pour une zone de $5.0^\circ \times 3.5^\circ$ à la résolution de SAR, soit 0.1° en longitude et en latitude, il faut environ 20 jours de calcul séquentiel sur une machine standard (Intel Core i7 à 2.20 GHz avec 8 Go de RAM). Pour réaliser une validation croisée en *N-folds*, par exemple avec $N = 20$, il faudrait environ 400 jours.

Pour réaliser ces expériences en un temps raisonnable, il est possible de paralléliser les traitements sur un *cluster* de machines. Celui utilisé est celui du Centre ERS d'Archivage et de Traitement (CERSAT), et sa particularité est d'interconnecter des nœuds de calcul par un réseau Gigabit milieu de gamme. Comme chaque point à HR est émulé indépendamment, une parallélisation triviale est suffisante, et comme celle-ci ne sollicite pas de communication inter-nœuds, ce type de cluster s'est révélé particulièrement adapté.

Ce cluster est composé d'environ 70 serveurs DELL, chacun étant équipé de bi-processeurs 6 cœurs (12 avec la technologie *hyperthread*), et de 64 Go de RAM. Ce qui fait plus que 1600 cœurs au total, et permet donc de descendre le temps de calcul à moins d'une demi journée.

Les données SAR sont volumineuses. Pour ne pas saturer la RAM dont l'usage croit avec le nombre de cœurs utilisés, les données d'apprentissage sont sauvegardées dans des fichiers de petite taille. Les données de l'émulation (les résultats) de chaque tâche sont sauvegardées dans des fichiers élémentaires en fonction du découpage de la parallélisation. Une fois finis tous les calculs, tous les résultats sont reconstitués pour obtenir l'émulation complète de chaque champ de vent dans le catalogue.

6.2.3 Mesures de performance

Les paramètres tels que les erreurs moyennes, les variances expliquées, les coefficients de corrélation ou les quantiles peuvent aider à quantifier et qualifier la capacité prédictive d'un modèle de régression. Par convention, la donnée d'observation SAR est appelée la référence.

Les erreurs moyennes d'une série temporelle mesurent la norme de la

différence entre les vecteurs émulés $\{u, v\}$ et les vecteurs référence $\{u_{\text{ref}}, v_{\text{ref}}\}$:

$$\overline{E} = \frac{1}{n_t} \sum_{t \in \tau} \sqrt{(u_t - u_{\text{ref},t})^2 + (v_t - v_{\text{ref},t})^2} \tag{6.1}$$

où τ est l'ensemble des instants t où les données de références sont disponibles et n_t est le nombre de références.

La variance est une mesure servant à caractériser la dispersion d'une distribution. Comparer la variance expliquée entre les émulations et les références permet de mesurer la capacité de conservation des variances par un modèle d'émulation. Prenons pour exemple la composante zonale u. La variance expliquée est obtenue par :

$$R_u^2 = 1 - \frac{\text{Var}(u - u_{\text{ref}})}{\text{Var}(u_{\text{ref}})} \tag{6.2}$$

où Var est la fonction de variance.

Le coefficient de corrélation mesure la corrélation linéaire entre deux variables. Le coefficient de corrélation entre la prédiction et la référence est donné pour la composante par :

$$r_u = \frac{\sigma_{u,\, u_{\text{ref}}}}{\sigma_u \cdot \sigma_{\text{ref}}} = \frac{\sum_i (u_i - \overline{u}) \cdot (u_{\text{ref},i} - \overline{u_{\text{ref}}})}{\sqrt{\sum_i (u_i - \overline{u})^2} \cdot \sqrt{\sum_i (u_{\text{ref},i} - \overline{u_{\text{ref}}})^2}} \tag{6.3}$$

Le coefficient de corrélation pour la composante méridionale v peut être calculé de la même manière.

De façon empirique, on peut dire que le quantile E_q d'ordre q est une valeur qui partage la série statistique ordonnée en deux sous-ensembles qui contiennent respectivement un nombre d'observations égal à nq et $n(1-q)$. L'équation suivante donne le quantile pour un niveau d'erreur d'émulation :

$$P(E > E_q) = q \tag{6.4}$$

C'est-à-dire que un pourcentage q de données ont une erreur d'émulation inférieure à E_q.

6.2.4 Points caractéristiques

Pour faciliter les illustrations et les comparaisons entre les différents modèles, une série de points (cf. Figure 6.2) est choisie sur la grille à HR, suivant deux critères : être représentatif des différentes zones, et être prélevé sur des endroits où la variabilité entre la basse et la haute résolution est élevée pour les points dans la zone côtière. Ce sont les mêmes points utilisés pour les analyses spatiales conjointes dans la section § 4.2.

Les données SAR en ces points sont utilisées pour évaluer les différents modèles.

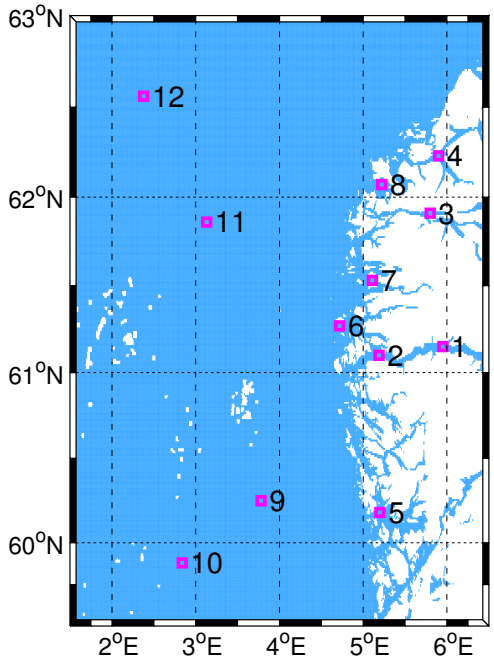


FIGURE 6.2 – *Rappel de la zone d'étude et des 12 points sélectionnés pour les évaluations des différents émulateurs.*

6.3 Synthèse des évaluations des modèles

La performance des modèles dépend des éléments suivants :

1. nombre de variables explicatives ;
2. type d'information utilisées ;
3. méthode de régression utilisée.

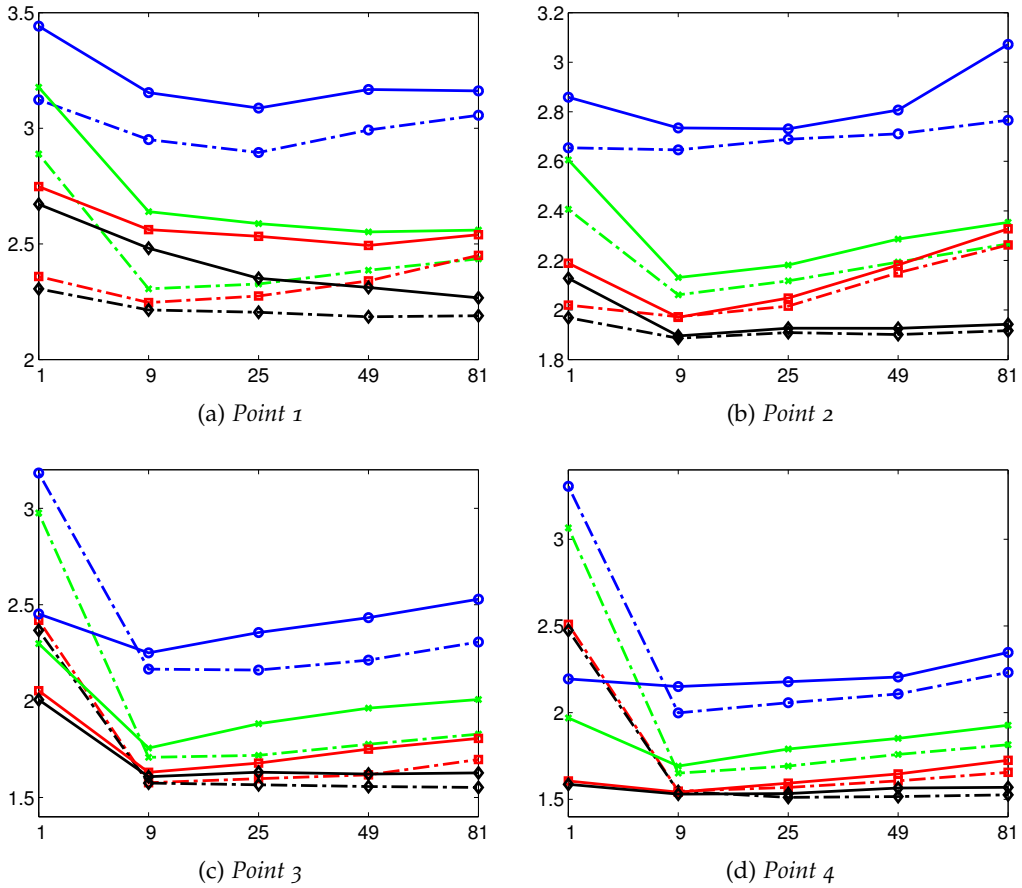
Les comparaisons de toutes les combinaisons permettent d'évaluer l'influence de chaque élément sur la performance des modèles.

6.3.1 Influences du nombre de variables explicatives

La première validation consiste à montrer les influences du nombre de variables explicatives extraites à partir d'un champ de vent à BR sur les différents modèles. Dans l'évaluation, chaque type d'information est paramétré avec un nombre de variables explicatives différent, hormis pour les informations globales.

Pour chaque point émulé, une validation croisée en 19 parties est utilisée pour obtenir les émulations de 758 échantillons (la taille du catalogue), ce qui correspond à environ 5% des données de validation à chaque opération. La figure 6.3 représente les erreurs moyennes des différents modèles aux 12 points choisis pour la validation (cf. Figure 6.2), en fonction : du type d'information avec points tirets pour les informations locales et ligne solide pour les informations non-locales ; du type de méthode de régression avec les cercles bleus pour le plus proche voisin, les croix vertes pour méthode analogue

par somme pondérée, les carrés rouges pour la méthode MLR et les diamants noirs pour la méthode SVR. L'axe des abscisses représente le nombre de points utilisés à BR pour les variables explicatives. L'échelle de l'axe des ordonnées de chaque figure n'est pas alignée pour permettre de zoomer les courbes au maximum et ainsi mieux comparer les courbes sur la même figure.



(a) *Point 1* (b) *Point 2*

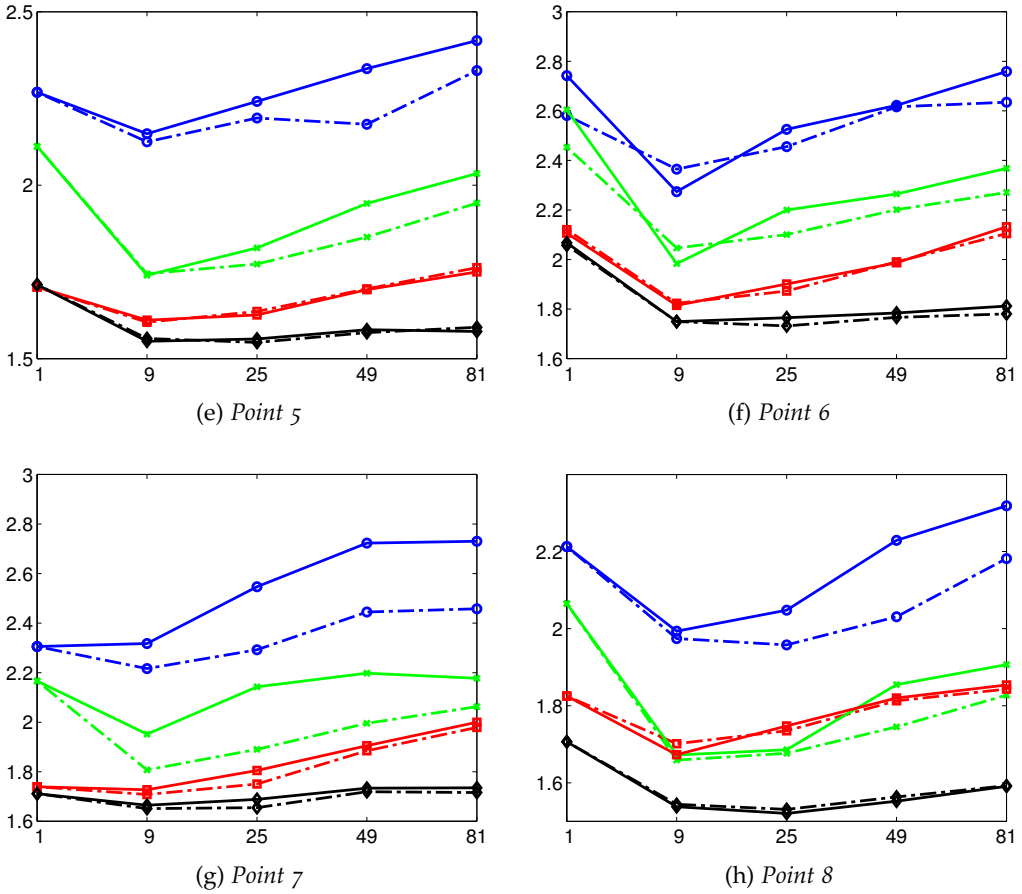
(c) *Point 3* (d) *Point 4*

FIGURE 6.3 – *Partie I : Zones fjords*

Influences du nombre de variables explicatives sur les performances des différentes méthodes de régression. L'axe des abscisses représente le nombre de points utilisés à BR pour les variables explicatives ; L'axe des ordonnées représente les erreurs moyennes (en $\mathrm{m\,s^{-1}}$*) résultant de la validation croisée. Les différentes courbes correspondent aux différentes méthodes de régression : voisin le plus proche (cercles bleus), méthode analogue par somme pondérée (croix vertes), Régression Linéaire Multiple (MLR) (carrés rouges) et Machine à Vecteurs de support pour la Régression (SVR) (diamants noirs). Pour chaque méthode, les deux types d'information sont comparées — informations locales (points tirets) ; informations non-locales (ligne solide).*

À partir de la figure 6.3, on peut remarquer que :

1. les erreurs moyennes des méthodes de régression varient entre $1.5\,\mathrm{m\,s^{-1}}$ et $3.8\,\mathrm{m\,s^{-1}}$. Globalement, la taille 9 donne des erreurs moyennes les plus petites. La taille 1 et 81 ont toutes les deux généralement de très mauvaises performances ;
2. pour chaque nombre de variables, le modèle SVR (diamants noirs) atteint une meilleur performance d'émulation en terme d'erreurs moyennes



(e) *Point 5* (f) *Point 6*

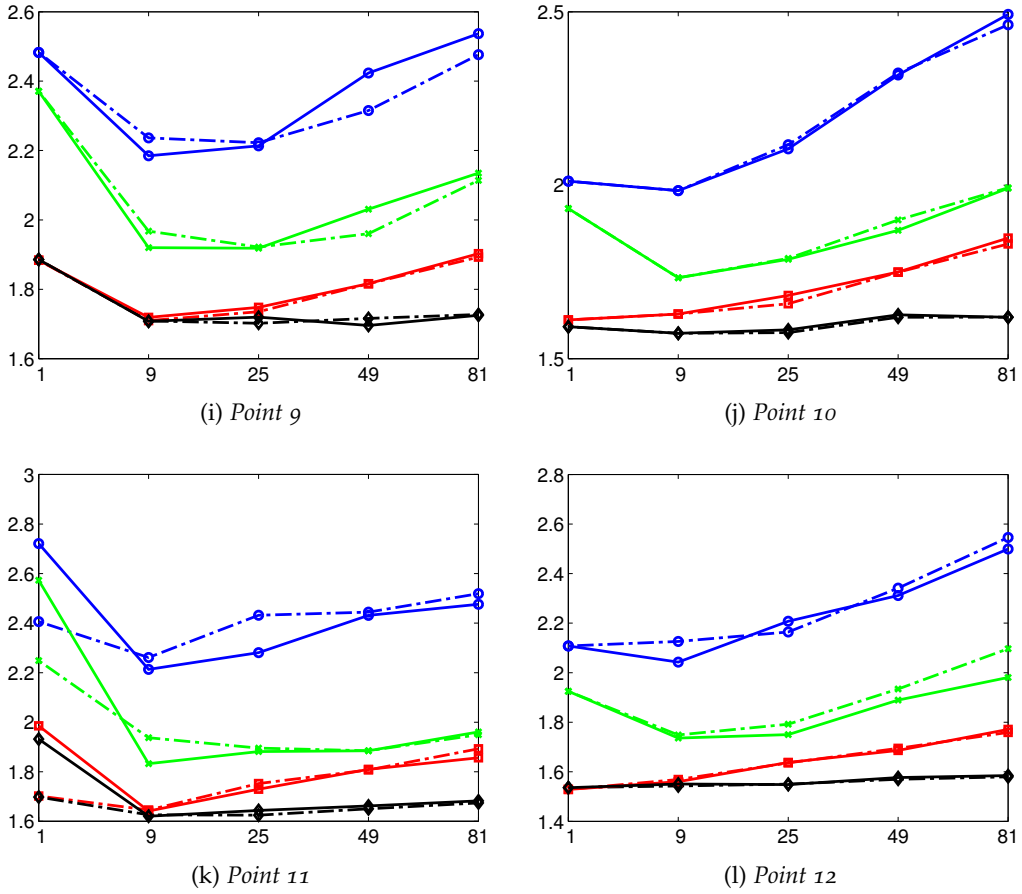
(g) *Point 7* (h) *Point 8*

FIGURE 6.3 – *Partie II : Zones côtières.*

(autour de $1.7\,\mathrm{m\,s^{-1}}$) par rapport aux trois autres méthodes, et elle reste beaucoup moins sensible au nombre de variables utilisées ;

3. au contraire, les méthodes analogues et MLR sont très sensibles au nombre de variables. Pour la méthode MLR, si le nombre de variables est optimal (ce qui correspond ici à au nombre 9), elle peut presque atteindre le même niveau de performance que le modèle SVR. Cela s'explique par une corrélation fortement linéaire avec les voisins directs (portée entre 50 km et 100 km). Plus des informations non linéairement corrélées sont ajoutées, plus la capacité prédictive de MLR décroît ;
4. même si le modèle SVR est moins sensible au nombre de variables sélectionnées, en moyenne, la taille 9 a tendance à donner de meilleur résultat.

En résumé, Le nombre de variables explicatives a plus d'influence sur les méthodes analogues et la méthode MLR, et moins sur la méthode SVR. Suivant ce critère, la méthode SVR apparaît beaucoup plus robuste par rapport aux trois autres méthodes. Globalement, la taille 9 correspond à un nombre optimal pour toutes les méthodes. Dans les validations suivantes, cette taille est utilisée pour chaque modèle.



(i) *Point 9* (j) *Point 10*

(k) *Point 11* (l) *Point 12*

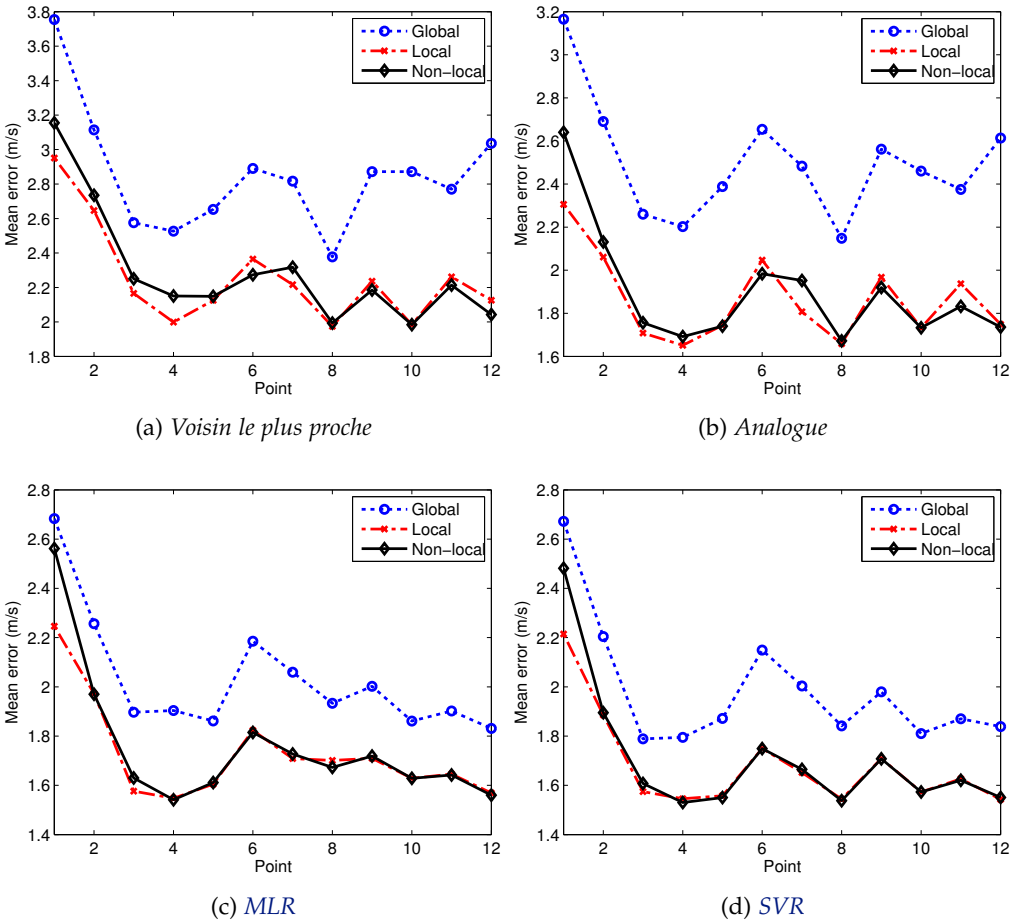
FIGURE 6.3 – *Partie III : Au large.*

6.3.2 Influence du type d'information

Les résultats précédents illustrent l'influence du nombre de variables explicatives. Pour comparer les différents types d'information, la figure 6.4 représente les erreurs d'émulation pour les informations globale (cercles bleus, pointillés), les informations locales (croix rouges, points tirets) et les informations non-locales (diamants noirs, ligne solide) pour chacune des méthodes de régression. Le nombre de points égal à 9 est utilisée pour les deux derniers types d'information.

À partir de la figure 6.4, on peut remarquer que généralement, les erreurs d'émulation sont beaucoup plus élevées en utilisant les informations globales par rapport à celles des deux autres types d'information. Les informations locales et non-locales ont des erreurs d'émulation très proches, sauf pour le point 1, où les informations locales ont des erreurs d'émulation plus petites.

Ces expériences montrent que l'émulation à haute résolution dépend plus des informations à basse résolution locale que globale. Vu que l'utilisation des informations globales a de plus mauvaises performances, on n'utilise plus que les informations locales et non-locales dans les comparaisons suivantes.



(a) *Voisin le plus proche* (b) *Analogue*

(c) *MLR* (d) *SVR*

FIGURE 6.4 – *Comparaison entre les différents types d'information : les informations globales (cercles bleus, pointilles), les informations locales (croix rouges, points tirets) et les informations non-locales (diamants noirs, ligne solide). En terme d'erreur moyenne* ($\mathrm{m\,s^{-1}}$) *pour les différentes approches de régression : Voisin le plus proche (a), méthode analogue par somme pondérée (b), Régression Linéaire Multiple (MLR) (c) et Machine à Vecteurs de support pour la Régression (SVR) (d).*

6.3.3 Comparaison entre méthodes de régression

Avant de comparer les différentes méthodes de régression, il est intéressant d'analyser la performance d'une méthode simple comme la méthode d'interpolation spatiale pour obtenir les émulations à HR. Nous considérons une méthode d'interpolation spatiale bilinéaire. La figure 6.5 compare les erreurs d'émulation entre la méthode d'interpolation et la méthode SVR non-locale. Elle montre que la méthode d'interpolation a des erreurs de prédiction plus élevées sur tous les points dans la zone côtière (points 1 à 8) et que le niveau d'erreurs diffère beaucoup d'un point à l'autre. Ce résultat peut être interprété par le fait que la haute résolution reconstruite dépend uniquement de la BR et qu'elle ne permet pas de tenir compte des variabilités sous mailles (sur une grille HR).

Ensuite, les différentes méthodes de régression proposées dans la sec-

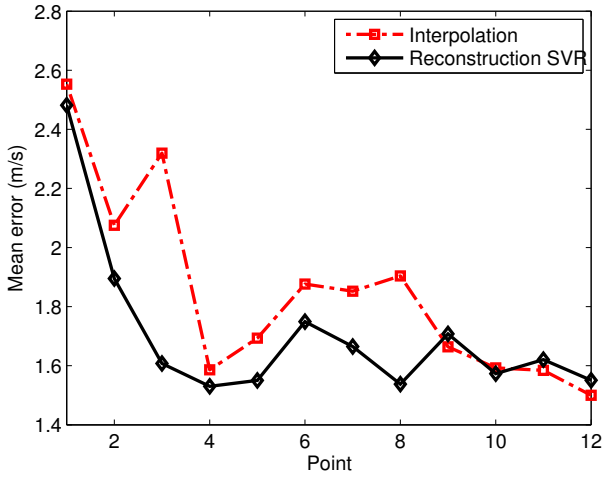


FIGURE 6.5 – *Comparaison des erreurs de prédiction entre la méthode d'interpolation (carrés rouge, points tirets) et la méthode Machine à Vecteurs de support pour la Régression (SVR) (diamants noirs, ligne solide).*

tion 6.1.2 sont comparées, en utilisant les informations locales et non-locales avec un nombre de variables optimal égal à 9. La figure 6.6 montre les erreurs d'émulation pour la méthode du plus proche voisin (cercles bleu), la méthode analogue par somme pondérée (croix vert), la méthode MLR (carrés rouge) et la méthode SVR (diamants noirs). La méthode SVR, avec les informations locales et non-locales, montre des meilleurs capacités de prédiction sur les 12 points, avec des erreurs de prédiction autour de $1.7\,\mathrm{m\,s^{-1}}$. On peut constater également que les erreurs de prédiction sont plus élevées pour les points 1 et 2. Tous les deux se situent dans le même *fjord*.

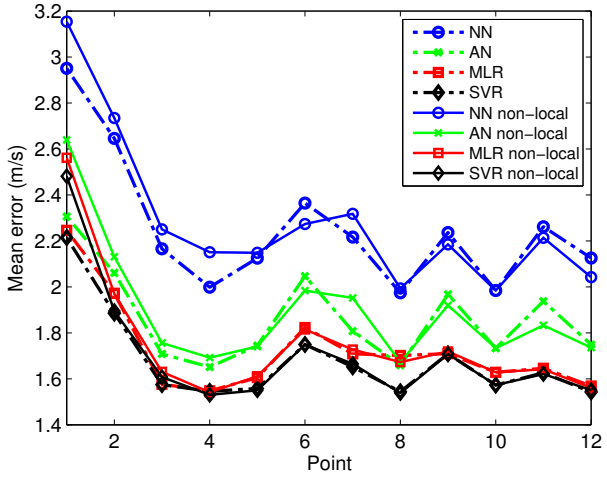
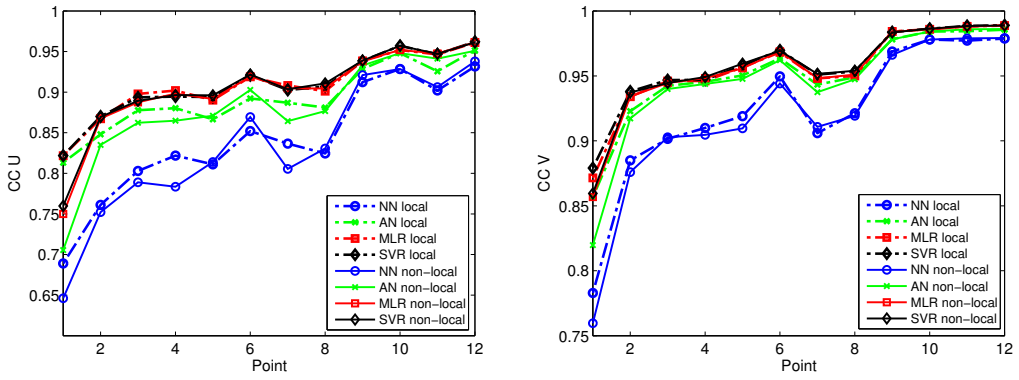


FIGURE 6.6 – *Comparaison entre les différents émulateurs en terme d'erreur moyenne* ($\mathrm{m\,s^{-1}}$) *pour les différentes approches : Le plus proche voisin (NN, cercles bleu), méthode analogue (AN, croix vert), Régression Linéaire Multiple (MLR) (carrés rouge) et Machine à Vecteurs de support pour la Régression (SVR) (diamants noirs). Pour chaque méthode de régression, deux types d'information sont utilisés : informations locales (points tirets) et informations non-locales (ligne solide).*

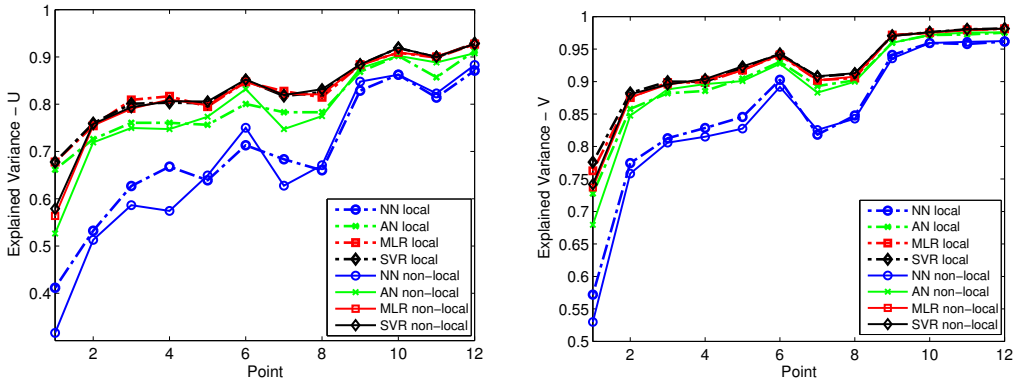
La figure 6.7 compare les coefficients de corrélation entre les vents émulés et

les références pour les composantes u et v des différents modèles. Elle montre que les émulations par les méthodes SVR et MLR, combinant les informations locales ou non-locales, sont très corrélées avec les observations, entre 0.82 et 0.95 pour la composante u et entre 0.85 et 0.99 pour la composante v. La comparaison des variances expliquées par les différents modèles (cf. Figure 6.8) a la même tendance que celle des coefficients de corrélation. Elles varient entre 80% et 90% pour la composante u et 90% et 98% pour la composante v, sauf pour le point 1 pour lequel la variance expliquée est plus petite.



(a) *Composante u* (b) *Composante v*

FIGURE 6.7 – *Comparaison des coefficients de corrélation entre les vents émulés et les références pour les composantes u (a) et v (b) des différents modèles. La légende est identique à celle de la figure 6.6.*

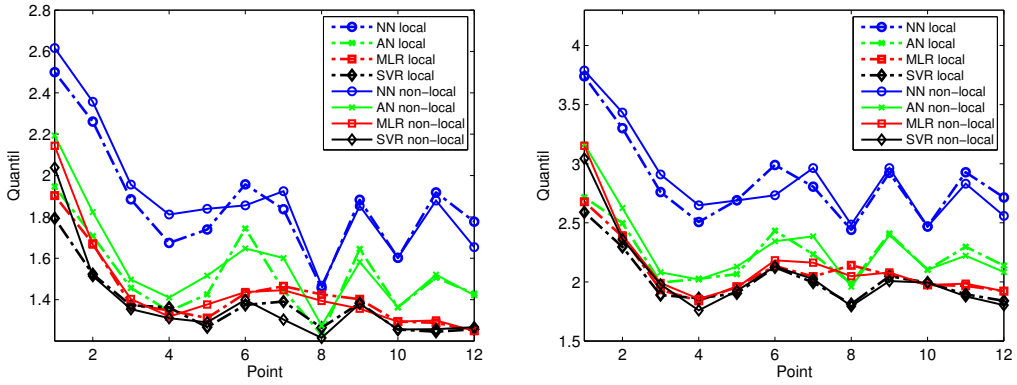


(a) *Composante u* (b) *Composante v*

FIGURE 6.8 – *Comparaison des variances expliquée entre les vents émulés et les références pour les composantes u (a) et v (b) des différents modèles. La légende est identique à celle de la figure 6.6.*

La figure 6.9 illustre la comparaison des quantiles des erreurs entre les différents modèles. Pour 50% des cas, les erreurs pour le meilleur modèle (SVR avec les informations non-locales) sont inférieures à $1.4\,\mathrm{m\,s^{-1}}$ (sauf pour le point 1 avec $2\,\mathrm{m\,s^{-1}}$). Pour 70% des cas, les erreurs pour le meilleur modèle sont inférieures à $1.8\,\mathrm{m\,s^{-1}}$ (sauf pour le point 1 avec $3\,\mathrm{m\,s^{-1}}$). Pour les points en zones côtières (points 4 à 8), les différences des quantiles des erreurs entre les

modèles basés sur la MLR et les modèles basés sur la SVR sont plus importantes qu'aux autres points. Par exemple, pour 50% des cas ($q = 0.5$, cf. Figure 6.9a), les erreurs au point 8 pour les modèles basés sur la SVR sont inférieures à $1.2\,\mathrm{m\,s^{-1}}$ et pour les modèles basés sur la MLR elles sont inférieures à $1.5\,\mathrm{m\,s^{-1}}$.



(a) $q = 0.5$ (b) $q = 0.7$

FIGURE 6.9 – *Comparaison des quantiles pour les différents modèles avec le quantile égale* 0.5 *(a) et* 0.7 *(b). La légende est identique à celle de la figure 6.6.*

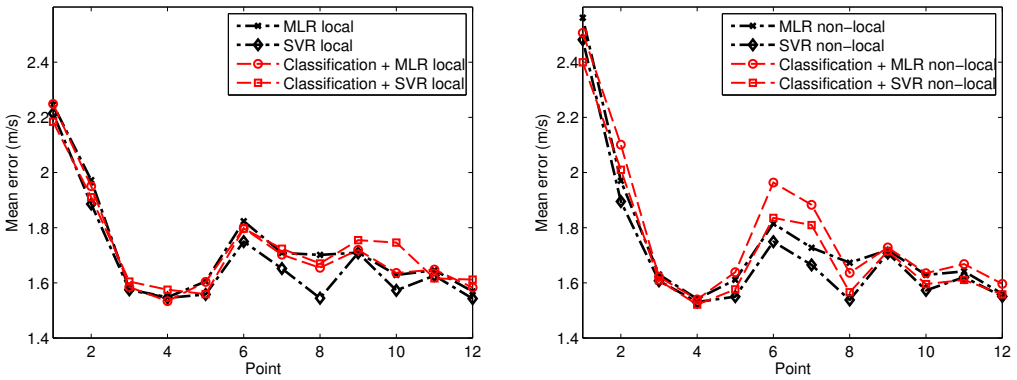
Les différents paramètres statistiques montrent que les modèles basés sur la méthode de régression SVR, combinant l'approche de sélection des variables explicatives locales et non-locales, donnent de meilleurs résultats. Ensuite viennent, par ordre décroissant en terme de performance, la méthode MLR, la méthode analogue par somme pondérée et enfin la méthode du plus proche voisin.

En zones *fjords*, les erreurs d'émulation restent beaucoup plus élevées que celles en zones côtières et au large pour tous les modèles proposés. La variabilité à HR en zones *fjords* dépend probablement d'autre types de paramètres en plus des vecteurs vent à BR.

6.3.4 Influence de la direction du vent

Cette partie consiste à évaluer si l'ajout d'une couche de classification avant la régression, par exemple, une classification sur la base de la direction de vent, peut améliorer la performance prédictive du modèle.

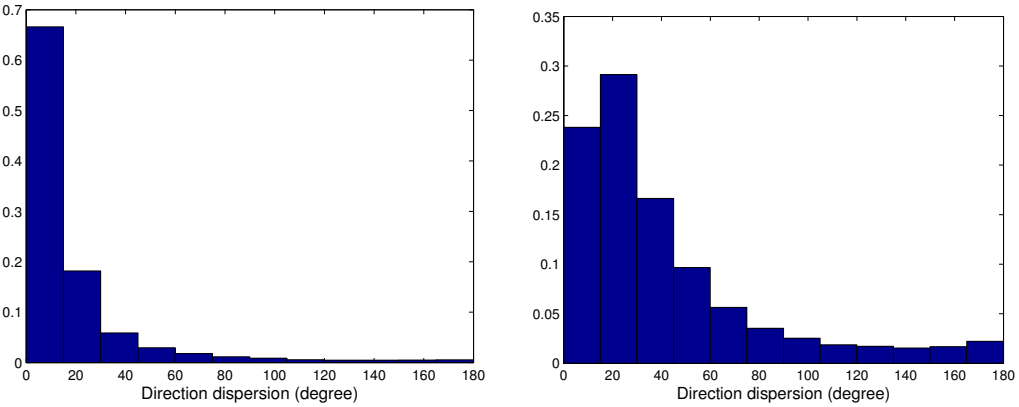
La figure 6.10 compare les erreurs moyennes entre les modèles sans et avec la classification par direction. Pour la comparaison, deux méthodes de régression, MLR et SVR, avec les informations locales (6.10a) et les informations non-locales (6.10b) sont utilisées. Pour les informations non-locales, l'emplacement des points critiques change en fonction de la direction (cf. 5.2.4.3). Le nombre de points utilisés pour former les variables explicatives est de 9. Les résultats montrent qu'il n'y a pas d'amélioration globalement en ajoutant la classification par direction. La classification associant la régression linéaire avec les informations locales améliore très légèrement la performance. Pour la régression SVR, elle conduit à une légère dégradation des résultats.



(a) *Informations locales* (b) *Information non-locales*

FIGURE 6.10 – *Comparaison des erreurs de prédiction entre les modèles sans classification (points tirets noirs) et avec classification (tiret rouge) pour MLR et SVR avec les information locales (a) et les informations non-locales (b).*

Le problème de classification des champs de vent par direction à BR est que les directions en chaque point sont différentes. Il est donc difficile de classer les vents par direction pour une zone donnée. La figure 6.11 montre les histogrammes des différences absolues (en degré) entre la direction locale et les directions de ses 8 voisins directs pour les deux zones différentes. La zone au large comprend tous les points qui ont une longitude plus petite que 4.0° et le reste des points est considéré comme en zone côtière. L'histogramme 6.11b montre que les différences sont beaucoup plus importantes dans la zone côtière qu'au large (cf. Figure 6.11a).



(a) *Offshore* (b) *Côtier*

FIGURE 6.11 – *Histogrammes des différences absolues (en degré) entre la direction locale et les directions de ces 8 voisins qui l'entourent pour une zone offshore (a) et pour une zone côtière (b).*

Pour une zone côtière, la classification des vents par direction est donc très difficile. Dans ce cas, il est plus pertinent de choisir les modèles de régression sans une couche de classification et de laisser les méthodes de régression gérer la non-linéarité.

6.3.5 Illustration des champs de vent HR émulés

On a choisi deux cas pour illustrer les résultats d'émulation par les différents modèles : celui du 12 mai 2008 (cf. 6.12) et celui du 4 mars 2009 (cf. 6.13). L'illustration permet de comparer les résultats des différents modèles. Pour le cas du 12 mai 2008, les émulations à HR par les modèles basés sur les méthodes analogues (en haut au milieu et à droite) sont plus proches de celles par les modèles basés sur les modèles MLR (en bas au milieu) et SVR (en bas à gauche) que pour le cas du 4 mars 2009. Cela peut probablement s'expliquer par le plus grand nombre de situations proches en direction et en intensité dans le catalogue pour le 12 mai 2008 comparativement au 4 mars 2009. Globalement, les méthodes MLR et SVR donnent des champs émulés mieux structurés et plus proches de l'observation SAR.

Pour les méthodes analogues, les hautes résolutions reconstruites ont moins d'effet de « tuiles » en utilisant les informations non-locales par rapport à l'utilisation des informations locales. L'émulation par les informations non-locales produit des champs plus lisses et nets. Cela signifie que l'approche par les informations non-locales peut être très intéressante pour l'émulation à HR dans le cas où une méthode analogue fonctionne aussi bien que les autres méthodes.

Du fait que les informations locales utilisent une fenêtre carrée centrée sur le point (p', q') à BR, tous les points sur la grille à HR qui correspond au point (p', q') sur la grille à BR ont les mêmes variables explicatives. Cela induit l'effet de « tuile » qui est moins visible en exploitant les informations non-locales. En effet l'emplacement des points critiques pour les informations non-locales varie d'un point à l'autre sur la grille à HR en fonction du calcul des entropies conditionnelles.

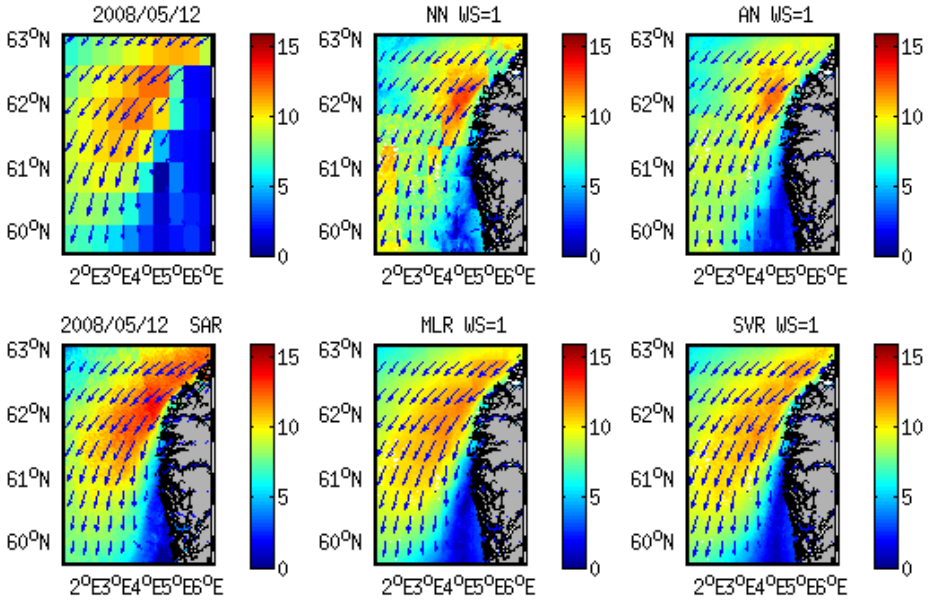
6.3.6 Synthèse

Cette section évalue les influence des différents éléments :

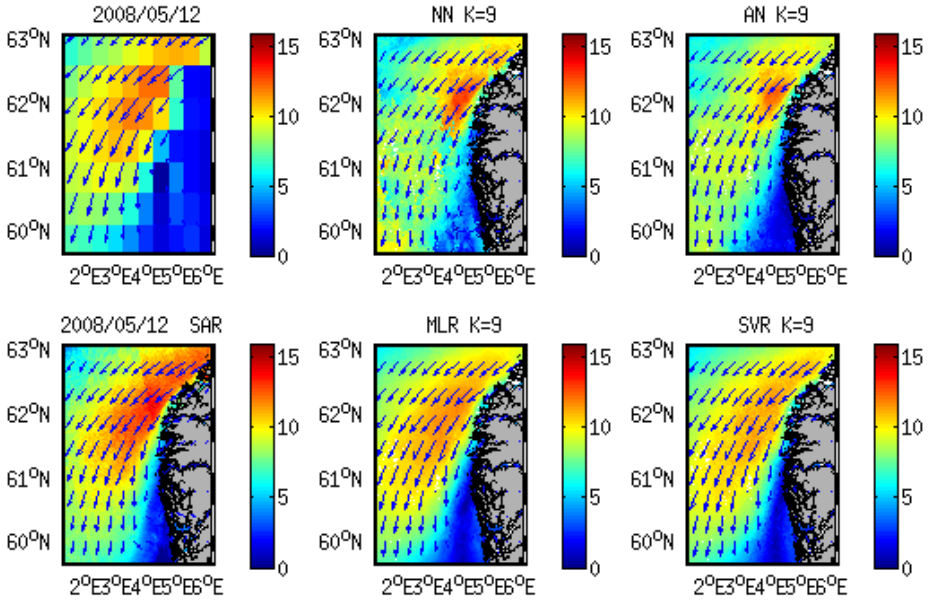
nombre de variables explicatives Le nombre des variables explicatives a plus d'influence sur les méthodes analogues et la méthode MLR que sur la méthode SVR. Le nombre qui correspond à 9 points utilisés pour construire le vecteur de variables explicatives conduit généralement aux erreurs d'émulation les plus petites. Ce nombre optimal est ensuite utilisé pour les comparaisons entre les différents modèles.

type d'information Chaque type d'information est évalué en association aux 4 méthodes de régression. Les 4 types de modèles de régression montrent que les informations globales ont des erreurs de construction beaucoup plus élevées par rapport aux informations locales et non-locales. Il y a peu d'apport du non-local par rapport au local sauf pour les méthodes analogues en termes d'artefact de reconstruction.

méthode de régression Les différents paramètres statistiques montrent que les

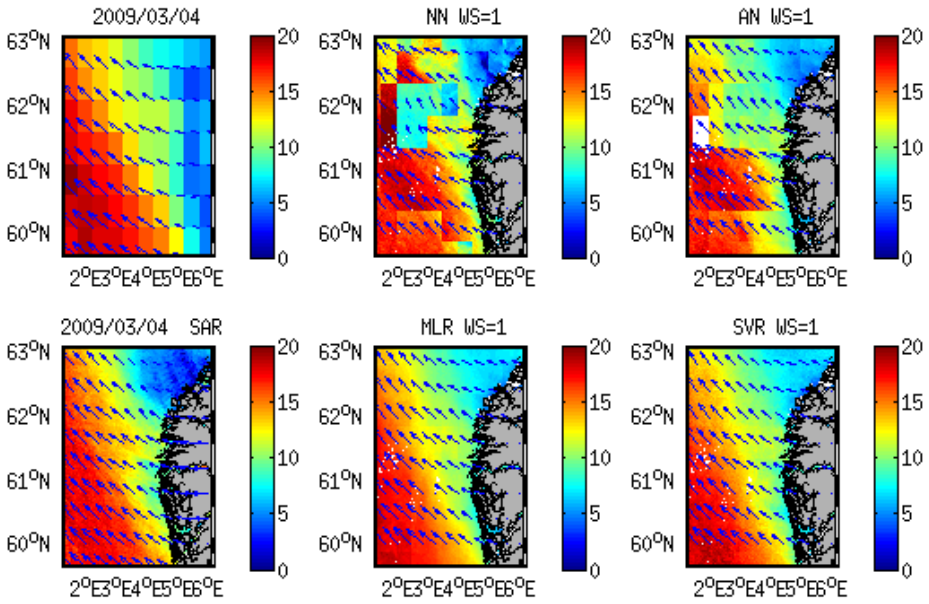


(a) *Informations locales*

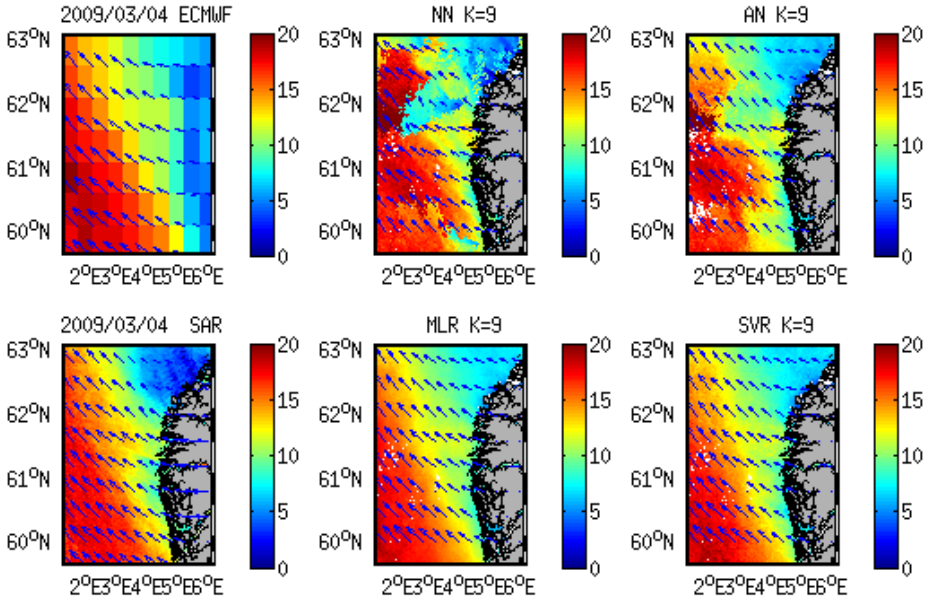


(b) *Informations non-locales*

FIGURE 6.12 – *Émulation du champ de vent ECMWF du 12 mai 2008 (en haut à gauche) par différentes méthodes de régression : plus proche voisin (NN, en haut au milieu), analogue par somme pondérée (AN, en haut à droite), Régression Linéaire Multiple (MLR) (en bas au mimieu) et Machine à Vecteurs de support pour la Régression (SVR) (en bas à droite), avec les informations locales (a) et les informations non-locales (b). Le champ SAR correspondant (en bas à gauche) est utilisé comme référence.*



(a) *Informations locales*



(b) *Informations non-locales*

FIGURE 6.13 – *Émulation du champ de vent ECMWF du 4 mars 2009 (en haut à gauche) par différentes méthodes de régression : plus proche voisin (NN, en haut au milieu), analogue par somme pondérée (AN, en haut à droite), Régression Linéaire Multiple (MLR) (MLR, en bas au mimieu) et Machine à Vecteurs de support pour la Régression (SVR) (en bas à droite), avec les informations locales (a) et les informations non-locales (b). Le champ SAR correspondant (en bas à gauche) est utilisé comme référence.*

modèles de régression par SVR donnent de meilleurs résultats, ensuite viennent, par ordre décroissant en terme de performance, la méthode MLR, la méthode analogue par somme pondérée et enfin la méthode du plus proche voisin.

classification par direction Les modèles comportant une couche de classification par direction sont comparés avec les meilleurs modèles sans classification. Les résultats montrent que la classification par direction ne conduit pas à des gains de performance, voir au contraire détériorent la qualité des émulations.

6.4 Évaluation des modèles optimaux

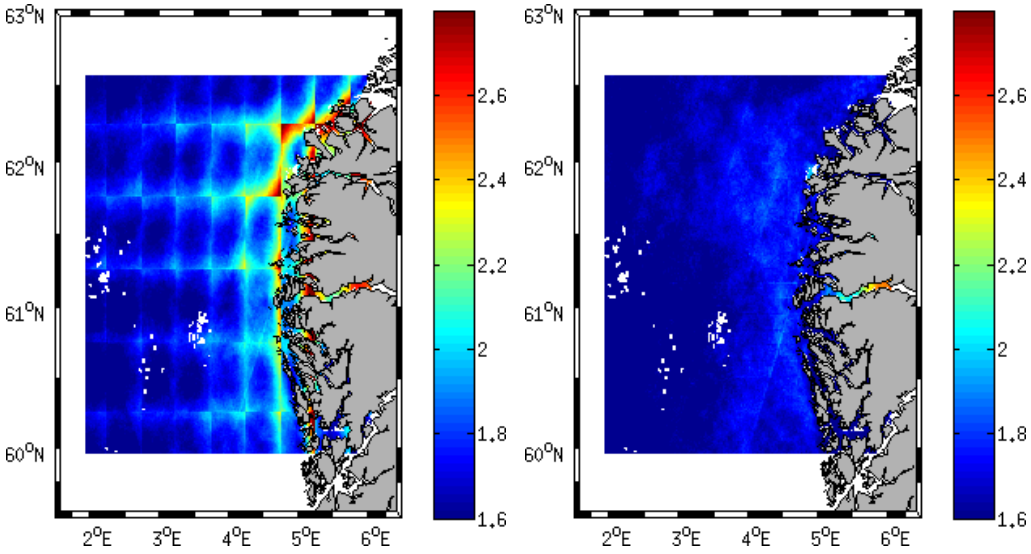
Finalement, les modèles de SVR sont considérés comme des modèles optimaux. L'utilisation des informations locales et non-locales donne des résultats très proches. La différence est que les informations non-locales permettent d'éviter l'effet de « tuile ». Les deux types de modèle sont utilisés pour l'émulation de l'ensemble des champs de vent dans le catalogue, mais seules les évaluations statistiques sur les modèles de SVR avec les informations non-locales sont illustrées.

La technique de validation croisée est utilisée pour obtenir l'émulation de toute le catalogue. Les même 19-*folds* sont utilisés ici pour obtenir les émulations de 758 champs de vents à HR pour la période 2005-2010. La zone émulée se limite à $E2.0^\circ$ – $E6.0^\circ$ et à $N60.0^\circ$ – $N62.5^\circ$ pour éviter la gestion des bords. L'évaluation du catalogue reconstruit se fait en amont au chapitre 4, les analyses statistiques conjointes entre les données ECMWF et les données SAR.

6.4.1 Comportement global

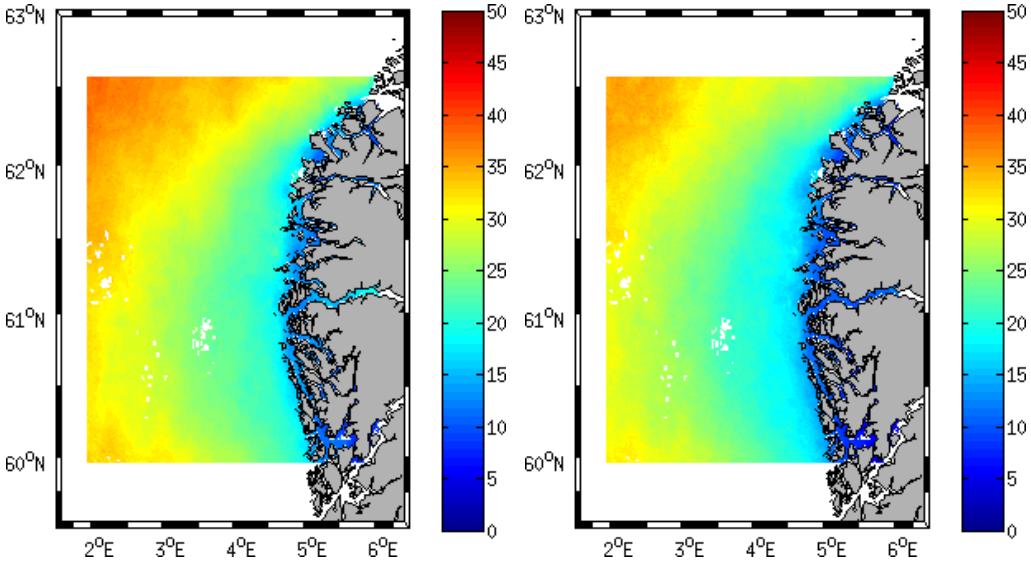
La figure 6.14 compare les erreurs moyennes entre les champs ECMWF et les champs SAR avec celles entre les champs émulés et les champs SAR. En chaque point, les erreurs moyennes sont calculées par l'équation (6.1). Cette figure montre que les erreurs d'émulation sont moins élevées que les différences entre les champs ECMWF et les champs SAR dans la zone côtière : elles sont d'entre $1.6\,\mathrm{m\,s^{-1}}$ et $1.8\,\mathrm{m\,s^{-1}}$. Dans le *fjord* à la latitude 61.15°, les erreurs d'émulation sont plus grandes. Cette zone coïncide avec les endroits présentant moins de données que le minimum nécessaire pour une bonne représentativité du catalogue, comme présenté en 3.3.4. De plus, la figure 4.1 dans la section 4.1 montre que la variabilité à HR dans cette zone est naturellement plus élevée.

La figure 6.15 compare les variances de la série temporelle en chaque point pour les deux composantes du SAR et de l'émulation. La variance de l'émulation est légèrement moins élevée que celle des observations, mais elles restent très comparables.



(a) *Distance entre ECMWF et SAR* (b) *Erreur d'émulation*

FIGURE 6.14 – *Comparaison des erreurs avant (a) et après (b) l'émulation. Les erreurs sont calculées avec toutes les données dans le catalogue ainsi que toutes les émulations du catalogue pour tous les points de la grille à HR.*

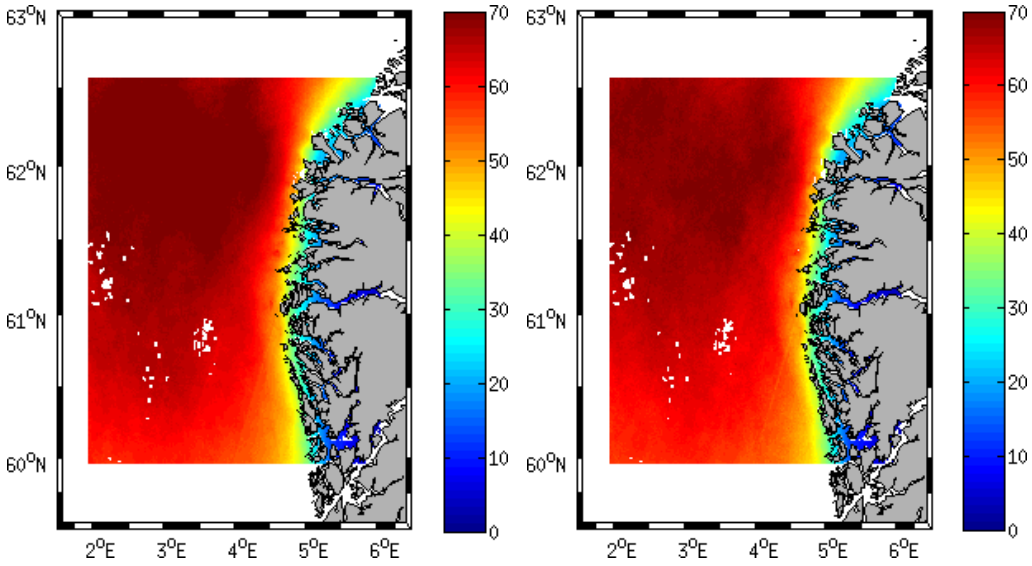


(a) *Composante u du champ SAR* (b) *Composante u du champ reconstruit*

FIGURE 6.15 – *Comparaison de la variance en chaque point pour la composant u du champ SAR (a) et pour la composante u du champ reconstruit (b).*

6.4.2 Roses des vents et diagrammes de dispersion

Puisque les deux composantes sont reconstruites séparément, la comparaison des roses des vents permet de voir la conservation de la relation de dépendance entre les composantes vent u et v. Ces roses des vents présentées dans la figure 6.16 peuvent être comparées avec celles du chapitre 4 (cf. § 4.2.1, Figure 4.4). Pour simplifier les illustrations, seuls 9 points parmi 12 sont utilisés. On observe ainsi que les roses des vents émulées sont beaucoup plus proches

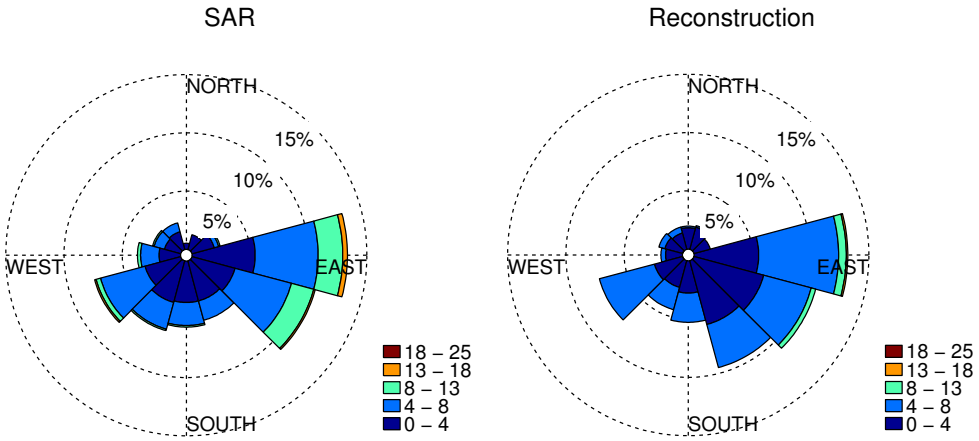


(c) *Composante v du champ SAR* (d) *Composante v du champ reconstruit*

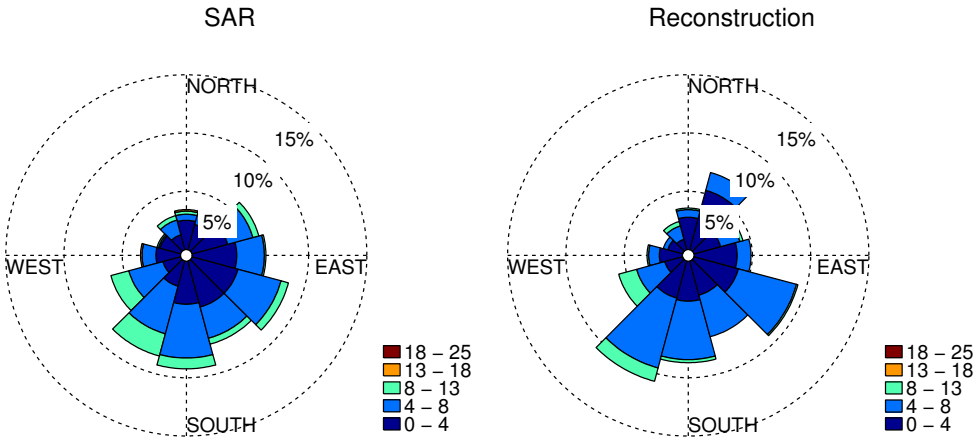
FIGURE 6.15 – *Comparaison de la variance en chaque point pour la composant v du champ SAR (a) et pour la composante v du champ reconstruit (b).*

de celles des observations par rapport à celles de l'ECMWF, quelques soient les zones. Les résultats montrent que la dépendance entre les composantes vent *u* et *v*est bien conservée même si elles sont émulées indépendamment.

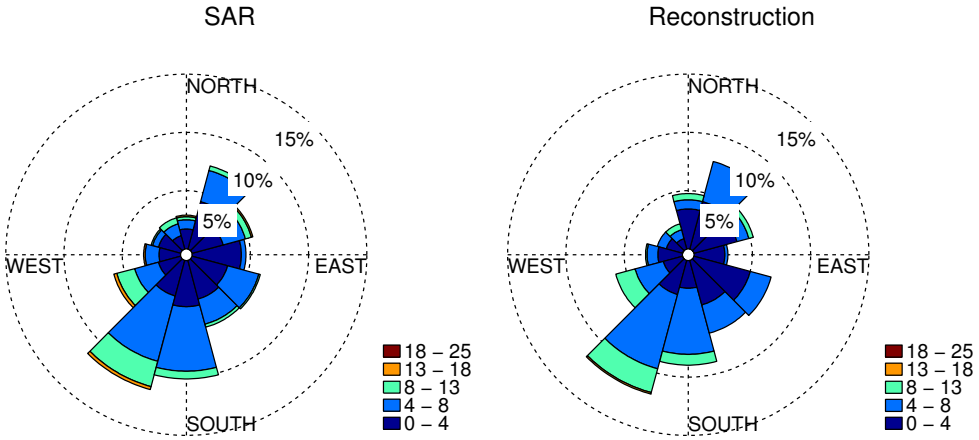
Les diagrammes de dispersion permettent de voir la corrélation entre les séries temporelles du SAR et de l'émulation. Ces diagrammes, illustrés à la figure 6.17, peuvent aussi être comparés avec ceux du chapitre 4 (cf. § 4.2.2, Figure 4.5). On observe que la relation entre les champs de vent SAR et les champs reconstruits (étoiles noires) s'approche beaucoup de l'identité (ligne rouge), par rapport à celle entre les champs SAR et les données ECMWF (cercles bleus).



(a) *Point 1*

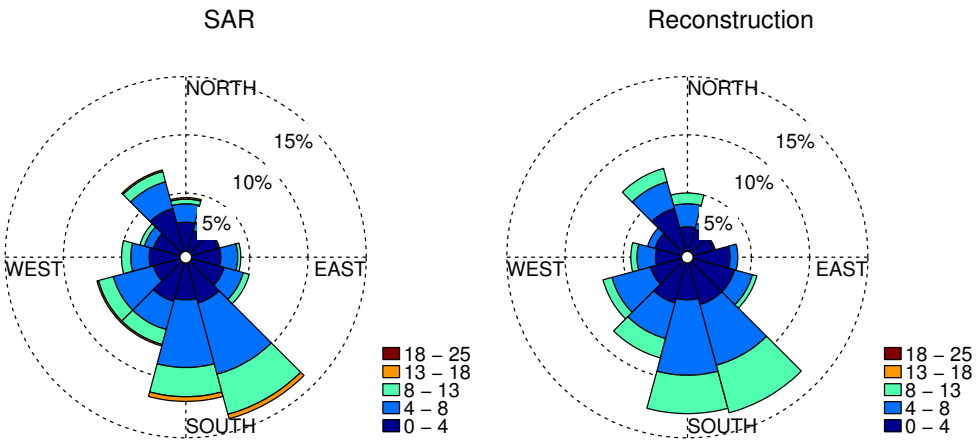


(b) *Point 3*

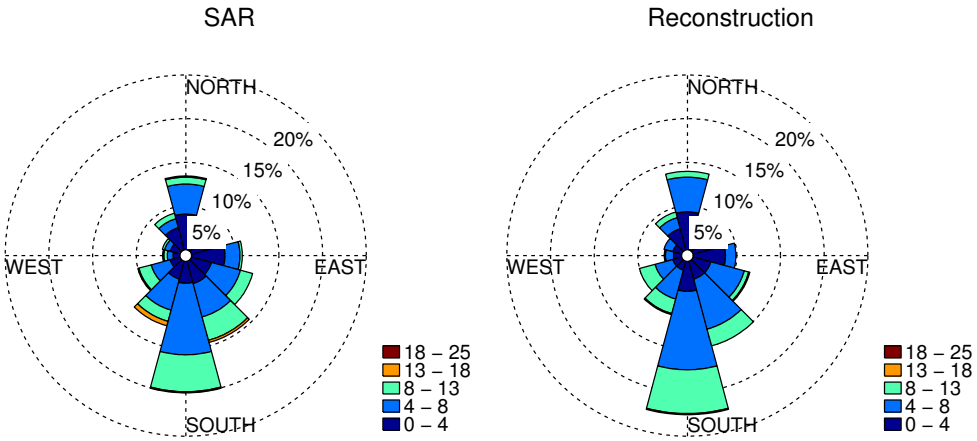


(c) *Point 4*

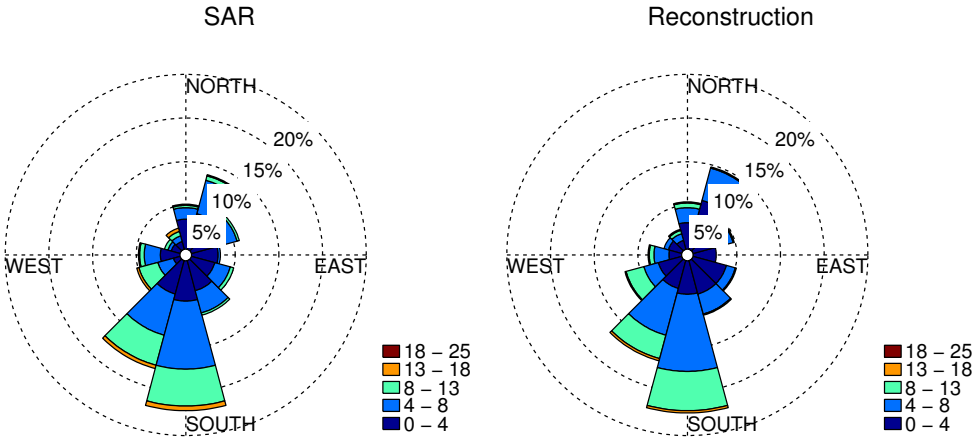
FIGURE 6.16 – *Zones fjords.*



(d) *Point 5*

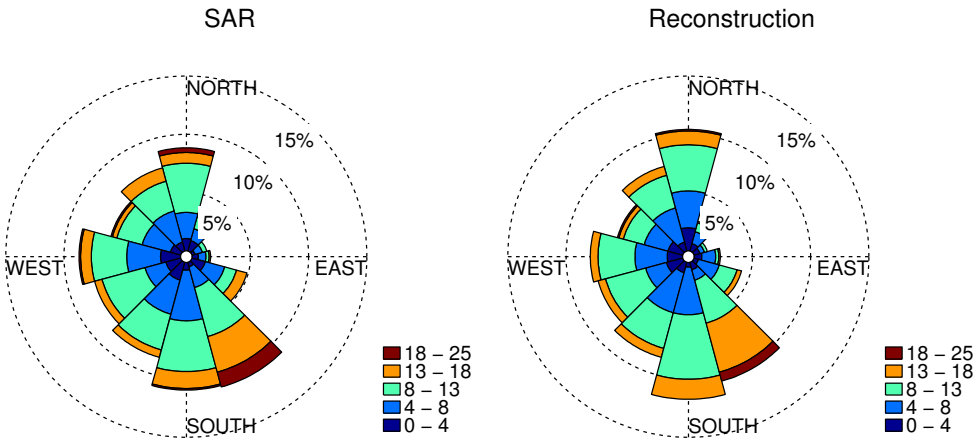


(e) *Point 7*

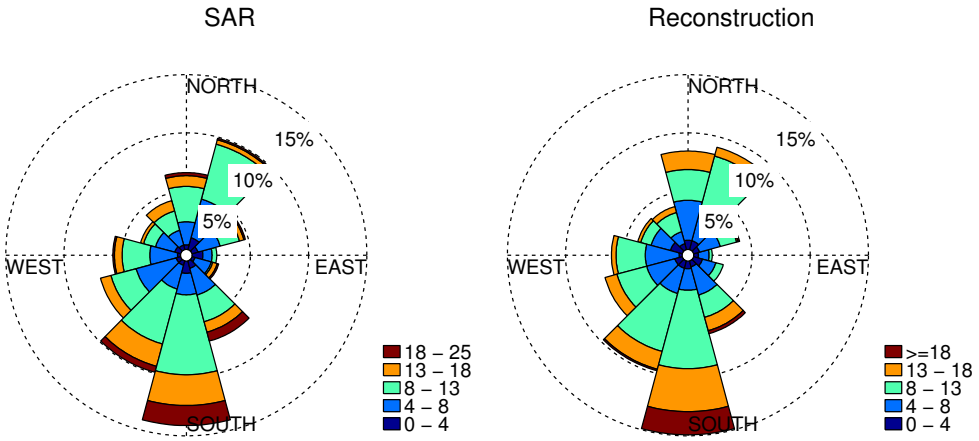


(f) *Point 8*

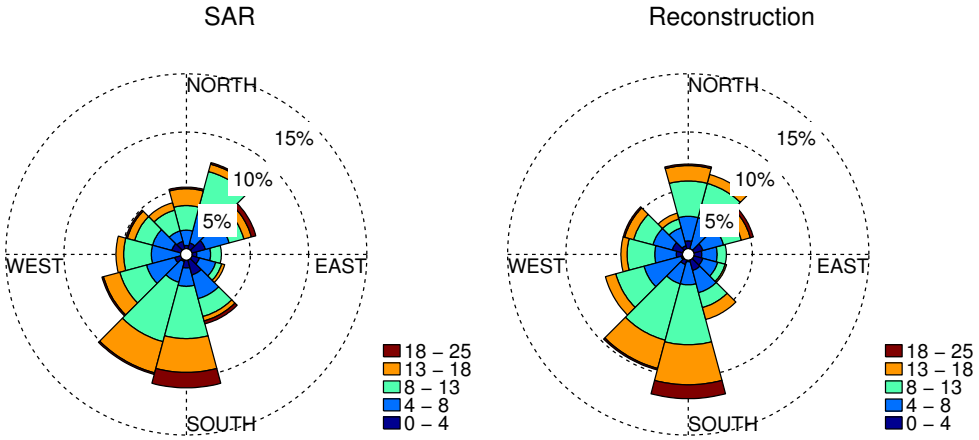
FIGURE 6.16 – *Zones côtières.*



(g) *Point 10*

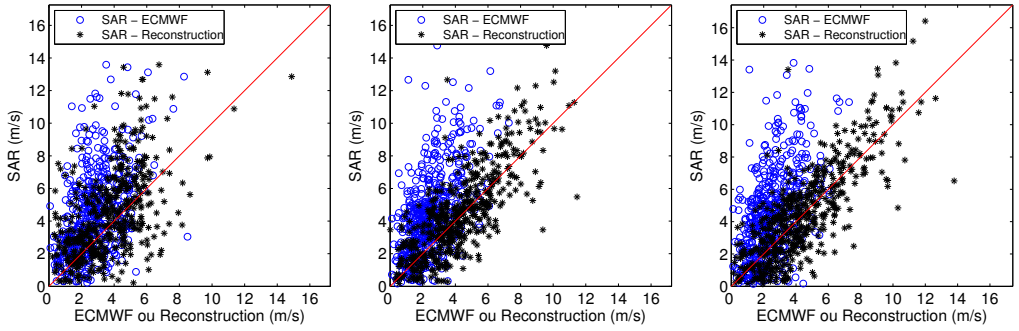


(h) *Point 11*



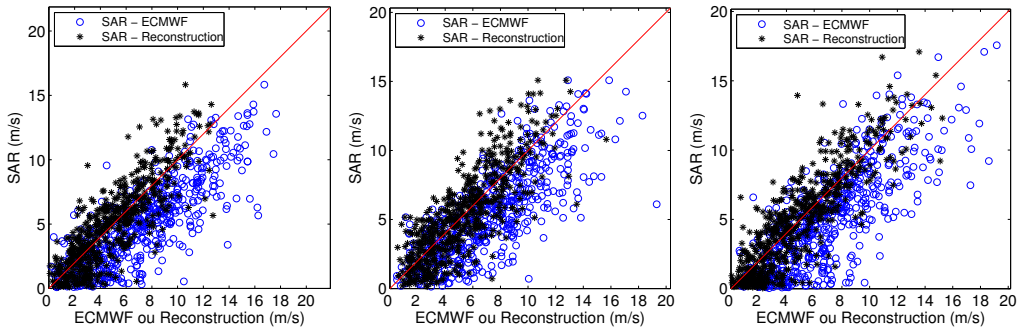
(i) *Point 12*

FIGURE 6.16 – *Au large.*
Roses des vents du SAR (à gauche) et de l'émulation (à droite).



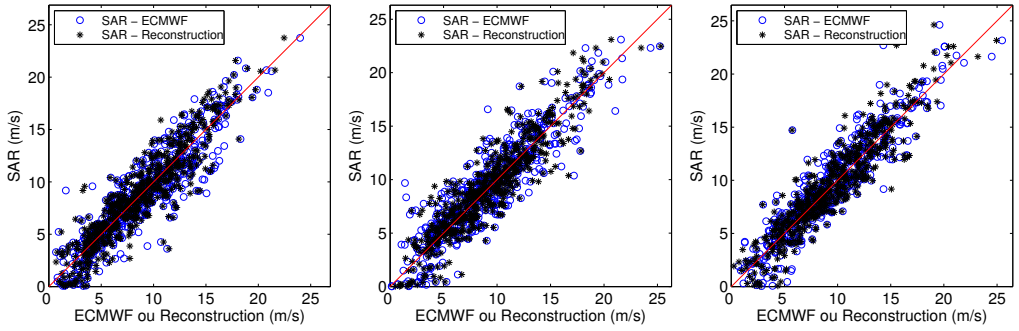
(a) *Point 1* (b) *Point 3* (c) *Point 4*

FIGURE 6.17 – *Zones fjords.*



(d) *Point 5* (e) *Point 7* (f) *Point 8*

FIGURE 6.17 – *Zones côtières.*



(g) *Point 10* (h) *Point 11* (i) *Point 12*

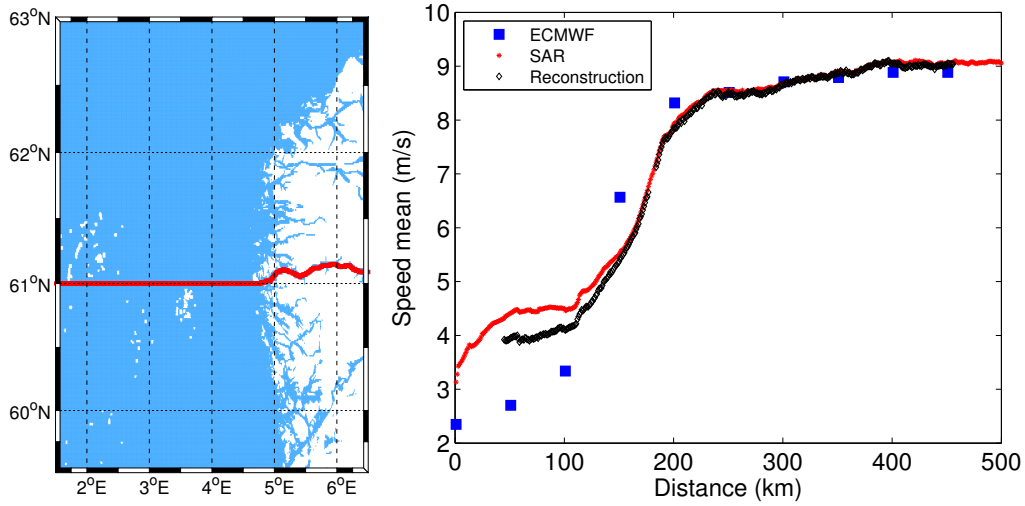
FIGURE 6.17 – *Au large.*

Diagramme de dispersion de la vitesse des vents entre les données SAR et les données ECMWF (cercles bleus) et entre les données SAR et les émulations (étoiles noires). L'axe des abscisses représente les vitesses des vents SAR (en $\mathrm{m\,s^{-1}}$*) et l'axe des ordonnées représente les vitesses de vent ECMWF ou de l'émulation (en* $\mathrm{m\,s^{-1}}$*).*

6.4.3 Comportement du *fjord* vers le large

La figure 6.18 montre le comportement des vents de l'ECMWF (carrés bleus), du SAR (étoiles rouges) et de l'émulation (diamants noirs) en fonction de la distance du *fjord* vers l'océan à partir du point de départ (*N*61.09°, *E*6.50°) vers l'ouest. La zone de validation est illustrée par la figure 6.18a. On

observe que la moyennes globale des vents reconstruits est très proche des vents observés par SAR.



(a) *Plan perpendiculaire à la côte* (b) *Distance*

FIGURE 6.18 – *La zone du fjord vers le large (a, la ligne rouge) et comportement global de la moyenne des vitesses des vents de l'ECMWF (carrés bleus), du SAR (étoiles rouges) et de l'émulation (diamants noirs) en* $\mathrm{m\,s^{-1}}$ *en fonction de la distance à partir du point de départ (N61.09°, E6.50°) vers l'ouest.*

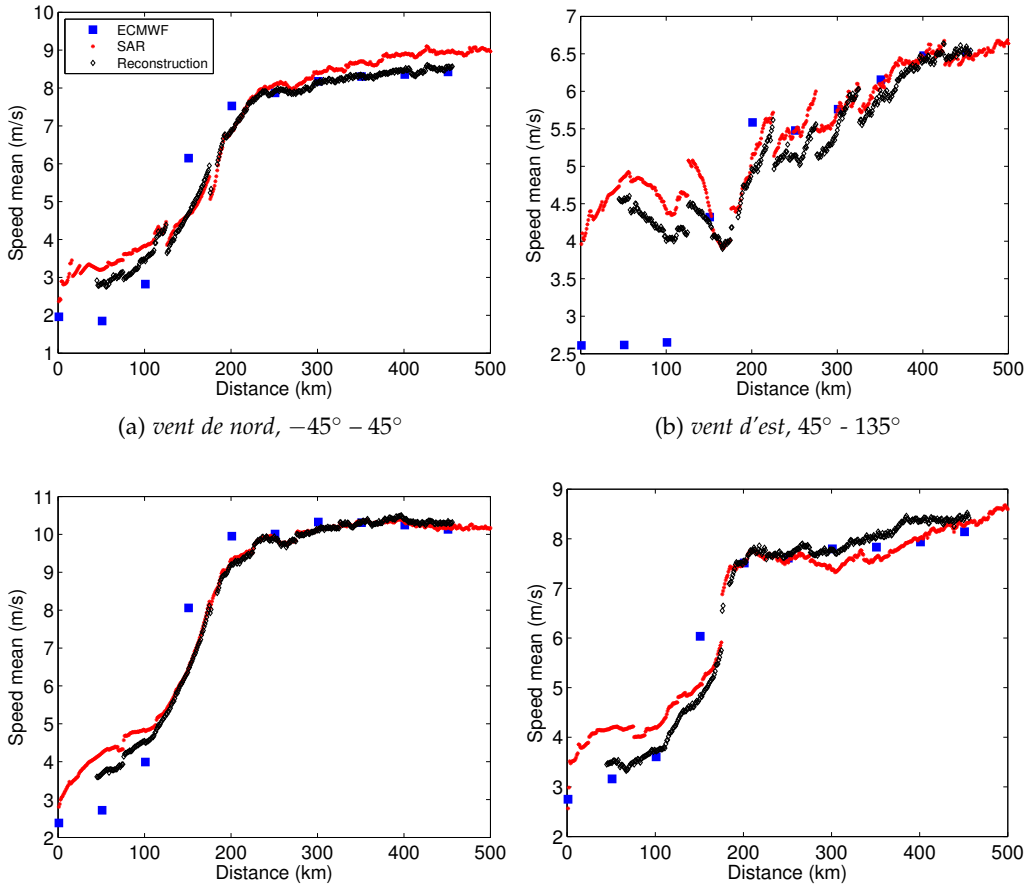
La figure 6.19 montre le comportement des vents par direction. Pour les vents de nord, d'est et de sud, la différence de la moyenne entre les champs reconstruits et les champs SAR est plus petite que celle entre les données ECMWF et les champs SAR. C'est moins le cas pour le vent d'ouest. Pour le vent d'est, l'effet de fluctuation est partiellement reconstruit. La fréquence d'oscillation est très proche de celle d'observation, mais la moyenne de l'intensité de vent reconstruit est plus petite que celle du vent observé par SAR.

6.4.4 Exemple de champs de vent

Dans cette partie, on donne quelques exemples d'émulation en utilisant les modèles basés sur la régression SVR avec le type d'information non-locale.

Les biais et les RMSE pour les composantes u et v, l'intensité et la direction pour les différents champs émulés, du 18 juillet 2009, du 16 février 2010, du 17 septembre 2005, du 5 février 2006, du 11 octobre 2006 et du 27 octobre 2007 (cas 1 à 6) sont donnés au tableau 6.1. Les résultats montrent que les biais et les RMSE restent très petits.

Les illustrations de la figure 6.20a à 6.22 montrent que les champs émulés sont très comparables à ceux observés par SAR. Pour les exemples du 17 septembre 2005 (cf. Figure 6.20c), du 5 février 2006 (cf. Figure 6.20) et du 11 octobre 2006 (cf. Figure 6.21), les vents ECMWF sont quasiment nuls dans les *fjords* et les champs reconstruits indiquent des vents plus forts (autour de $7\,\mathrm{m\,s^{-1}}$).



(a) *vent de nord*, $-45° - 45°$ (b) *vent d'est*, 45° - 135°

(c) *vent de sud*, $135° - 225°$ (d) *vent d'ouest*, $225° - 315°$

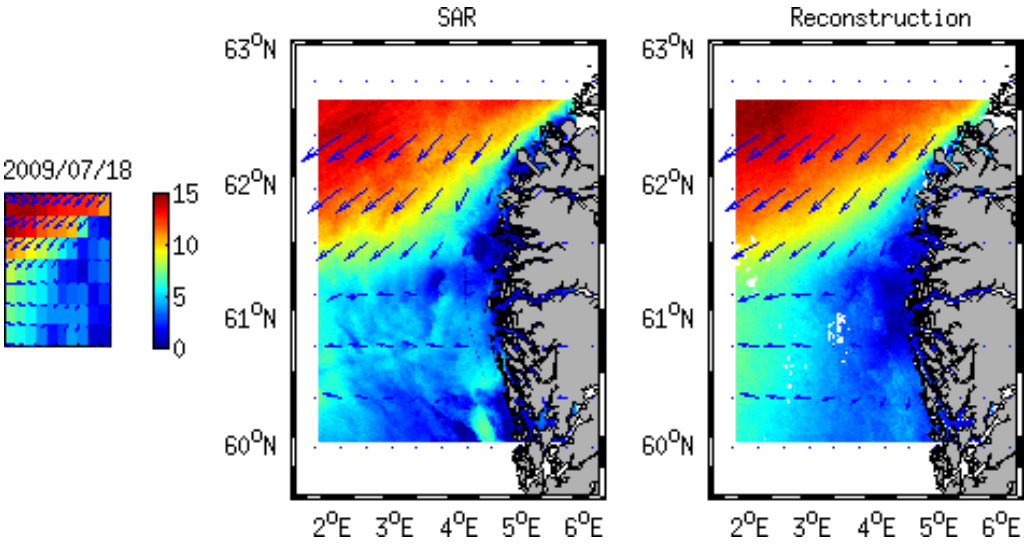
FIGURE 6.19 – *Comportement des vents de l'ECMWF, du SAR et de l'émulation en fonction de la distance pour 4 directions.*

6.5 CONCLUSION

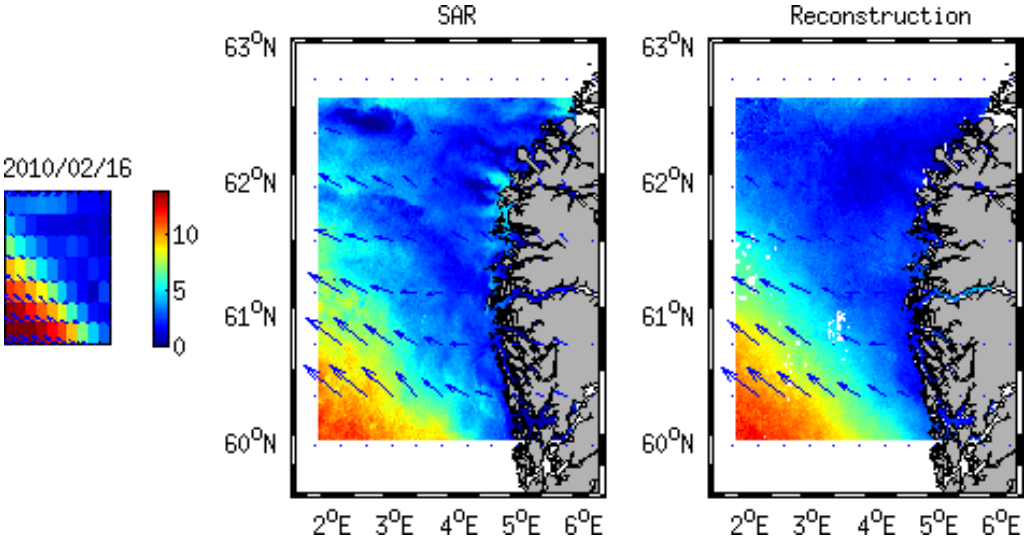
Ce chapitre évalue les différents modèles de régression à l'aide de paramètres statistiques. Les évaluations menées montrent que les informations locales et non-locales ont des performances assez proches, et donnent de meilleurs résultats que les informations globales. Au niveau des méthodes de régression, la méthode SVR a une capacité prédictive meilleure que toutes les autres méthodes : analogues, MLR et modèles basés sur la classification. Elle est beaucoup moins sensible au nombre de variables explicatives utilisées, et donc plus robuste, par rapport à la méthode de MLR par exemple. Les modèles basés sur la méthode SVR sont utilisés pour l'émulation de l'ensemble des champs de vent dans le catalogue. On observe que les champs de vent émulés restent très comparables aux observations SAR. Les caractéristiques statistiques comme la distribution de direction et d'intensité, la moyenne, la variance de composante u et v sont bien reproduites.

Tableau 6.1 – *Biais et RMSE pour les composantes u et v, l'intensité et la direction pour les différents champs émulés. Les cas 1 à 6 correspondent aux champs émulés du 18 juillet 2009, du 16 février 2010, du 17 septembre 2005, du 5 février 2006, du 11 octobre 2006 et du 27 octobre 2007 respectivement.*

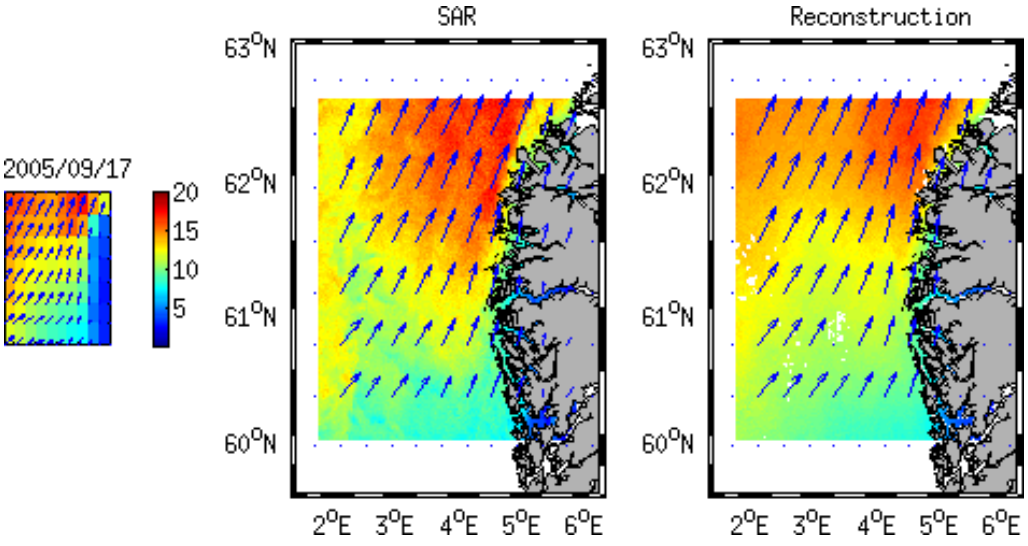
	Cas 1	Cas 2	Cas 3	Cas 4	Cas 5	Cas 6
u biais ($\mathrm{m\,s^{-1}}$)	0.56	-0.08	0.10	-0.65	0.32	0.11
u RMSE ($\mathrm{m\,s^{-1}}$)	1.16	0.99	0.90	1.20	1.10	0.73
v biais ($\mathrm{m\,s^{-1}}$)	0.13	0.08	-0.06	-0.10	-0.11	-0.08
v RMSE ($\mathrm{m\,s^{-1}}$)	0.63	0.51	0.73	0.60	0.41	0.94
Intensité biais ($\mathrm{m\,s^{-1}}$)	-0.34	0.16	-0.03	-0.65	-0.27	-0.03
Intensité RMSE ($\mathrm{m\,s^{-1}}$)	1.15	1.00	1.00	1.24	1.01	1.05
Direction biais (degré)	-2.41	-0.32	0.68	0.26	1.32	0.51
Direction RMSE (degré)	17.46	11.03	2.98	4.72	5.35	3.16



(a) *Émulation du 18 juillet 2009.*



(b) *Émulation du 16 février 2010.*



(c) *Émulation du 17 septembre 2005.*

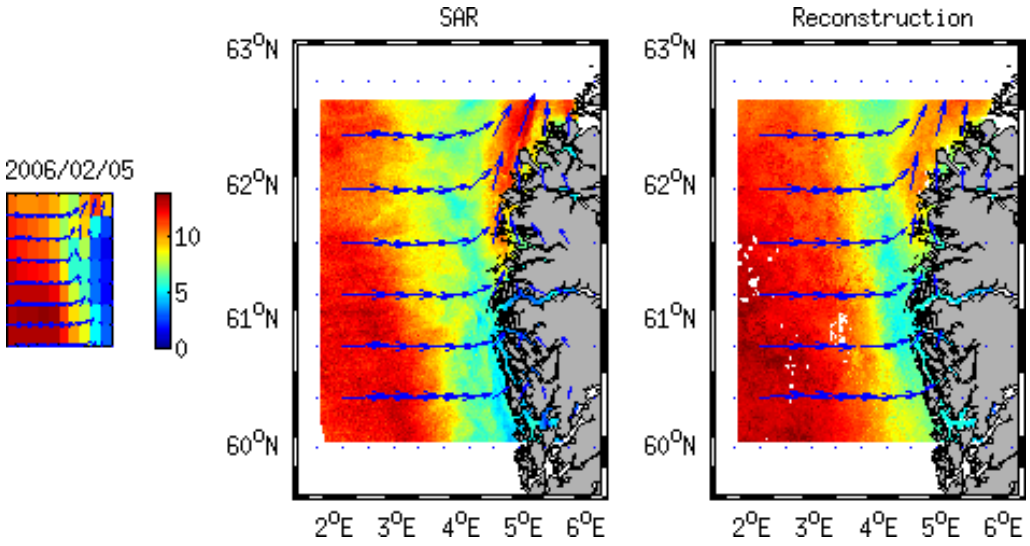


FIGURE 6.20 – *Exemple de l'émulation du 5 février 2006.*

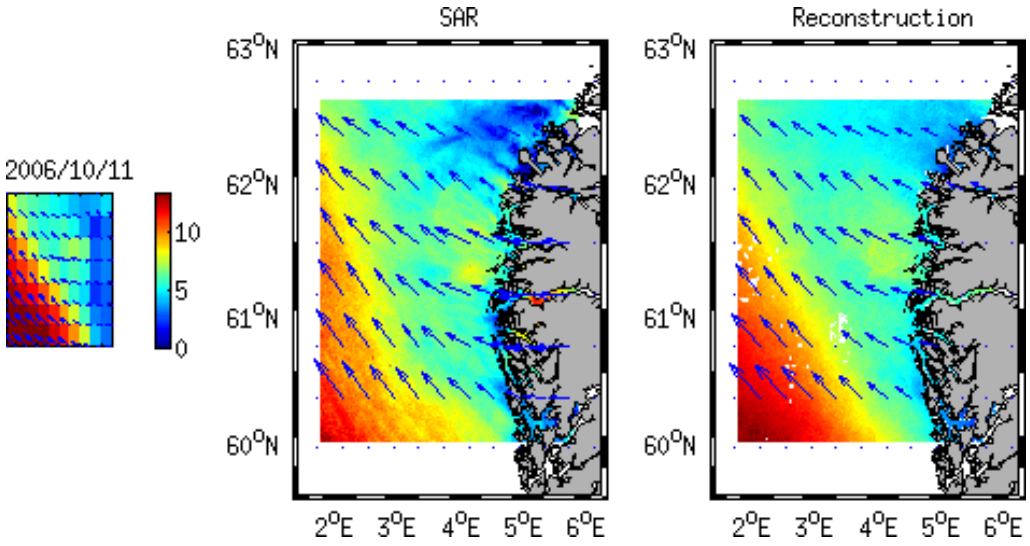


FIGURE 6.21 – *Exemple de l'émulation du 11 octobre 2006.*

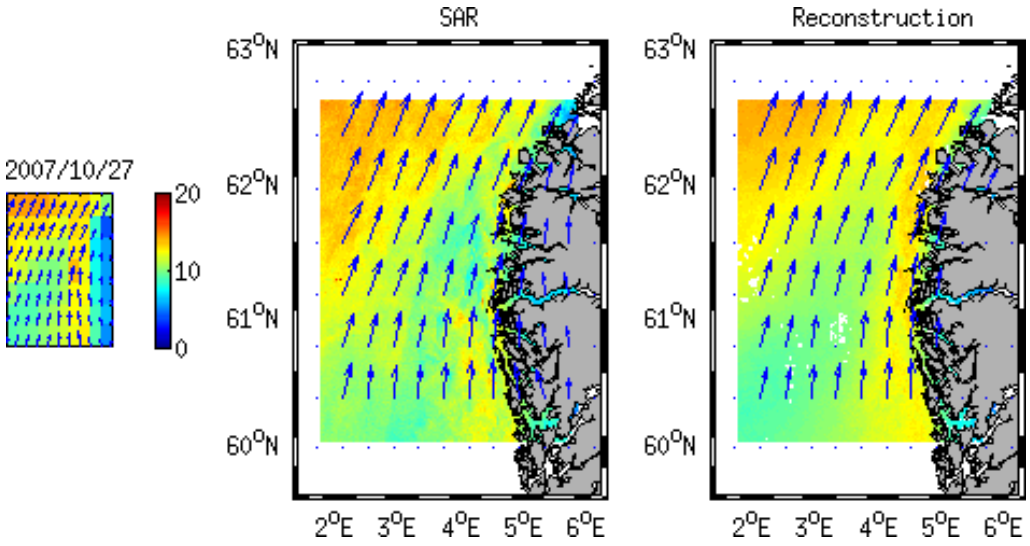


FIGURE 6.22 – *Exemple de l'émulation du 27 octobre 2007.*

Conclusions et perspectives

7

Le besoin de données à très haute-résolution spatio-temporelle pour les paramètres océaniques ne cesse d'augmenter. Ni les observations satellite ni les sorties des modèles numériques ne fournissent directement des données à très HR spatio-temporelle. Cette thèse propose des méthodes statistiques comme alternative pour obtenir les données des champs de vent à haute résolution spatio-temporelle. Le principe de ces techniques est d'émuler les champs de vent à HR SAR, à partir des données à BR spatiale mais à HR temporelle, issues des sorties du modèle numérique ECMWF, et de « catalogues » qui consistent en des couples de données historiques co-localisées à basse et à haute résolution spatiale. Dans cette thèse, la HR spatiale est modélisée directement comme une fonction de transfert appliquée à la BR. Les modèles de la fonction de transfert sont calibrés à l'aide de données historiques. Pour une nouvelle donnée BR disponible, l'application de la fonction de transfert apprise fournit un champ HR émulé à la résolution spatiale du champ SAR (0.1° en longitude et en latitude).

Pour établir la relation entre la basse et la haute résolution, nous avons proposé une architecture de modèle qui consiste à spécifier en chaque point sur la grille à HR une fonction de transfert, au lieu d'une seule fonction de transfert pour tous les points dans la zone d'étude. Ce schéma prend en compte les résultats des analyses spatiales conjointes entre les données de vent ECMWF et SAR. Les analyses montrent que les caractéristiques de la distribution de la variabilité à HR sont très locales et que la variabilité à HR est beaucoup plus élevée en zones côtières et *fjords* qu'au large. Compte tenu du fait que la relation entre les champs ECMWF et SAR apparaît moins linéaire en zones côtières et *fjords*, les modèles de régression linéaire sont distingués de ceux de régression non-linéaire.

Dans un soucis de généralisation, les modèles considérés sont formulés comme des modèles de régression du champ HR à partir de variables explicatives issues du champ BR. Pour cela, deux éléments sont distingués :

Les variables explicatives Trois types d'informations sont utilisés et comparés : les informations globales qui utilisent les M premières composantes principales pour représenter la situation globale de toute la zone d'étude ; les informations locales qui utilisent les variables aux points locaux à l'intérieur d'une fenêtre carrée centrée sur le point émulé ; les

informations non-locales qui utilisent les champs à BR aux K premiers points parmi tous les points de la grille à BR qui ont les entropies conditionnelles les moins élevées sur la situation à HR au point émulé.

Les méthodes de régression les méthodes analogues et la méthode Régression Linéaire Multiple (MLR) sont introduites et comparées avec une approche plus récente, plus complexe et plus robuste : la méthode de régression non-linéaire de type Machine à Vecteurs de support pour la Régression (SVR). Nous reformulons ces approches de façon unifiée à l'aide d'une fonction de noyau pour mieux les distinguer. La méthode SVR diffère de la méthode MLR par le type de noyau utilisé et par le choix des données de référence, et elle diffère des méthodes analogues classiques par la calibration de ses coefficients de régression. Elle est préférée aux autres types de régression en raison de son optimisation du choix des données de référence et des coefficients de régression, et de sa meilleure capacité de généralisation.

Un modèle est composé d'un type d'information et d'une méthode de régression. Les évaluations expérimentales qui ont été menées pour les différents modèles sur une série de points caractéristiques montrent que :

- le nombre de variables explicatives a une influence sur la performance prédictive des modèles. Les méthodes analogues et la méthode MLR sont beaucoup plus sensibles au nombre de variables que la méthode SVR. Globalement, pour la zone d'étude, un ensemble de 9 points, ce qui correspond pour les informations locales à une fenêtre locale d'environ $150\,\mathrm{km} \times 150\,\mathrm{km}$ fournit les erreurs d'émulation les plus petites ;
- les informations globales ont des erreurs de construction beaucoup plus élevées par rapport aux informations locales et non-locales. On peut avoir un gain de $0.5\,\mathrm{m\,s^{-1}}$ par endroits en moyenne pour les modèles de MLR et SVR et de plus que $1.0\,\mathrm{m\,s^{-1}}$ pour les modèles analogues. Entre les informations locales et non-locales, les différences sont petites ;
- les modèles de régression utilisant la méthode SVR ayant des erreurs de prédiction autour de $1.7\,\mathrm{m\,s^{-1}}$ donnent de meilleurs résultats, ensuite viennent, par ordre décroissant en terme de performance, la méthode MLR, la méthode analogue par somme pondérée et enfin la méthode du plus proche voisin ;
- exploitant une étape préalable de classification de la situation basse-résolution, les régressions multi-modèles n'améliorent pas la performance prédictive des modèles.

Le modèle proposé, basé sur la méthode SVR, a permis de reconstruire des champs très comparables aux observations SAR pour l'ensemble des zones d'étude. Les roses des vents montrent que la dépendance des deux composantes vent u et v est bien reconstituée, même si celles-ci sont émulées séparément. Les erreurs d'émulation sont inférieures à environ $1.7\,\mathrm{m\,s^{-1}}$ pour tous les points, sauf dans l'un des *fjords*. Les modèles développés sont particulièrement

intéressants en zones côtiers où les biais, pourtant relativement importants, sont bien corrigés.

La contribution principale de cette thèse est donc d'avoir développé des méthodologies originales pour les émulations de vent à HR, en associant les données récentes et une méthodologie d'apprentissage avancée. Les champs de vent émulés montrent que les modèles développés peuvent offrir une alternative pour obtenir les données de vent à haute résolution spatio-temporelle et on peut envisager un produit opérationnel permettant l'émulation (à la volée) de champs de vent HR à partir des données ECMWF.

Les données de vent à HR SAR sont très volumineuses, mais grâce à un traitement distribué sur un *cluster* de calcul, toutes les émulations de cette thèse ont pu être menées à bien.

Perspectives

Dans cette thèse, les variables explicatives utilisées se restreignent aux champs de vent à BR. Il semble que l'utilisation des autres types de paramètre comme le gradient de température, la température de l'air, *etc*. peuvent améliorer les résultats de reconstruction [Walmsley et al. (2001)]. Des informations jusqu'à la plus grande échelle peuvent également être ajoutées comme variables explicatives pour représenter les situations globales.

Seule la variabilité HR spatiale est prise en compte dans l'émulation des champs de vent à HR. L'annexe B montre que les vents ont des variabilités mensuelles et saisonnières marquées. Des analyses plus approfondies sur la différence entre les champs ECMWF et SAR en fonction de la saison et du mois peuvent être envisagées pour étudier la nécessité de prendre en compte la variabilité temporelle dans l'émulateur.

Dans cette étude, nous avons choisi une zone côtière. Dans les zones plus éloignées de la côte où la distribution de la variabilité spatiale à HR peut être identique d'un point à l'autre, les modèles par point spécifique ne sont plus nécessaires. Les émulations par des modèles globaux peuvent être proposées et comparées avec celles reposant sur les modèles par point spécifique. Si les modèles globaux sont suffisants pour l'émulation à HR *offshore*, il reste néanmoins à évaluer la distance de la côte à partir de laquelle les modèles par point spécifique ne sont plus aussi pertinents. Pour l'émulation sur une zone plus grande et globale, il est donc possible de combiner les deux types de modèle. Dans une optique de simplification des modèles, il serait intéressant de proposer des modèles permettant de combiner des modèles globaux et locaux.

Les modèles proposés dans cette thèse ne prennent en compte que la composante déterministe de la relation entre basse et haute résolution (une fois le modèle calibré, le champ de vent pour une situation à BR spécifique est unique). Or, les dynamiques géophysiques comprennent également une composante stochastique. L'extension des modèles proposés pour prendre

en compte cette composante stochastique et son conditionnement par des informations BR paraît être un axe de recherche très intéressant. Différentes approches complémentaires peuvent être envisagées, notamment l'introduction d'effets aléatoires en modélisant les paramètres des modèles de régression (coefficients, paramètres des noyaux) comme des variables aléatoires, et/ou le développement de modèles aléatoires dont la composante moyenne serait donnée par les modèles déterministes proposés dans cette thèse.

La méthodologie proposée pour l'émulation des champs de vent à HR est générique. Elle peut être appliquée à d'autres types de paramètres atmosphériques et océaniques, comme la température de l'eau, la couleur de l'eau, la précipitation *etc*. Elle peut également être appliquée pour un emboîtement de résolution. Au lieu d'utiliser les données à deux résolutions spatiales différentes (BR et HR), ce qui est le cas de cette thèse, les données issues de différents capteurs ou de modèles numériques à N_r types de résolutions différentes seraient utilisées. Les modèles de régression seraient alors calibrés entre deux résolutions successives pour effectuer le changement d'échelle en douceur.

Une perspectives pour les applications des données émulées à HR serait de les utiliser comme données d'entrée pour forcer les modèles numériques à plus haute résolution. Les sorties des modèles peuvent ensuite être comparées avec celles simulées par d'autres moyens.

Acronymes

ACP Analyse en Composantes Principales

ALADIN Aire Limitée, Adaptation dynamique, Développement InterNational

AMS American Meteorological Society

ARPEGE Action de Recherche de Petite Échelle Grande Échelle

AROME Application de la Recherche à l'Opérationnel à Méso-Échelle

ASAR Advanced SAR

CART Classification and Regression Trees

CERSAT Centre ERS d'Archivage et de Traitement

CLS Collecte Localisation Satellites

BR Basse Résolution

HR Haute Résolution

ECMWF European Center for Medium-range Weather Forecast

ENVISAT ENVIronmental SATellite

EOF Fonctions Orthogonales Empiriques

ESA Agence Spatiale Européenne

GCM General Circulation Models

GMF Geophysical Model Function

GPS Global Positioning System

LOS Laboratoire d'Océanographie Spatiale

OLS Ordinary Least Square

PNT Prévision Numérique du Temps

QuikSCAT Quick Scatterometer

NASA National Aeronautics and Space Administration

NDBC National Data Buoy Center

NCEP National Center for Environmental Prediction

NESDIS National Environmental Satellite, Data, and Information Service

NOAA National Oceanic and Atmospheric Administration

NWP Numerical Weather Prediction

RAR Real Aperture Radar

RBF Radial Basis Function

RCM Regional Climate Model

RFE Recursive Feature Elimination

RMSE Root-Mean-Square Error

SAR Radar à Synthèse d'Ouverture

SVM Machine à Vecteurs de Support

SVR Machine à Vecteurs de support pour la Régression

SVs Vectors de Support

LR Régression Linéair

MLR Régression Linéaire Multiple

VC Vapnik-Chervonenkis

VHF Very High Frequency

Bibliographie

Debasish Basak, Srimanta Pal, et Dipak Chandra Patranabis. Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10) :203–224, 2007. (Cité pages 17 et 75.)

Phillipe Beaucage, Anna Glazer, Julien Choisnard, Wei Yu, Monique Bernier, Robert Benoit, et Gaëtan Lafrance. Wind assessment in a coastal environment using synthetic aperture radar satellite imagery and a numerical weather prediction model. *Canadian Journal of Remote Sensing*, pages 368–377, 2007. (Cité page 3.)

Mohamed Bassam Ben Ticha. *Fusion de données satellitaires pour la cartographie du potentiel éolien offshore*. PhD thesis, Ecole des Mines de Paris, 2007. (Cité pages 11, 14, 59 et 81.)

Rasmus E. Benestad, Inger Hanssen-Bauer, et Deliang Chen. *Empirical-statistical downscaling*. World Scientific Pub Co Inc, 2008. (Cité page 10.)

Kristin P. Bennett et O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1 (1) :23–34, 1992. (Cité page 76.)

Philippe Besse et Béatrice Laurent. *Apprentissage Statistique : prévision et data mining*. 2012. (Cité pages 63 et 78.)

Chabi Biaou, Angelbert. *De la méso-échelle à la micro-échelle : désagrégation spatio-temporelle multifractale des précipitations*. PhD thesis, École Nationale Supérieure des Mines de Paris, Décembre 2004. (Cité page 9.)

H Björnsson et SA Venegas. A manual for eof and svd analyses of climatic data. *CCGCR Report*, 97(1), 1997. (Cité page 64.)

Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001. (Cité page 11.)

Christopher S. Bretherton, Catherine Smith, et John M. Wallace. An intercomparison of methods for finding coupled patterns in climate data. *Journal of climate*, 5(6) :541–560, 1992. (Cité page 64.)

Christophe Cassou, Marie Minvielle, Laurent Terray, et Claire Périgaud. A statistical–dynamical scheme for reconstructing ocean forcing in the atlantic.

part I : weather regimes as predictors for ocean surface variables. *Climate Dynamics*, 36(1) :19–39, 2011. (Cité page 10.)

Chih-Chung Chang et Chih-Jen Lin. LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2 :1–27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. (Cité page 87.)

CS Cheng, G. Li, Q. Li, et H. Auld. Statistical downscaling of hourly and daily climate scenarios for various meteorological variables in south-central canada. *Theoretical and Applied Climatology*, 91(1) :129–147, 2008. (Cité page 10.)

Vladimir Cherkassky et Yunqian Ma. Selection of meta-parameters for support vector regression. Dans *Artificial Neural Networks*, ICANN, pages 687–693. Springer, 2002. (Cité page 79.)

Vladimir Cherkassky et Yunqian Ma. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, 17(1) :113–126, 2004. (Cité page 79.)

Vladimir Cherkassky et Filip M Mulier. *Learning from data : concepts, theory, and methods*. Wiley.com, 2007. (Cité page 79.)

J. H. Christensen, B. Hewitson, A. Busuioc, A. Chen, X. Gao, R. Held, R. Jones, R. K. Kolli, WK Kwon, R. Laprise, et al. Regional climate projections. Dans *Climate Change, 2007 – The Physical Science Basis : Working group I Contribution to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Chapitre 11, pages 847–940. Cambridge University Press, 2007. (Cité page 10.)

J Corte-Real, B Quian, et H Xu. Circulation patterns, daily precipitation in portugal and implications for climate change simulated by the second hadley centre gcm. *Climate Dynamics*, 15(12) :921–935, 1999. (Cité page 10.)

John C Curlander et Robert N McDonough. *Synthetic aperture radar*, volume 199. Wiley New York, 1991. (Cité page 25.)

Knut-Frode Dagestad, Hansen Morten W., Johannessen Johnny A., et al. Development and validation of a sar wind emulator. Rapport technique, Nansen Environmental and Remote Sensing Center, 2009. (Cité page 40.)

Gérard Dreyfus, Jean-Marc Martinez, Manuel Samuelides, Mirta B Gordon, Fouad Badran, et Sylvie Thiria. *Apprentissage statistique : Réseaux de neurones-Cartes topologiques-Machines à vecteurs supports*. Editions Eyrolles, 2011. (Cité pages 11 et 16.)

George M Furnival et Robert W Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4) :499–511, 1974. (Cité page 61.)

K. Goubanova, V. Echevin, B. Dewitte, F. Codron, K. Takahashi, P. Terray, et M. Vrac. Statistical downscaling of sea-surface wind over the peru–chile upwelling region : diagnosing the impact of climate change from the IPSL-CM4 model. *Climate Dynamics*, pages 1–14, 2010. ISSN 0930-7575. (Cité pages 4, 5, 10, 11, 12, 16 et 64.)

Yann Guermeur et Hélène Paugam-Moisy. Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines. *Apprentissage automatique*, pages 109–138, 1999. (Cité pages 17, 77 et 78.)

Isabelle Guyon et André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3 :1157–1182, 2003. (Cité page 63.)

Isabelle Guyon, Jason Weston, Stephen Barnhill, et Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3) :389–422, 2002. (Cité page 61.)

Kevin Hamilton. Comprehensive meteorological modelling of the middle atmosphere : A tutorial review. *Journal of Atmospheric and Terrestrial physics*, 58(14) :1591–1627, 1996. (Cité page 22.)

Trevor Hastie, Robert Tibshirani, et Jerome Friedman. *The elements of statistical learning*, volume 2. Springer, 2009. (Cité pages 63 et 64.)

H Hersbach, A Stoffelen, et S De Haan. An improved C-band scatterometer ocean geophysical model function : CMOD5. *Journal of Geophysical Research : Oceans (1978–2012)*, 112(C3), 2007. (Cité page 24.)

Radan Huth. Statistical downscaling in central europe : Evaluation of methods and potential predictors. *Climate Research*, 13(2) :91–101, 1999. (Cité page 10.)

Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005. (Cité page 64.)

Richard W Katz. Use of conditional stochastic models to generate climate change scenarios. *Climatic Change*, 32(3) :237–255, 1996. (Cité page 10.)

V. Kecman. *Learning and soft computing : support vector machines, neural networks, and fuzzy logic models*. MIT press, 2001. (Cité pages 10 et 77.)

Claude Kergomard. *La télédetection aéro-spatiale*. École Nationale Supérieure Paris, 2013. (Cité page 24.)

John W Kidson et Craig S Thompson. A comparison of statistical and model-based downscaling techniques for estimating local climate variations. *Journal of Climate*, 11(4), 1998. (Cité page 10.)

Ron Kohavi et George H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1) :273–324, 1997. (Cité page 61.)

Clifford Lau. *Neural networks : theoretical foundations and analysis*. IEEE press, 1991. (Cité page 11.)

Jean-Yves Le Vourc'h, Claude Fons, et Marcel Le Stum. *Météorologie générale et maritime*. École nationale de la météorologie, Météo France, 2001. (Cité page 26.)

M. G. H Ligda. radar storm observation. pages 1265–1282, 1951. (Cité page 20.)

Portabella Marcos. *Wind field retrieval from satellite radar systems*. PhD thesis, University of Barcelona, 2002. (Cité page 24.)

Armel Martin. *Influence des ondes de gravité de montagne sur l'écoulement de grande échelle en présence de niveaux critiques*. PhD thesis, Paris VI, Paris, France, 2008. (Cité page 55.)

Eric Martin, B Timbal, et E Brun. Downscaling of general circulation model outputs : simulation of the snow climatology of the french alps and sensitivity to climate change. *Climate Dynamics*, 13(1) :45–56, 1996. (Cité page 11.)

René Mayençon. *Météorologie marine*. Éditions maritimes et d'outre-mer, Avril 1982. (Cité pages 5, 27, 46 et 52.)

SW McCandless et Christopher R Jackson. Principles of synthetic aperture radar. *SAR Marine User's Manual*, pages 1–23, 2004. (Cité pages 25, 127 et 128.)

Marie Minvielle. *Méthode de désagrégation statistico-dynamique adaptée aux forçages atmosphériques pour la modélisation de l'océan atlantique : développement, validation et application au climat présent*. PhD thesis, Sciences de l'Univers, de l'environnement et de l'espace, 2009. (Cité pages 4, 11, 12, 14, 16, 81 et 82.)

Frank Monaldo, Vincent Kerbaol, Pablo Clemente-Colón, B. Furevik, J. Horstmann, J. Johannessen, X. Li, W. Pichel, T. D. Sikora, D. J. Thomson, et al. The SAR measurement of ocean surface winds : an overview. Dans *Proceedings of the Second Workshop Coastal and Marine Applications of SAR*, pages 2–12, 2003. (Cité pages 3, 22 et 36.)

Isidoro Orlanski. A rational subdivision of scales for atmospheric processes. 56 (5) :527–530, Mai 1975. (Cité pages 19 et 20.)

S. C. Pryor et R. J. Barthelmie. Analysis of the effect of the coastal discontinuity on near-surface flow. Dans *Annales Geophysicae*, volume 16, pages 882–888. Springer, 1998. (Cité page 27.)

W. C. De Rooy et K. Kok. A combined physical-statistical approach for the downscaling of model wind speed. *Weather and forecasting*, 19(3) :485–495, 2004. (Cité page 10.)

Bernhard Schölkopf et Alexander J. Smola. *Learning with kernels : Support vector machines, regularization, optimization, and beyond*. MIT press, 2001. (Cité pages 5, 16, 17, 75 et 78.)

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 :379–423, 623–656, 1948. (Cité page 66.)

Merrill Skolnik. *Radar Handbook, Third Edition*. McGraw-Hill Education, 2008. ISBN : 9780071485470. (Cité page 127.)

Alex J. Smola et Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3) :199–222, 2004. (Cité pages 75 et 76.)

Scott Spak, Tracey Holloway, Barry Lynn, et Richard Goldberg. A comparison of statistical and dynamical downscaling for surface temperature in north america. *Journal of Geophysical Research*, 112, 2007. (Cité page 9.)

Ad Stoffelen et David Anderson. Scatterometer data interpretation : Estimation and validation of the transfer function CMOD4. *Journal of Geophysical Research*, 102(C3) :5767–5780, 1997. (Cité page 25.)

B Timbal et BJ McAvaney. An analogue-based method to downscale surface air temperature : application for australia. *Climate Dynamics*, 17(12) :947–963, 2001. (Cité page 11.)

Vladimir Vapnik et Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2) :264–280, 1971. (Cité page 75.)

Vladimir Vapnik, Steven E Golowich, et Alex Smola. Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, pages 281–287, 1997. (Cité pages 5, 11, 16 et 17.)

R. Vautard. Multiple weather regimes over the north atlantic analysis of precursors and successors. *Monthly weather review*, 118(10) :2056–2081, 1990. (Cité page 10.)

Hans von Storch, Bruce Hewitson, et Linda Mearns. Review of empirical downscaling techniques. Dans *RegClim Spring Meeting*, 2000. (Cité page 9.)

John Walmsley, Rebecca Barthelmie, et William Burrows. The statistical prediction of offshore winds from land-based data for wind-energy applications. *Boundary-Layer Meteorology*, 101 :409–433, 2001. (Cité pages 4, 10, 11, 14, 16, 27, 45, 81, 82 et 117.)

Wenjian Wang, Zongben Xu, Weizhen Lu, et Xiaoyun Zhang. Determination of the spread parameter in the gaussian kernel for classification and regression. *Neurocomputing*, 55(3) :643–663, 2003. (Cité page 80.)

R. L. Wilby, S. P. Charles, E. Zorita, B. Timbal, P. Whetton, et L. O. Mearns. Guidelines for use of climate scenarios developed from statistical downscaling methods. *IPCC Task Group on Data and Scenario Support for Impact and Climate Analysis*, Août 2004. (Cité pages 9 et 10.)

R. L. Wilby et T. M. L. Wigley. Downscaling general circulation model output : a review of methods and limitations. *Progress in Physical Geography*, 21(4) : 530, 1997. (Cité pages 9 et 10.)

Daniel S Wilks. Multisite downscaling of daily precipitation with a stochastic weather generator. *Climate Research*, 11(2) :125–136, 1999. (Cité page 10.)

Eduardo Zorita, James P. Hughes, Dennis P. Lettemaier, et Hans von Storch. Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *Journal of Climate*, 8(5) :1023–1042, 1995. (Cité pages 10 et 11.)

Eduardo Zorita et Hans von Storch. The analog method as a simple statistical downscaling technique : comparison with more complicated methods. *Journal of Climate*, 12(8) :2474–2489, 1999. (Cité pages 4, 5, 11, 16, 74 et 75.)

Résolution spatiale d'image SAR

A

La résolution d'un radar est définie comme sa capacité à distinguer deux cibles très proches l'une de l'autre. Elle est traditionnellement divisée en deux parties : la résolution en distance et la résolution en azimut Skolnik (2008). Considérons une image radar composée d'un ensemble de valeurs $A(x, y)$, où x correspond à l'axe de la direction de mouvement d'un satellite et y est la direction d'éclairage du radar. La résolution dans la direction y est la résolution en distance et celle dans la direction x est la résolution en azimut (figure A.1).

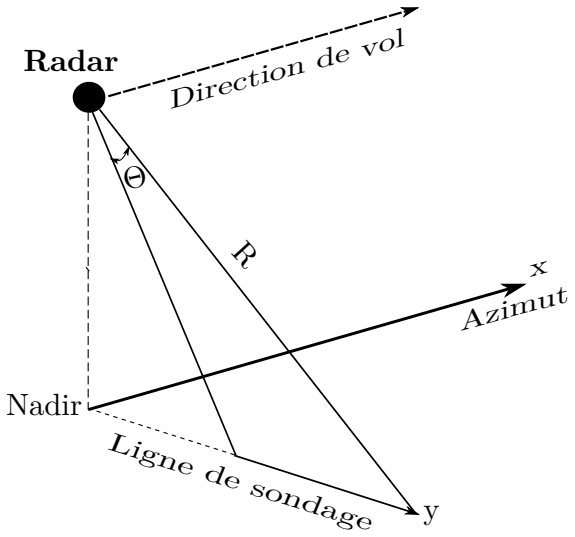


Figure A.1 – *Géométrie du système Radar aéroporté*

La résolution en distance est la capacité d'un système radar à distinguer les différents cibles dans la direction d'éclairage du radar mais à des distances différentes. Elle dépend de la largeur de l'impulsion émise, du type et de la taille des cibles et de l'efficacité du récepteur. Théoriquement elle peut être calculée grâce à la formule suivante McCandless et Jackson (2004) :

$$r_d \geq \frac{c_0 \tau}{2} \tag{A.1}$$

où τ est le largeur d'une impulsion et c_0 est la vitesse de la lumière. Une interprétation de la formule A.1 est que les signaux réfléchis par deux cibles

qui ont une distance plus petite que la distance parcourue pendant le temps $\frac{\tau}{2}$ par les impulsions commencent à se superposer à la réception du radar. Il n'est plus possible de distinguer les deux retours au site radar. Selon la formule A.1, pour augmenter la résolution en distance, il faut diminuer la largeur d'une impulsion.

La résolution en azimut s'appelle aussi résolution angulaire. La résolution angulaire est l'écart angulaire minimum pour que le radar soit capable de distinguer deux objets identiques qui se présentent à la même distance. Quand deux objets identiques se présentent dans le même lobe du radar à la même distance, leurs réflexions vont retourner vers le radar en même temps. Pour un système de Real Aperture Radar (RAR), la résolution en azimut est déterminée par la largeur de lobe principale d'antenne Θ (figure A.2) McCandless et Jackson (2004) :

$$S_a \geq 2R \sin \frac{\Theta}{2} \tag{A.2}$$

où R est la distance entre le radar et l'objet. Plus le lobe est étroit, plus la directivité de l'antenne est importante. La largeur de lobe d'antenne est elle-même proportionnée à la longueur d'onde des impulsions transmises et inversement proportionnée à la taille de l'antenne.

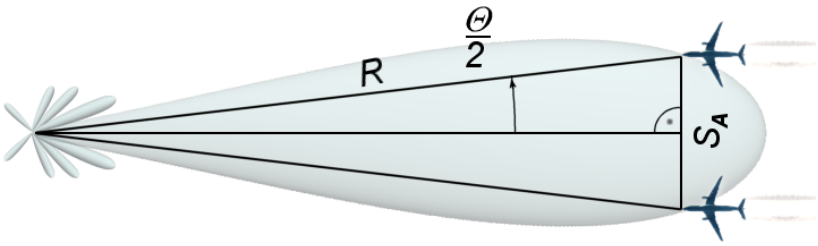


FIGURE A.2 – *Résolution angulaire d'un système RAR, Charly Whisky*©.

Pour augmenter la résolution en azimut d'un système RAR, 3 solutions sont possibles : diminuer la distance radar-objet (R) ; diminuer la longueur d'onde des impulsions, ou augmenter la longueur de l'antenne. Néanmoins, aucune de ces 3 solutions n'est facilement réalisable dans l'espace.

Le Système de radar à synthèse d'ouverture (SAR) consiste en la combinaison des mesures radar et de techniques de traitement d'image pour « synthétiser » le longueur de l'antenne, contrairement au système RAR qui augmente physiquement le longueur de l'antenne pour raffiner la résolution en azimut. SAR utilise de multiples sondages successifs, décalés dans le temps et l'espace, pour obtenir un sondage composite. L'antenne du radar, relativement petite, donne du sol un signal qui est la résultante, en amplitude et phase, de tous les échos générés par tous les points éclairés par l'impulsion émise : c'est l'intégrale (au sens mathématique du terme) de l'espace éclairé. Le signal reçu est donc un point de la transformée de Fourier du sol éclairé. Comme le radar se déplace avec son porteur, avion ou satellite, il reçoit d'autres points de

cette transformée. Il suffit d'enregistrer tous ces points et d'en faire ensuite la transformée inverse pour reconstituer le relief en deux dimensions du sol.

Finalement, la résolution en azimut pour un système SAR est proportionnée au longueur de l'antenne :

$$S_a = \frac{D_{\mathrm{AT}}}{2} \tag{A.3}$$

Contrairement au système RAR, plus la longueur de l'antenne est petite, plus l'ouverture du radar est grande, plus la résolution en azimut est fine.

Enfin, les résolutions en distance et angulaire conduisent à la notion de cellule de résolution spatiale. Le sens de cette cellule est très clair : il est impossible de distinguer deux cibles se trouvant à l'intérieur d'une même cellule de résolution.

Analyse temporelle

B

Six années de données de vent recueillies sur la station Troll sont utilisées pour tracer les roses des vents mensuelles. Les données peuvent être recueillies dans la base de donnée de l'Institut Météorologique Norvégien (http://eklima.met.no/). La figure B.1 montre une rose des vents globale pour la période de 2006 à 2011. Les autres roses de vents sont tracées pour chaque mois de l'année. Les résultats montrent que les vents ont des caractéristiques mensuelles très marquées.

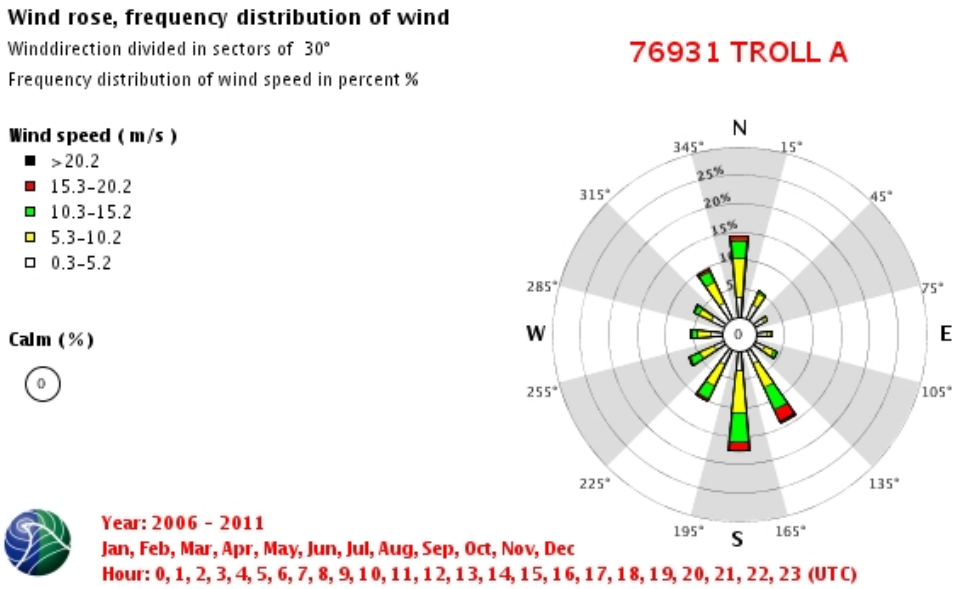
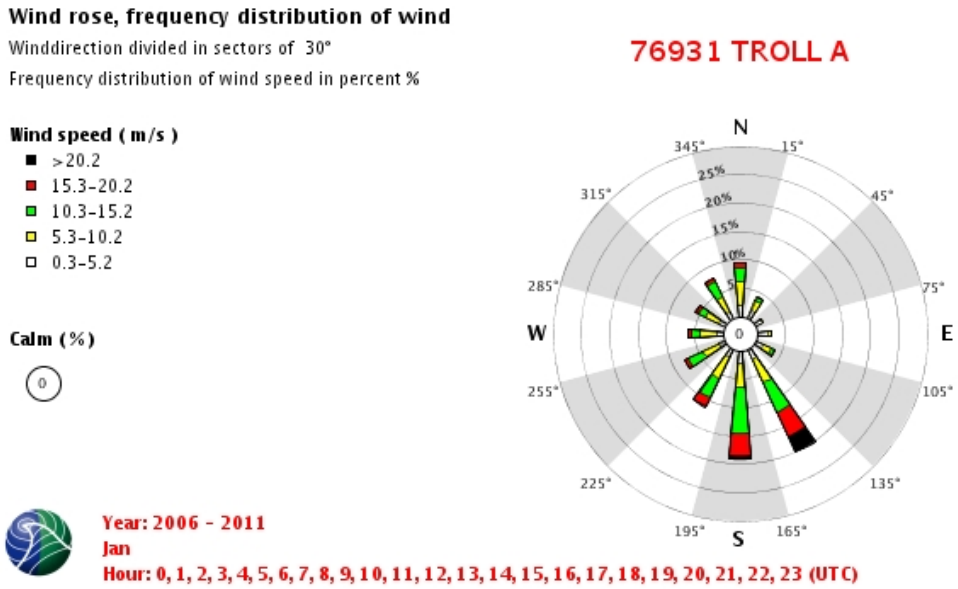
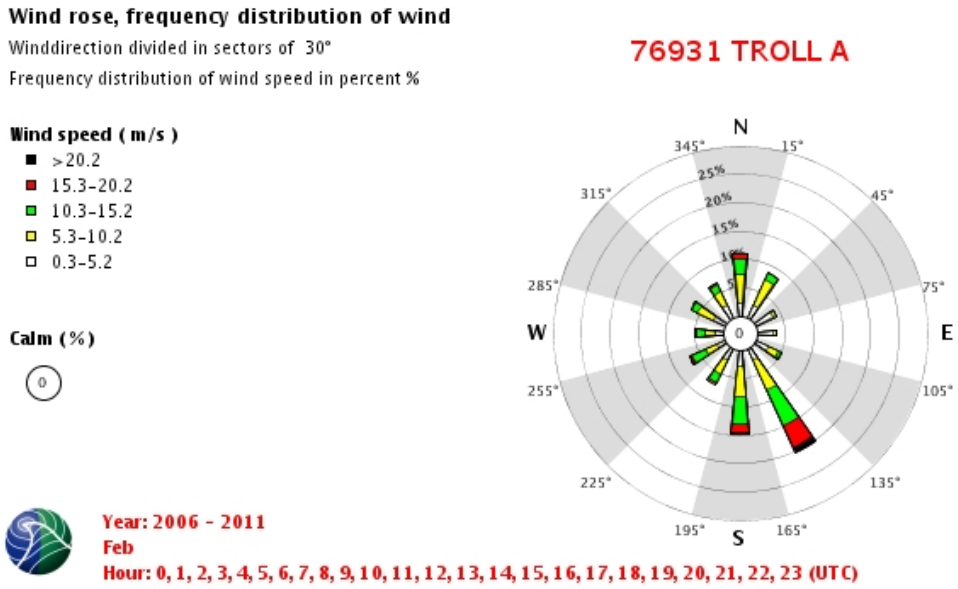


Figure B.1 – *Rose des vents globale, de 2006 à 2011, sur la station Troll (N60.60°, E3.70°), Norvège.*

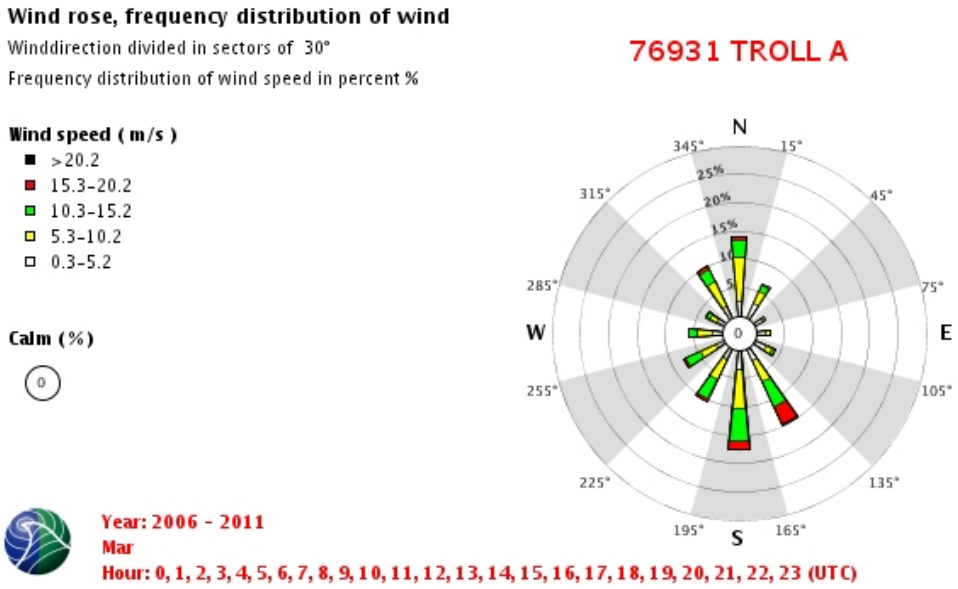


(a) *Rose des vents : janvier*

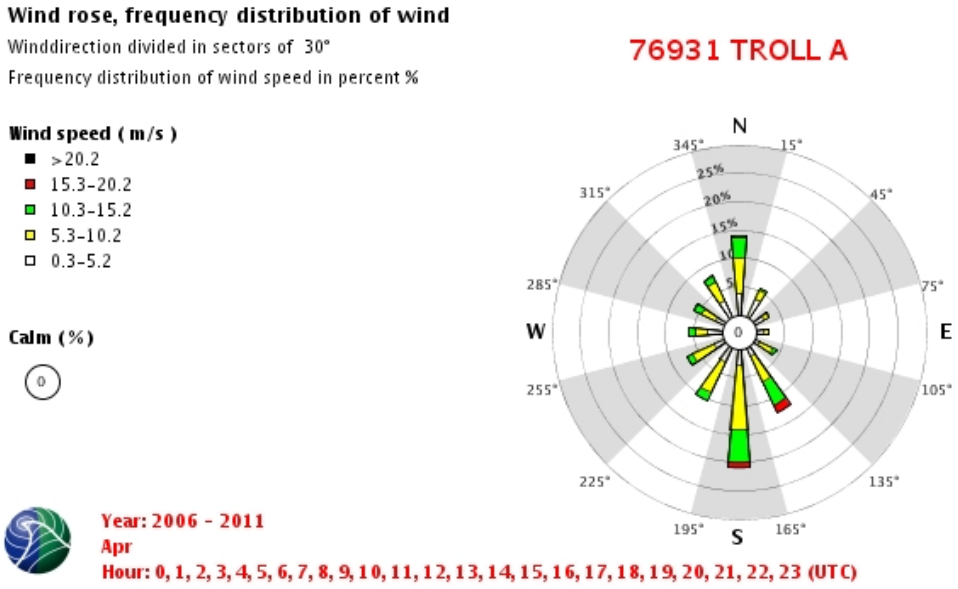


(b) *Rose des vents : février*

FIGURE B.2 – *Rose des vents du mois de janvier et du mois de février sur la station Troll.*

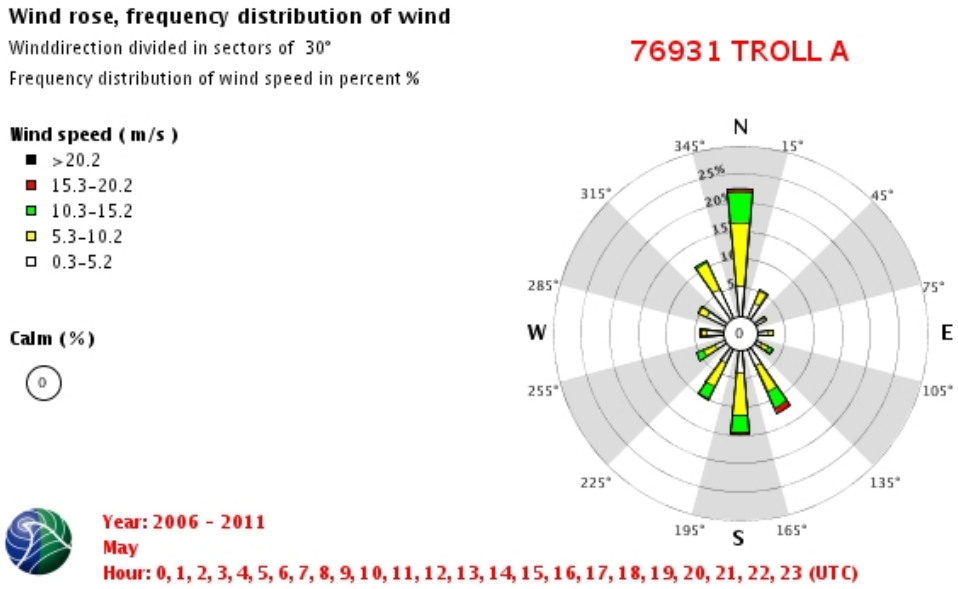


(c) *Rose des vents : mars*

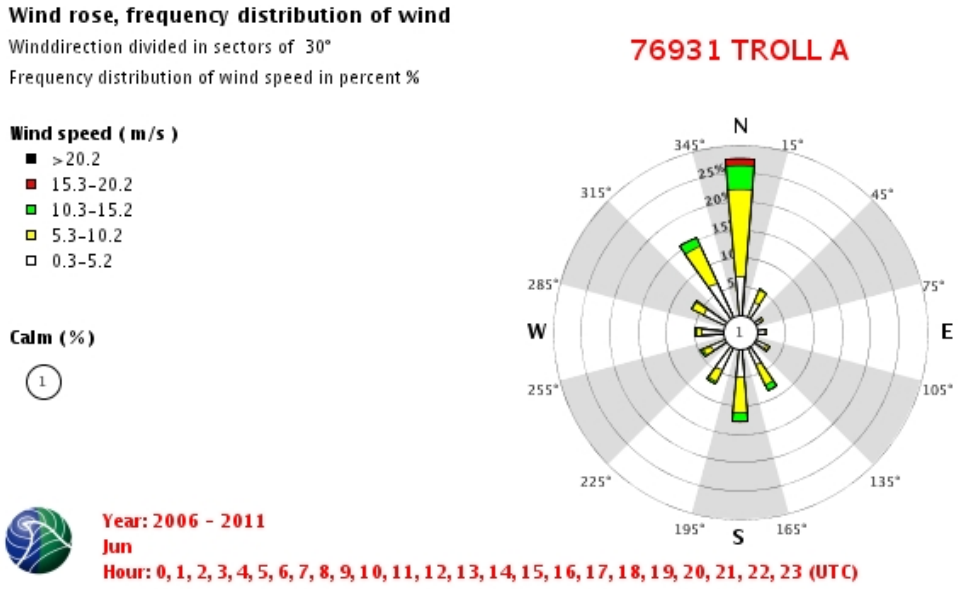


(d) *Rose des vents : avril*

FIGURE B.2 – *Rose des vents du mois de mars et du mois d'avril sur la station Troll.*

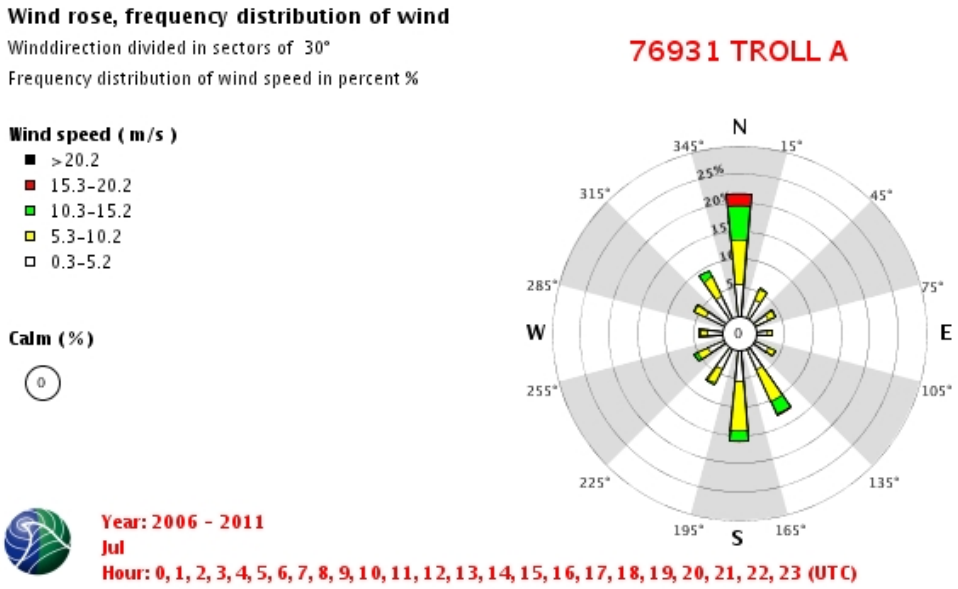


(e) *Rose des vents : mai*

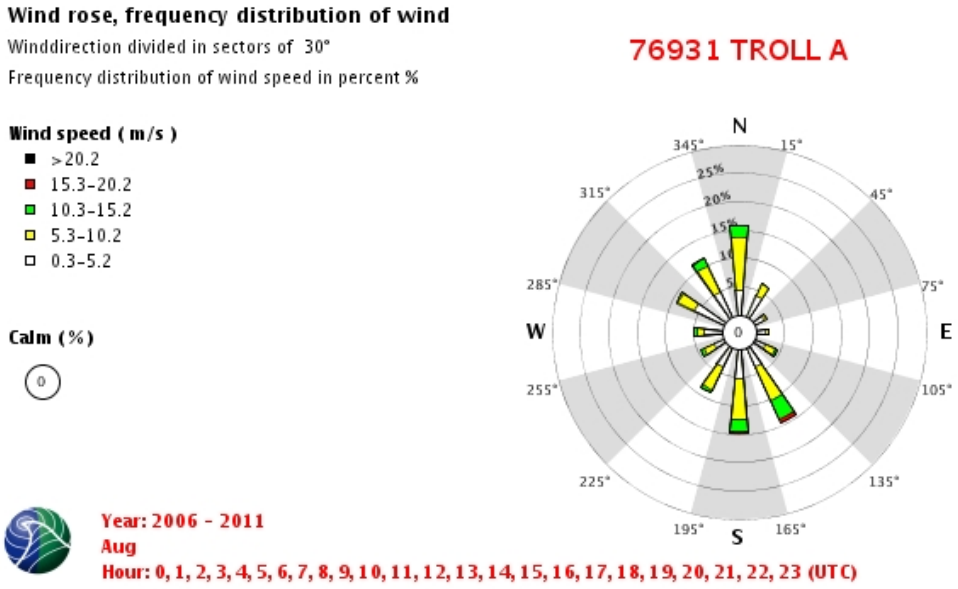


(f) *Rose des vents : juin*

FIGURE B.2 – *Rose des vents du mois mai et du mois juin sur la station Troll.*

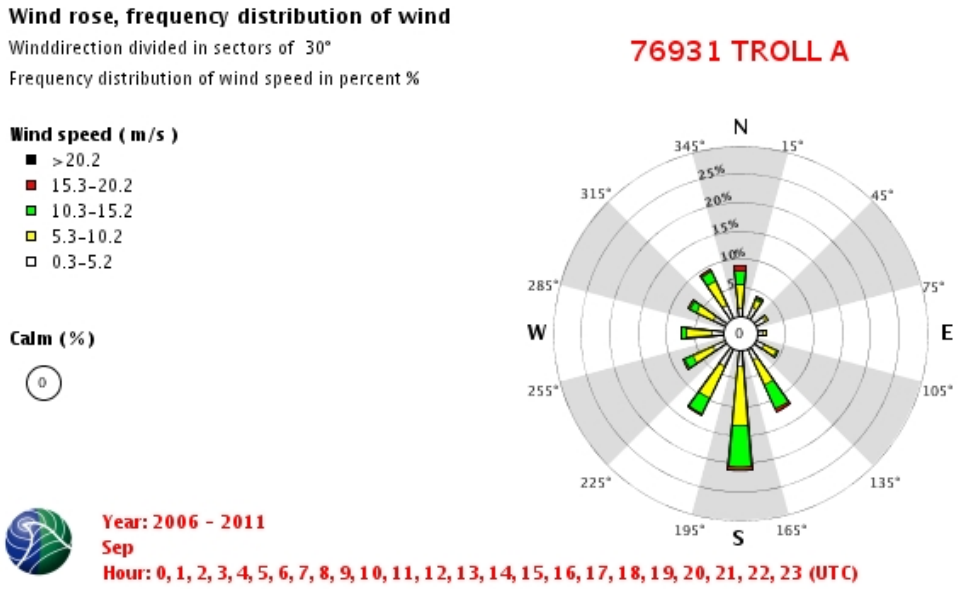


(g) *Rose des vents : juillet*

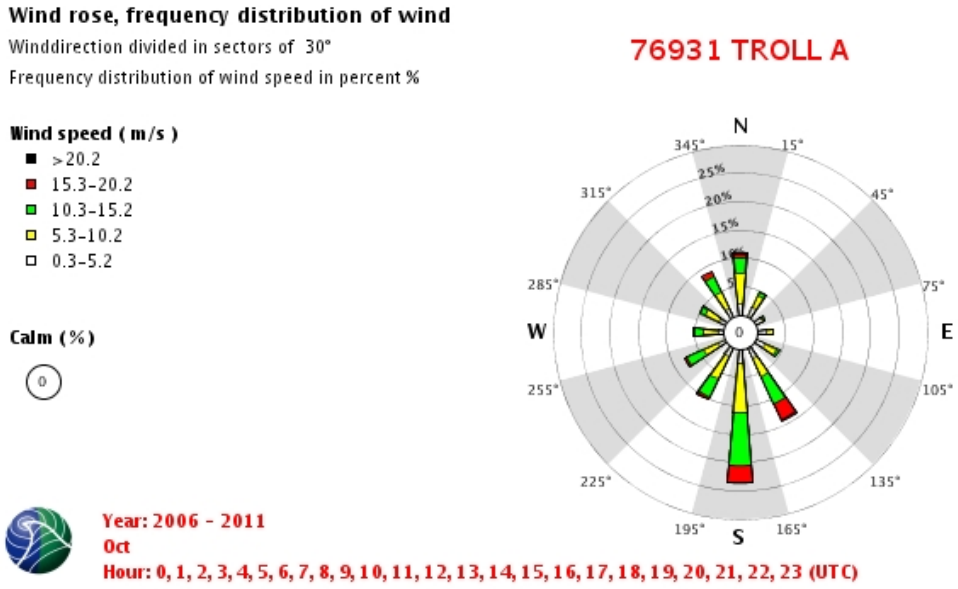


(h) *Rose des vents : août*

FIGURE B.2 – *Rose des vents du mois juillet et du mois août sur la station Troll.*

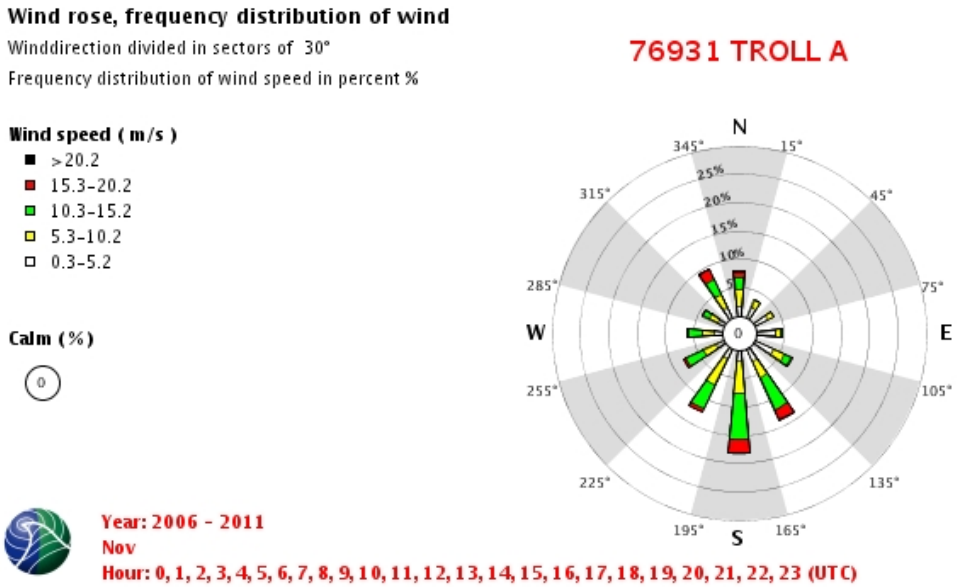


(i) *Rose des vents : septembre*

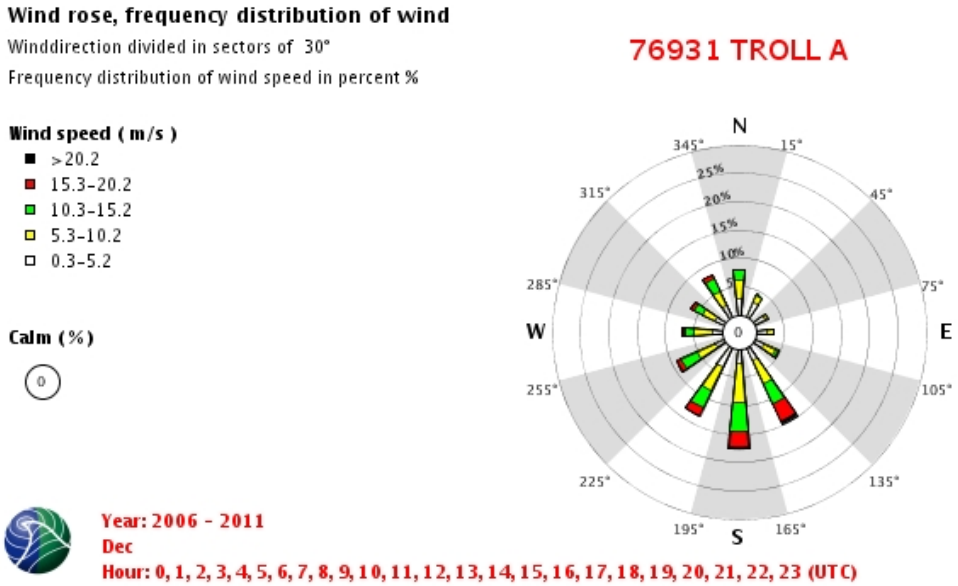


(j) *Rose des vents : octobre*

FIGURE B.2 – *Rose des vents du mois septembre et du mois octobre sur la station Troll.*



(k) *Rose des vents : novembre*



(l) *Rose des vents : décembre*

FIGURE B.2 – *Rose des vents du mois novembre et du mois décembre sur la station Troll.*

Glossaire

C

– Vent

Le vent est défini comme un déplacement de l'air. Un vent de surface à 10 m représente le mouvement horizontal de l'air à une hauteur de 10 mètres au dessus de la surface. L'information recueillie sur le vent inclut sa direction, sa vitesse et sa nature. Dans les dix premiers mètres au-dessus du sol, le vent a tendance à prendre de la vitesse et à tourner avec la hauteur.

Pour les mesures sur site, les vents sont normalement mesurés sur des sites dégagés, nivelés et les plus éloignés possible d'obstacles au déplacement de l'air tels que les arbres, les immeubles ou les collines. À la plupart des stations principales, on mesure généralement le vent en prenant la moyenne sur une ou deux minutes à chaque observation, les valeurs étant fournies par un anémomètre. À d'autres sites de mesure du vent, les valeurs peuvent être fournies par les enregistrements des anémomètres. Les périodes de moyennage peuvent varier d'une minute à une heure. Une rafale de vent extrême est la pointe de vent instantanée observée aux cadrans de l'anémomètre ou tirée d'un graphique d'enregistrement en continu.

– Direction du vent

Direction dans laquelle souffle le vent, par rapport au nord vrai ou géographique (360 degrés au compas) ; par exemple, un vent d'est souffle de l'est et non vers l'est. Elle représente la direction moyenne au cours d'une période de deux minutes, cessant à l'heure de l'observation, et est arrondie à la dizaine de degrés la plus proche ou exprimée en fonction de l'un des 16 points du compas (N, NE, WNW, etc.). Si elle est exprimée en dizaines de degrés, 9 signifie 90 degrés vrais ou un vent d'est et 36 signifie 360 degrés vrais ou un vent soufflant du pôle Nord géographique. Une valeur de zéro (0) indique un vent calme.

– Vitesse du vent

Vitesse de déplacement de l'air, exprimée en kilomètres par heure (km/h), en mètre par seconde ($m\,s^{-1}$) ou en nœuds. Généralement elle est observée à 10 mètres au-dessus du sol. Elle représente la vitesse moyenne au cours de la période de deux minutes cessant à l'heure de l'observation. Facteurs de conversion : 1 nud = $1.85\,km\,h^{-1}$ = $0.51 m\,s^{-1}$ et $1\,km\,h^{-1}$ = 0.54 nuds = $0.27 m\,s^{-1}$.

– Météorologie

État de l'atmosphère à un moment particulier. Il s'agit des variations à court terme ou instantanées de l'atmosphère, par opposition aux changements à long terme ou climatiques.

– **Climatologie** La climatologie, par opposition à la météorologie, est une science concernée par le long terme. Elle étudie l'état physique moyen de l'atmosphère — à travers la pression atmosphérique, la température, le vent, l'humidité, les précipitations, l'ensoleillement, etc. — et les variations de cet état dans le temps et l'espace. Ainsi peut-elle examiner les caractéristiques et l'évolution du climat global (moyenné sur toute la Terre), mais aussi discriminer et classifier à différentes échelles du climat divers types de climat, dont elle s'efforce alors de préciser les localisations géographiques, les fluctuations à court et à long terme et, pour finir, les causes de leur répartition dans l'espace et de leur évolution dans le temps.

– **European Center for Medium-range Weather Forecast (ECMWF)**
ECMWF est le Centre européen pour les prévisions météorologiques à moyen terme. C'est une organisation internationale financée par 25 États européens. Il a pour objectif d'élaborer des méthodes numériques de prévision météorologique de portée moyenne, la préparation régulière de prévisions météo à moyen terme à distribuer aux services météo des États membres, la recherche scientifique et technique pour l'amélioration de ces prévisions ainsi que la collecte et le stockage de données météo appropriées.

Les données ECMWF sont divisées en 3 grandes catégories : analyses, prévisions instantanées et prévisions cumulées. Les données d'analyses combinent les données de prévisions à court terme avec les observations pour produire le meilleur ajustement de deux. Les données sont disponibles plusieurs fois par jours. Les prévisions instantanées sont produites par le modèle de prévision, à partir d'une analyse. Elles sont disponibles aux différentes étapes de temps (heures) à partir de l'heure d'une analyse (notons que les prévisions ne sont pas toutes initiées par les données d'analyses). Les paramètres des prévisions cumulées sont accumulés depuis le début de la prévision. On peut diviser la valeur par la durée de prévision pour obtenir une moyenne sur la durée d'accumulation.

Liens utiles

D

Les centres de prévision météorologique modernes fournissent des prévisions de meilleure qualité que par le passé et constituent une autre source de données extrêmement utiles. Voici une liste non exhaustive de ces centres :

1. Société de Collecte Localisation Satellites
   http://www.boost-technologies.com/esa/sar_wind.html
2. European Centre for Medium-Range Weather Forecasts (ECMWF)
   http://www.ecmwf.int/
3. Institut Météorologique Norvégien
   http://eklima.met.no/
4. National Center for Environmental Prediction
   http://www.ncep.noaa.gov/
5. MetEd
   https://www.meted.ucar.edu/
6. Ocean Surface Winds Team (OSWT) of the Center for Satellite Application and Research
   http://manati.star.nesdis.noaa.gov/
7. Physical Oceanography Distributed Active Archive Center of Jet Propulsion Laboratory
   http://podaac.jpl.nasa.gov/
8. European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT)
   http://www.eumetsat.int/Home/index.htm
9. World meteorological organization
   http://www.wmo.int/

Résumé

Dans cette thèse, on s'intéresse à l'émulation de champs de vent de la surface de la mer à haute résolution. La disponibilité de ces champs à haute résolution est critique dans de nombreuses situations, par exemple la gestion du littoral, les structures offshore, le suivi de nappes de pétrole, etc. Les satellites, en particulier les systèmes à SAR (Synthetic Aperture Radar) peuvent observer la surface de mer à une résolution spatiale de quelques mètres. Les champs de vent SAR produits de manière opérationnelle atteignent une résolution spatiale de moins de 1 km. Cependant, les échantillonnages de la surface de l'océan par SAR sont irréguliers en temps en en espace pour une zone donnée. En revanche, les modèles numériques comme ECMWF (European Center for Medium-range Weather Forecast) peuvent fournir des champs de vent toutes les 6 heures, mais avec une résolution spatiale très basse (~50 km * 50 km).
Nous combinons les données ECMWF et SAR pour produire des champs de vent à haute résolution spatio-temporelle. Pour cela, l'émulation à haute résolution est formulée comme l'apprentissage statistique d'une fonction de transfert entre la basse résolution et la haute résolution. La fonction de transfert est calibrée par des couples de données historiques co-localisées à basse et à haute résolution. On utilise alors des informations locales, non-locales ou globales associées à des méthodes de régression linéaire ou non-linéaire. L'approche est validée sur une zone côtière de Norvège, caractérisée par la complexité de sa topographie.
Nous montrons que les modèles non-linéaires SVR (machine à Vecteurs de Support pour la Régression), utilisant les informations locales ou non-locales, donnent de meilleurs résultats par rapport aux modèles basés sur des méthodes plus classiques, comme analogues ou MLR (Régression Linéaire Multiple). Les champs émulés par la méthode SVR sont très proches des observations SAR. Les biais pourtant importants en zones côtières sont bien corrigés. En outre, les champs émulés préservent les propriétés statistiques des champs de vent SAR.

Mots-clés : Apprentissage supervisé, Émulation statistique, Vent côtier, Haute résolution, Machine à Vecteurs de support pour la Régression (SVR), Radar à Synthèse d'Ouverture (SAR), Prévision du modèle numérique

Abstract

This PhD thesis addresses the reconstruction of high resolution sea surface wind fields. The availability of such high resolution fields is critical for numerous issues, e.g. coastal management, offshore structures, oil spill disaster tracking, etc. Satellites, especially from Synthetic Aperture Radar (SAR) systems, can monitor the ocean surface at a spatial resolution of a few meters. SAR wind fields are operationally produced with spatial resolutions of less than 1 km. However, satellite SAR systems involve highly irregular sampling of the ocean surface and, for a given region, SAR wind fields may be delivered with a low temporal resolution, typically every 7-to-10 days for temperate zones. By contrast, numerical model predictions, such as European Center for Medium-range Weather Forecast (ECMWF) wind fields, are typically delivered with a high temporal resolution (e.g. every 6 h), but with a low spatial resolution (~50 km * 50 km).

The question of the combination of numerical model predictions and SAR wind fields naturally arises to deliver high resolution wind fields at sea surface anywhere and anytime. Here, we state this issue as the statistical learning of transfer functions between low resolution model predictions and the associated high resolution SAR fields. We investigate the extent to which such regression functions can be determined from a set of co-located low resolution and high resolution fields. Both global,
local and non-local information schemes as well as linear and non-linear regression methods are considered. As a case-study, we carry out numerical experiments for a coastal area off Norway, which involves complex low resolution to high resolution situations.
We show that machine learning models based on non-linear Support Vector Regression (SVR) method, combined either with local or non-local information, perform better than more classical models, e.g. those based on analog method or Multiple Linear Regression (MLR). The SVR based models produce high resolution variability very close to the reference local variability observed by SAR, especially in coastal area. Furthermore, the reconstructed wind fields preserve the statistical distribution properties of SAR wind fields

Keywords : Machine learning, Downscaling, Coastal wind, High-resolution, Support vector regression (SVR) Synthetic aperture radar (SAR), Numerical model predictions

TELECOM Bretagne

n° d'ordre : 2014telb0319
Télécom Bretagne
Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3
Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00