Stochastic data assimilation of the random shallow water model loads with uncertain experimental measurements

L. Mathelin, Christophe Desceliers, M.Yussuf Hussaini

Stochastic data assimilation of the random shallow water model loads with uncertain experimental measurements ⋆

L. Mathelin [a], C. Desceliers [b] & M.Y. Hussaini [c]

[a] *LIMSI - CNRS, BP 133, 91403 Orsay, France*

[b] *Université Paris Est, Laboratoire Modélisation et Simulation Multi-Echelle, MSME UMR 8208 CNRS, 77454 Marne-la-vallée, France*

[c] *Computational Science & Engineering, Department of Mathematics, Florida State University, Tallahassee, FL 32306-4510, USA*

Abstract

This paper is concerned with the estimation of a parametric probabilistic model of the random displacement source field at the origin of seaquakes in a given region. The observation of the physical effects induced by statistically independent realizations of the seaquake random process is inherent with uncertainty in the measurements and a stochastic inverse method is proposed to identify each realization of the source field. A statistical reduction is performed to drastically lower the dimension of the space in which the random field is sought and one is left with a random vector to identify. An approximation of the vector components is determined using a Polynomial Chaos decomposition, solution of an optimality system to identify an optimal representation. A second order gradient-based optimization technique is used to efficiently estimate this statistical representation of the unknown source while accounting for the non-linear constraints in the model parameters. This methodology allows the uncertainty associated with the estimates to be quantified and avoids the need for repeatedly solving the forward model.

Key words: Parametric estimation, stochastic inverse method, Polynomial Chaos, data assimilation, maximum likelihood.

⋆ Mathelin L., Desceliers C., Hussaini M.Y., Stochastic data assimilation of the random shallow water model loads with uncertain experimental measurements, Computational Mechanics, **47**(6) , 603-616 (2011)

Email addresses: mathelin@limsi.fr, christophe.desceliers@univ-paris-est.fr

1 Introduction

1.1 Scientific context

It is no surprise that statistical estimation of stochastic processes has received considerable attention from researchers in the past and still continues to be a very active field as real-world problems are stochastic in nature. It is almost always necessary to quantify a stochastic process in the sense of giving it a tractable representation for further use in subsequent analysis or modeling. Such a representation allows one to reduce the stochastic process to a finite set of parameters describing it as closely as possible in a certain statistical sense. Indeed, techniques for propagating and quantifying uncertainty in numerical simulations are now widespread and they often require the description of the corresponding boundary and/or initial conditions. Their description may come from observational data of experiments, which leads to the problem of the identification of a stochastic process properties. In the general case, this stochastic process must be described in a specific space for its use as input data in a subsequent model. Often, one cannot directly observe the random process but one can only observe the effects it induces. The problem is then to identify the underlying stochastic process (in other words, infer its statistical properties) from a set of observational data.

In this work, we are interested in deriving an approximate probabilistic description of the sea bottom displacement field caused by earthquakes and modeled as a random process $S(\boldsymbol{x}, t; \theta) \in \Omega_{\boldsymbol{x}} \times T \times \Theta$, where $\Omega_{\boldsymbol{x}}$ is the spatial domain, T is the temporal domain and Θ is the space of elementary events θ. The stochastic framework is the probability space $(\Theta, \sigma_\Theta, P_\Theta)$, where σ_Θ is a σ-algebra and P_Θ is a probability measure. The sea bottom motion field is different from event to event but the local geophysical properties cause the field to remain close to an "average" field. For instance, the amplitude and spatial extent of the motion field may vary but the global patterns remain similar because they are related to active tectonic plate locations, such as the contact line between two plates. A statistically well characterized approximation of the displacement random field can be subsequently used for simulating the uncertain future evolution of the tectonic plates or as an input parameter for inferring magma motions underneath the outer layer of the Earth. However, in the identification step of the sea bottom displacement stochastic field, the model we rely on to predict the output field (which is observed) from the input source field involves poorly known parameters and is then itself inherent with uncertainty. Further, the observation sensor on the ocean monitoring satellite is of finite accuracy and also introduces uncertainty in the measurements. Thus, we also aim to quantify the impact of the random parameters related to the physical model and sensor uncertainty on the estimate of the source field

$S(\boldsymbol{x}, t; \theta)$.

Among the usual methods, one of the most popular approaches is to rely on the maximum likelihood principle combined with an *a priori* functional form (*e.g.*, Polynomial Chaos) for the source field probabilistic description (*e.g.*, see Eggermont & LaRiccia (2001), Ghanem & Spanos (1990), Ghanem & Spanos (2003)). The probabilistic model of the source-field and all uncertain variables involved in the problem are then statistically sampled, the model is solved for the output field and the constitutive parameters of the source field are adjusted so that the likelihood of the observed data is maximized. This procedure leads to the optimal, in the likelihood sense, estimate of the source field in the chosen functional form. However, it does not allow one to provide confidence intervals to the identified form, nor can it give insights into the parameters most affecting the identification accuracy. Further, as the optimization step requires a large number of cost function evaluations, this approach is thought to become intractable in the general case.

Another popular approach for identifying the most probable parameter values of a system is the Bayesian approach. It relies on conceptual grounds similar to the maximum likelihood approach, but incorporates prior information in the parameters to estimate, usually leading to a better conditioned problem. Further, it results in probability distribution functions rather than in point-estimates. It has been successfully applied to various fields (see for instance Wang & Zabaras (2004), Marzouk *et al.* (2007), Marzouk & Najm (2009), Zabaras & Ganapathysubramanian (2008), Soize & Ghanem (2009) or Koutsourelakis (2009)). However, it is not easy to incorporate several sources of uncertainty into the Bayesian framework, as necessary here. Further, the resulting identified field is often highly sensitive to the prior distribution, conflicting with the motivation of this work where we aim at providing confidence intervals for the identified source field.

In fact, there is a specific point to bear in mind: in the problem at hand, the uncertainty in the model and the sensor is *epistemic* as opposed to *aleatory* for the random process $S(\boldsymbol{x}, t; \theta)$. This different nature of uncertainty allows one to separate their impact and one can make use of this to quantify the uncertainty associated with the identified stochastic process. A maximum likelihood-related approach does not account for this distinction and would result in a point-wise estimate. Further, the estimate would be erroneous because it is biased by the uncertainty in the model and the sensor.

1.2 Proposed method

In this paper, we propose to solve the stochastic data assimilation problem with the objective to identify a model of the stochastic input linked to the stochastic model of the physical problem at hand. This means that the stochastic model of the input will depend on the probabilistic models of both the uncertain constitutive parameters of the physical problem and the output (sensor uncertain measurements). Indeed, the poorly known parameters are parameterized with a set of known random variables $\boldsymbol{\xi}$ so that the probabilistic model of the identified source field is determined up to this $\boldsymbol{\xi}$-parameterization. Similarly, the identified source field is determined up to the stochastic parameterization of the measurement model.

The identification of the source stochastic properties is achieved through a combination of inverse problems and statistical reduction. For the m-th realization θ_m of a seismic event, the solution of an inverse problem allows one to estimate the deterministic source field $S^m \equiv S(\boldsymbol{x}, t; \theta_m)$ giving rise to the m-th realization of the observed data. A statistical reduction procedure is applied to the collection of identified source fields S^m at a given time t^m, $m = 1, \ldots, m_{obs}$, and allows one to approximate it in a Karhunen-Loève-like expansion form. One could think of letting the $\boldsymbol{\xi}$-parameterization affect the empirical correlation matrix, which would then result in eigenvectors and eigenvalues directly parameterized in terms of $\boldsymbol{\xi}$. However, this approach is hardly feasible since there is no guarantee of smoothness of the resulting eigenvectors and eigenvalues with respect to $\boldsymbol{\xi}$. Instead, the parameterization is incorporated directly in the correlation matrix determination: the $\boldsymbol{\xi}$-parameterization is described in a given functional form, say polynomial, so that it reduces to a finite set of variables (say, the polynomial coefficients). It results that the m-th identified source field is modeled as uncertain and corresponds to the uncertain input which, combined with uncertainty in the constitutive parameters of the physical model and uncertainty in the measurements, will result in the m-th measured data field. Upon describing these latter uncertainties with Polynomial Chaos (PC), the identified stochastic field is identified with a PC expansion and results in a vector-valued field of the PC coefficients.

The correlation matrix is then derived from the set of m_{obs} vector-valued fields and the resulting eigenvectors lie in a $\Omega_{\boldsymbol{x}} \times \Omega_{\boldsymbol{\xi}}$ space. A functional form, *e.g.*, Polynomial Chaos-based, is then prescribed for an approximation of the corresponding random variables, and the procedure is then conceptually similar to that of Desceliers *et al.* (2006) and Desceliers *et al.* (2007) with a specific distinction being that no uncertainty was considered in the measurements in these papers. As a consequence, their approach was to maximize the likelihood of a deterministic field in contrast with the present work where uncertainty in the measurements and in some exogenous parameters leads to rely on a

random field.

In the end, the identified probabilistic model of the source comprises both the epistemic and the aleatoric uncertainty. One is then able to estimate the uncertainty with which the stochastic process at the origin of the ocean surface motion is determined, hence defining the confidence one may have in the identified model based on the confidence one has in the physical model and the observational data. Once an optimal identification is found, the impact of the poorly known, $\boldsymbol{\xi}$-parameterized, sensor and model parameters onto the identified source field probabilistic model can be quantified from the eigenvectors' dependence on $\boldsymbol{\xi}$.

The paper is organized as follows. Section 2 presents the general seaquakes source problem considered in this paper for statistical identification. The different sources of uncertainty involved in the problem and their models are detailed in section 3. In section 4, the stochastic inverse problem is derived. It allows one to derive the source field corresponding to a given observed realization of the ocean surface field. This step is carried out for each realization m (seismic event). From a collection of m_{obs} retrieved source fields ($1 \leq m \leq m_{obs}$), an optimization procedure is used for the final identification of the statistical source field properties and is presented in section 5. The numerical implementation of the solution method is given in section 6. Some results are finally shown in section 7 both for the stochastic inverse technique and the full stochastic identification problem. Final conclusions are drawn in section 8.

2 Shallow water model

2.1 Formulation

Due to tectonic forces, the ocean bottom is sometimes subjected to sudden motions when and where the material can not sustain the enormous pressure and slips and deforms to release the strain. These seaquakes induce ocean surface perturbations that can be monitored and subsequently used to estimate some parameters of the corresponding ocean bottom displacement field, here modeled as a source term S. The ocean surface behavior is assumed to be governed by the Shallow Water Equations, hereafter denoted SWE, and the bottom motion field is represented by three source terms denoted by $\underline{S^u}$, $\underline{S^v}$ and $\underline{S}$. The shallow water flow is then described by the following set of equations:

$$\frac{D\,u}{D\,t} = f\,v - g_{\mathcal{G}}\,\frac{\partial w}{\partial x} - b_D\,u + \underline{S^u}, \tag{1}$$

$$\frac{D\,v}{D\,t} = -f\,u - g_{\mathcal{G}}\,\frac{\partial w}{\partial y} - b_D\,v + \underline{S^v}, \tag{2}$$

$$\frac{\partial w}{\partial t} = -\frac{\partial\left((\underline{H}+w)\;u\right)}{\partial x} - \frac{\partial\left((\underline{H}+w)\;v\right)}{\partial y} + \underline{S}, \tag{3}$$

where D/Dt denotes the substantive derivative and the fluid density is assumed to be constant as well as the free surface pressure. Here, f is the term corresponding to the Coriolis force, b_D is the viscous drag coefficient, $\boldsymbol{u} = \{u, v\}$ is the fluid velocity vector, w is the deviation of the ocean surface from its position at rest, $g_{\mathcal{G}}$ is the gravity constant and $\underline{H}$ is the ocean depth field. Depending on the latitude Φ_{lat}, one has $f = 2\,\omega_f\,\sin(\Phi_{lat})$, where ω_f is the angular rotational rate. Without loss of generality, it is chosen that the source field only acts along the w direction ($\underline{S^u} = 0$, $\underline{S^v} = 0$) and that the drag and the Coriolis forces can be neglected. Boundary conditions are prescribed along the edge of the domain $\Omega_{\boldsymbol{x}} \subset \mathbb{R}^2$ as

$$\boldsymbol{u} \cdot \boldsymbol{n} = 0, \tag{4}$$

where $\boldsymbol{n}$ is the local normal vector to the boundary.

2.2 Solution of the shallow water model

The problem is discretized using a set of non-overlapping spectral elements and results in an N_x-degree-of-freedom spatial representation for the ocean surface perturbation and the source field can be written as

$$\underline{S}^{N_x}(\boldsymbol{x}, t) = \sum_{k=1}^{N_x} \underline{S_k}(t)\,e_k(\boldsymbol{x}), \tag{5}$$

where $e_1(\boldsymbol{x}), \ldots, e_{N_x}(\boldsymbol{x})$ are the usual finite element interpolation functions related to a finite element mesh of domain $\Omega_{\boldsymbol{x}}$. More details about the numerical treatment of the SWE problem are provided in section 7.1. The above equation is rewritten as

$$\underline{S}^{N_x}(\boldsymbol{x}, t) = \mathbf{e}(\boldsymbol{x})^T\,\underline{\boldsymbol{S}}(t), \tag{6}$$

where

$$\mathbf{e}(\boldsymbol{x}) \equiv (e_1(\boldsymbol{x}) \ldots e_{N_x}(\boldsymbol{x}))^T \quad \text{and} \quad \underline{\boldsymbol{S}}(t) \equiv \left(\underline{S_1}(t) \ldots \underline{S_{N_x}}(t)\right)^T. \tag{7}$$

The SWE solution $w^{N_x}(\boldsymbol{x}, t)$ depends on $\boldsymbol{x}$, t, $\underline{\boldsymbol{S}}$ and $\underline{H}$ so that a functional g

can be formally introduced, such as:

$$w^{N_x}(\boldsymbol{x}, t) = g(\boldsymbol{x}, t; \underline{\boldsymbol{S}}, \underline{H}). \tag{8}$$

3 Uncertainty sources

3.1 Ocean depth field

From the SWE, it is seen that the ocean depth field H directly impacts the surface perturbations, w. However, the local ocean depth field is poorly known and limited knowledge is available to describe it. It is modeled as an uncertain field depending on m_{H} stochastic germs $\widetilde{\xi}_1, \ldots, \widetilde{\xi}_{m_{\mathrm{H}}}$ that are modeled as m_{H} independent uniform random variables lying on $[-1,\ 1]$. Without loss of generality, one here takes $m_{\mathrm{H}} = 1$:

$$H(\boldsymbol{x}; \widetilde{\xi}_1, \ldots, \widetilde{\xi}_{m_{\mathrm{H}}}) \equiv \underline{H}(\boldsymbol{x}) + c_0\, \widetilde{\xi}_1\, H_1(\boldsymbol{x}), \tag{9}$$

with $\underline{H}(\boldsymbol{x})$ the mean depth field and $c_0 \geq 0$ drives the variance of $H(\boldsymbol{x})$. More complex descriptions may be used such as a Karhunen-Loève decomposition if the covariance kernel is known.

3.2 Stochastic Shallow Water Model

To improve on the SWE model, intrinsic uncertainties related to the seaquake source have to be taken into account. As seen above, the goal is to identify the bottom displacement field statistics given only observations of the ocean surface. The representation of the source field is then modeled by a random field S indexed by $\Omega_{\boldsymbol{x}} \times T$ written as

$$S^{N_x}(\boldsymbol{x}, t) = \mathbf{e}(\boldsymbol{x})^T \boldsymbol{S}(t), \tag{10}$$

where $\mathbf{e}(\boldsymbol{x})$ is the vector used in the representation introduced in section 2.2 and $\boldsymbol{S}$ is now an $\mathbb{R}^{N_x}$-valued stochastic process indexed by T.

Consequently, from Eq. (8), the representation of the ocean surface deviation from its position at rest is now a random field W^{N_x} indexed by $\Omega_{\boldsymbol{x}} \times T$ written as

$$W^{N_x}(\boldsymbol{x}, t) = g(\boldsymbol{x}, t; \boldsymbol{S}, H). \tag{11}$$

3.3 Satellite measurements

The surface displacement of the ocean is modeled as a random field W indexed by $\Omega_{\boldsymbol{x}} \times T$ and which is correlated with the ocean depth and the source of the seaquake. A tractable and practical way for getting information about the ocean bottom displacement is through observations of the ocean surface response to a seaquake event. The available observations comprise measurements of the height of a large area of the ocean surface in time as obtained from dedicated satellites such as Topex-Poseidon or the recently launched Jason-2. Let $w^1_{obs}(\boldsymbol{x}, t), \ldots, w^{m_{obs}}_{obs}(\boldsymbol{x}, t)$ be measurements of ocean height at a given time t and at position $\boldsymbol{x}$ from m_{obs} statistically independent seismic events. These measurements constitute the experimental database and are used to retrieve information about the source of the seaquakes. The set of observations is modeled as m_{obs} statistically independent realizations of a random field $W_{obs}(\boldsymbol{S}, H)$ indexed by $\Omega_{\boldsymbol{x}} \times T$. As with any measure, the surface observations are uncertain and it is assumed that the uncertainties related to the measurement process are known and that the random field $W_{sat}(\boldsymbol{S}, H)$ can be written as

$$W_{obs}(\boldsymbol{S}, H) = W(\boldsymbol{S}, H) + c_{sat}\,\xi, \tag{12}$$

where ξ is a uniform random variable with values on $[-1,\, 1]$ and the constant c_{sat} represents the amount of uncertainty associated with the measures. Let the random field $W_{sat}(\boldsymbol{S}, H, \xi)$ indexed by $\Omega_{\boldsymbol{x}} \times T$ be the random observation of surface ocean displacement field constructed with the stochastic shallow water model. We then have

$$W_{sat}(\boldsymbol{S}, H, \xi) = g(\boldsymbol{x}, t; \boldsymbol{S}, H) + c_{sat}\,\xi, \tag{13}$$

4 Optimal random source field for each observation of the experimental database

4.1 Algebraic representation

We introduce the q-th order Polynomial Chaos expansion (hereafter denoted PC, Ghanem & Spanos (2003)) of $\boldsymbol{S}^m(t)$ for all $t \in T$ writing as

$$\boldsymbol{S}^m(t) \approx \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=0}^{q} \mathbf{a}^m_{\boldsymbol{\alpha}}(t)\, L_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{14}$$

where $\boldsymbol{\xi} = (\widetilde{\xi}_1, \ldots, \widetilde{\xi}_{m_{\mathrm{H}}}, \xi)$ is an $\mathbb{R}^{m_{\mathrm{H}}+1}$-valued vector of uniform random variables on $[-1,\, 1]$, $\boldsymbol{\alpha} = (\alpha_1 \ldots \alpha_{m_{\mathrm{H}}+1})$ with $|\boldsymbol{\alpha}| \equiv \sum_{k=1}^{m_{\mathrm{H}}+1} |\alpha_k|$ are the $\mathbb{N}^{m_{\mathrm{H}}+1}$-valued multi-indexes and $\mathbf{a}^m_{\boldsymbol{\alpha}}(t)$ are the $\mathbb{R}^{N_x}$-valued coefficients of the PC ex-

pansion of $\boldsymbol{S}^m(t)$. The polynomials $L_{\boldsymbol{\alpha}}$ are the multivariate Legendre polynomials such that

$$L_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \prod_{j=1}^{m_{\mathrm{H}}+1} L_{\alpha_j}\left(\{\boldsymbol{\xi}\}_j\right), \tag{15}$$

where L_{α_j} is a Legendre polynomial of degree α_j. Then Eqs. (10) and (14) yields

$$S^{N_x}(\boldsymbol{x}, t) \approx \mathbf{e}(\boldsymbol{x})^T \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=0}^{q} \mathbf{a}^m_{\boldsymbol{\alpha}}(t)\, L_{\boldsymbol{\alpha}}(\boldsymbol{\xi}). \tag{16}$$

4.2 Cost function for identification

For each experimental observation w^m_{obs}, $m = 1, \ldots, m_{obs}$, we introduce an optimal stochastic process $\boldsymbol{S}^{m,\mathrm{opt}}$ indexed by T, written as

$$\boldsymbol{S}^{m,\mathrm{opt}}(t) = \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=0}^{q} \mathbf{a}^{m,\mathrm{opt}}_{\boldsymbol{\alpha}}(t)\, L_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{17}$$

such that $\|W_{sat}(\boldsymbol{S}^{m,\mathrm{opt}}, H, \xi) - w^m_{obs}\|^2_S$ is minimal, where, for any second-order random field F indexed by $\Omega_{\boldsymbol{x}} \times T$, the norm $\|\cdot\|_S$ is defined as $\|F\|^2_S = \int_T \int_{\Omega_{\boldsymbol{x}}} E\left(F(t, \boldsymbol{x})|^2\right) d\boldsymbol{x}\, dt$ and where $E(\cdot)$ is the mathematical expectation operator. Consequently, the vector $\mathbf{a}^{m,\mathrm{opt}}(t) = (\mathbf{a}^{m,\mathrm{opt}}_{\boldsymbol{\alpha}}(t)\,,\, \boldsymbol{\alpha}\,;\, |\boldsymbol{\alpha}| \leq q) \in \mathbb{R}^{N^a_{\mathrm{chaos}}}$ is solution of an optimization problem

$$\mathbf{a}^{m,\mathrm{opt}} = \underset{\mathbf{a}^m \in \mathcal{A}}{\arg\min} \left\| W_{sat}\left(\sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=0}^{q} \mathbf{a}^m_{\boldsymbol{\alpha}}\, L_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), H, \xi\right) - w^m_{obs} \right\|^2_S, \tag{18}$$

where $\mathcal{A}$ is the set of all mappings $\mathbf{a}^m$ from T to $\mathbb{R}^{N^a_{\mathrm{chaos}}}$ such that $\mathbf{a}^m(t) = (\mathbf{a}^m_{\boldsymbol{\alpha}}(t)\,,\, \boldsymbol{\alpha}\,;\, |\boldsymbol{\alpha}| \leq q)$. The length of $\mathbf{a}^{m,\mathrm{opt}}$ is $N^a_{\mathrm{chaos}} = (P+1)\, N_x$ where $(P+1) = (q + m_{\mathrm{H}} + 1)!/(q!\,(m_{\mathrm{H}} + 1)!)$ is the number of terms in the PC expansion, Eq. (14).

Formulating Eq. (18) in words, one looks for the set of coefficients $\mathbf{a}^{m,\mathrm{opt}}$ such that the corresponding stochastic source field, combined with the uncertainty in the ocean depth field and the measurement, leads to the same (deterministic) ocean surface displacement field *for all statistical realizations*. Then, for each experimental observations, we construct the stochastic field $S^{m,\mathrm{opt}}$ for $m = 1, \ldots, m_{obs}$, such that

$$S^{m,\mathrm{opt}}(\boldsymbol{x}, t) = \mathbf{e}(\boldsymbol{x})^T \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=0}^{q} \mathbf{a}^{m,\mathrm{opt}}_{\boldsymbol{\alpha}}(t)\, L_{\boldsymbol{\alpha}}(\boldsymbol{\xi}). \tag{19}$$

The unconstrained optimization problem, Eq. (18), is here solved using a second-order quasi-Newton memory-limited technique, Gilbert & Lemaréchal (1989), relying on the gradient of the functional computed from the solution of the adjoint Shallow Water Equations. Iteratively solving the primal and dual SWE allows to converge to $\mathbf{a}^{m,\mathrm{opt}}$ within just a few tens of iterations. No regularization term was found necessary thanks to the weakly non-linear character of the governing equations.

5 Global optimal random source field for the experimental database

5.1 Algebraic representation

For *each* observation w_{obs}^m, $m = 1, \ldots, m_{obs}$, the polynomial expansion of a random field $\boldsymbol{S}^{m,\mathrm{opt}}$ is constructed by solving Eq. (18). While the random fields S, S^{N_x} and $S^{m,\mathrm{opt}}$ are indexed on the time and space domain, we are only interested in describing their spatial structure at the time t^m when their mean mechanical energy is maximum. For a given stochastic source $S^{m,\mathrm{opt}}$, t^m is then

$$t^m \equiv \arg\max_{t \in T} \left\| E\{S^{m,\mathrm{opt}}(\boldsymbol{x}, t)\} \right\|_{\Omega_{\boldsymbol{x}}}^2. \tag{20}$$

In the rest of the paper, we will hence drop the dependence in time and all time dependent quantities will be evaluated at $t = t^m$.

In this section, a *global* optimal representation of the random fields $S^{m,\mathrm{opt}}$ is constructed by considering each coefficient $\mathbf{a}^{m,\mathrm{opt}}$ as the m-th independent realization of a random vector $\boldsymbol{A}$ such that $\boldsymbol{A} \equiv (\mathbf{A}_{\boldsymbol{\alpha}}\,,\, \boldsymbol{\alpha}\,;\, |\boldsymbol{\alpha}| \leq q)$, where $\mathbf{A}_{\boldsymbol{\alpha}}$ is an $\mathbb{R}^{N_x}$-valued random vector. We then define S^{opt} as

$$S^{\mathrm{opt}}(\boldsymbol{x}) = \mathbf{e}(\boldsymbol{x})^T \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=0}^{q} \mathbf{A}_{\boldsymbol{\alpha}}\, L_{\boldsymbol{\alpha}}(\boldsymbol{\xi}). \tag{21}$$

In this section, an algebraic representation of random vector $\mathbf{A}_{\boldsymbol{\alpha}}$ is constructed using Polynomial Chaos expansion. Since such a construction can be computationally expensive, a statistical reduction of the random vector $\boldsymbol{A}$ is first introduced. Different approaches may be thought of to derive a suitable representation of a random vector. Among them, one can cite Principal Component Analysis (also termed Karhunen-Loève), as will be considered in this work; factor analysis, useful when noise is present in the identification process; projection pursuit, which basically tries to find "meaningful" directions in the data (Webb (2002)); and Independent Component Analysis (ICA), where a basis leading to the least dependent components is searched for (Hyvärinen

(1999)). In the present case, the Karhunen-Loève approach is the method of choice since one is basically interested in reducing the size of the dataset to the most significant modes in the L^2-sense rather than deriving a basis spanned by a potentially large number of independent random variables.

Let $[C]$ be the $(N^a_{\text{chaos}} \times N^a_{\text{chaos}})$ positive-definite symmetric real matrix estimated for large value of $m_{obs} > N^a_{\text{chaos}}$ by

$$[C] \simeq \frac{1}{m_{obs}} \sum_{m=1}^{m_{obs}} \left(\mathbf{a}^{m,\text{opt}} - \underline{\mathbf{A}}\right) \left(\mathbf{a}^{m,\text{opt}} - \underline{\mathbf{A}}\right)^T, \tag{22}$$

with

$$\underline{\mathbf{A}} \equiv \frac{1}{m_{obs}} \sum_{m=1}^{m_{obs}} \mathbf{a}^{m,\text{opt}}. \tag{23}$$

Let $\lambda^1 \geq \ldots \geq \lambda^{N^a_{\text{chaos}}} > 0$ and $\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^{N^a_{\text{chaos}}}$ respectively be the eigenvalues and the normalized eigenvectors of matrix $[C]$ such that $\|\boldsymbol{\psi}^1\|_{\Omega_{\boldsymbol{x}}} = \ldots = \|\boldsymbol{\psi}^{N^a_{\text{chaos}}}\|_{\Omega_{\boldsymbol{x}}} = 1$. Then, a N_{red}-term statistically reduced representation of the random vector $\boldsymbol{A}$ is expressed in terms of the random vector $\boldsymbol{B}^T = (B_1 \ \ldots \ B_{N_{\text{red}}})$ such that

$$\boldsymbol{A} \approx \underline{\mathbf{A}} + [\psi]\,[\Lambda]\,\boldsymbol{B}, \tag{24}$$

where $[\psi]$ is an $N^a_{\text{chaos}} \times N_{\text{red}}$ orthonormal matrix with $[\psi]_{jk} = \{\boldsymbol{\psi}^j\}_k$, $[\Lambda]$ is a $N_{\text{red}} \times N_{\text{red}}$ matrix with $[\Lambda]_{jk} = \sqrt{\lambda^k}\,[I]_{jk}$ and where $[I]$ is the identity matrix of $\mathbb{R}^{N_{\text{red}}}$. We then have

$$\boldsymbol{B} = [\Lambda]^{-1}\,[\psi]^T\,(\boldsymbol{A} - \underline{\mathbf{A}}). \tag{25}$$

From the zero-mean orthonormal properties of $[\psi]$, it follows that:

$$E(\boldsymbol{B}) = \mathbf{0}, \quad E\left(\boldsymbol{B}\,\boldsymbol{B}^T\right) = [I]. \tag{26}$$

Since $\mathbf{a}^{m,\text{opt}}$ is assumed to be the m-th realization of $\boldsymbol{A}$, the realization $\boldsymbol{B}(\theta_m)$ of $\boldsymbol{B}$ is such that, for all $m = 1, \ldots, m_{obs}$,

$$\boldsymbol{B}(\theta_m) = [\Lambda]^{-1}\,[\psi]^T\,\left(\mathbf{a}^{m,\text{opt}} - \underline{\mathbf{A}}\right). \tag{27}$$

In order to derive a tractable statistical representation of $\boldsymbol{B}$, it is decomposed into a PC expansion:

$$\boldsymbol{B} \approx \widehat{\boldsymbol{B}} \equiv \sum_{\boldsymbol{\beta},|\boldsymbol{\beta}|=0}^{p} \mathbf{b}_{\boldsymbol{\beta}}\,H_{\boldsymbol{\beta}}(\boldsymbol{\zeta}), \tag{28}$$

where, for a given $\nu \in \mathbb{N}$, $\boldsymbol{\zeta}$ is an $\mathbb{R}^\nu$-valued normal random variable of probability density $p_{\boldsymbol{\zeta}}$, $\boldsymbol{\beta}^T \equiv (\beta_1 \, \ldots \, \beta_\nu)$ with $|\boldsymbol{\beta}| = \sum_{k=1}^{\nu} |\beta_k| \leq p$ are the $\mathbb{N}^\nu$-valued multi-indexes and $H_{\boldsymbol{\beta}}$ are the unit-norm multivariate Hermite polynomials, such that

$$H_{\boldsymbol{\beta}}(\boldsymbol{\zeta}) = \prod_{j=1}^{\nu} H_{\beta_j}\left(\{\boldsymbol{\zeta}\}_j\right), \tag{29}$$

with H_{β_j} the Hermite polynomial of degree β_j. Then, from Eqs. (26), it can be deduced that the constraints $\mathbf{b}_{\boldsymbol{\beta}}(t)$ must obey:

$$\mathbf{b}_0 = \mathbf{0}, \qquad \sum_{\boldsymbol{\beta}, |\boldsymbol{\beta}|=0}^{p} \mathbf{b}_{\boldsymbol{\beta}} \, \mathbf{b}_{\boldsymbol{\beta}}{}^T = [I]. \tag{30}$$

5.2 Cost function for identification

Let $\mathbf{b} = (\mathbf{b}_{\boldsymbol{\beta}} \, , \, \boldsymbol{\beta} \, ; \, |\boldsymbol{\beta}| \leq p)$ be the vector which consists of all the coefficients of the Polynomial Chaos expansion of $\boldsymbol{B}$. Let $b \mapsto p_{\widehat{B}_j}(b; \mathbf{b})$ be the probability density function of the j-th random component $\widehat{B}_j = \{\widehat{\boldsymbol{B}}\}_j$. Then, the coefficients $\mathbf{b}_{\boldsymbol{\beta}}$ of the polynomial chaos expansion in Eq. (28) are identified as the vector $\mathbf{b} \in \mathbb{R}^{N^b_{\text{chaos}}}$, $N^b_{\text{chaos}} \equiv N_{\text{red}} \, (p+\nu)! / \, (p! \, \nu!)$, such that the most likely (highest probability) approximated representation of $\boldsymbol{B}$, Eq. (28), is that derived from the observations through Eq. (27). In other words,

$$\mathbf{b} \equiv \underset{\mathbf{b}' \, \in \, \mathcal{B}}{\arg\max} \prod_{m=1}^{m_{obs}} p_{\widehat{\boldsymbol{B}}} \left(\boldsymbol{B}_m ; \mathbf{b}'\right) = \underset{\mathbf{b}' \, \in \, \mathcal{B}}{\arg\max} \, \mathcal{L}(\mathbf{b}'), \tag{31}$$

which constitutes an optimization problem constrained by the fact that $\mathbf{b}$ lies on the manifold $\mathcal{B}$ defined by Eq. (30).

Finally, we can summarize the main equations at hand,

$$S^{\text{opt}}(\boldsymbol{x}) = \mathbf{e}(\boldsymbol{x})^T \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=0}^{q} \mathbf{A}_{\boldsymbol{\alpha}} \, L_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \quad \text{[Eq. (21)]}\,, \tag{32}$$

with $\boldsymbol{A} = (\mathbf{A}_{\boldsymbol{\alpha}} \, , \, |\boldsymbol{\alpha}| \leq q)$ and

$$\boldsymbol{A} \approx \underline{\mathbf{A}} + [\psi] \, [\Lambda] \sum_{\boldsymbol{\beta}, |\boldsymbol{\beta}|=0}^{p} \mathbf{b}_{\boldsymbol{\beta}} \, H_{\boldsymbol{\beta}}(\boldsymbol{\zeta}), \quad \text{[Eqs. (24) and (28)]}\,, \tag{33}$$

where $\mathbf{b} \equiv (\mathbf{b}_{\boldsymbol{\beta}} \, , \, \boldsymbol{\beta} \, ; \, |\boldsymbol{\beta}| \leq p)$ is solution of the optimization problem defined by Eq. (31).

6 Statistical identification

6.1 Regularized optimization problem

The probability density function $p_{\widehat{\boldsymbol{B}}}(\boldsymbol{B}(\theta_m); \mathbf{b}')$ reads

$$p_{\widehat{\boldsymbol{B}}}(\boldsymbol{B}(\theta_m); \mathbf{b}') = \int_{\mathbb{R}^\nu} \delta_{\mathbf{0}} \left(\boldsymbol{B}(\theta_m) - \sum_{\boldsymbol{\beta}, |\boldsymbol{\beta}|=0}^{p} \mathbf{b}'_{\boldsymbol{\beta}} \, H_{\boldsymbol{\beta}}(\boldsymbol{s}) \right) p_{\boldsymbol{\zeta}}(\boldsymbol{s}) \, d\boldsymbol{s}, \tag{34}$$

where $\delta_{\mathbf{0}}$ is the N_{red}-dimensional Dirac distribution,

$$\delta_{\mathbf{0}}(\boldsymbol{s}) : \quad \mathbb{R}^{N_{\text{red}}} \ni \boldsymbol{s} \mapsto \delta_{\mathbf{0}}(\boldsymbol{s}) = \prod_{i=1}^{N_{\text{red}}} \delta(s_i) \in \mathbb{R}. \tag{35}$$

As seen from Eqs. (31) and (34), the determination of the optimum $\mathbf{b}'$ implies the computation of ν-D integrals. The integrand is inexpensive to evaluate since it only involves basic operations on the PC series, and a Monte-Carlo method can be used to compute the integrals. For the subsequent use of gradient-based methods, a derivable integrand is desirable. Consider a surrogate delta operator defined as

$$\delta_{\mathbf{0}}\left(\widetilde{\boldsymbol{s}}\right) \approx \frac{a}{a^2 + d(\widetilde{\boldsymbol{s}})^{2\, n_\delta}}, \qquad d(\widetilde{\boldsymbol{s}}) \in \mathbb{R}^+, \; \widetilde{\boldsymbol{s}} \in \mathbb{R}^{N_{\text{red}}}, \tag{36}$$

with $0 < a \ll 1$, $n_\delta \in \mathbb{N}$ and where $d(\widetilde{\boldsymbol{s}})$ is the distance in $\mathbb{R}^{N_{\text{red}}}$ between $\widehat{\boldsymbol{B}}$ and $\boldsymbol{B}(\theta_m)$. Note that this is in a similar spirit as the approximation of probability distribution functions using a kernel density estimation approach.

However, as the dimension N_{red} of the random vector $\widehat{\boldsymbol{B}}$ grows, the support of the delta function $\delta_{\mathbf{0}}$ or its surrogate decreases exponentially and thus requires an increasing number of Monte-Carlo samples to accurately evaluate the integrals. This makes this approach difficult even for moderate length random vectors, say for $N_{\text{red}} \gtrsim 5$. Effectively sampling such integrals has attracted a great deal of attention in the past two decades and usually relies on an adaptive defensive importance sampling based on some approximation of the integrand, see Owen & Zhou (1998) among others. Another approach is chosen here and a modified expression of the cost function is defined in Eq. (31) is considered:

$$\log \mathcal{L} \equiv - \sum_{m=1}^{m_{obs}} \log \mathcal{L}_m, \quad \mathcal{L}_m \equiv \prod_{j=1}^{N_{\text{red}}} p_{\widehat{B}_j}\left(B_j(\theta_m); \mathbf{b}'\right). \tag{37}$$

This alternative formulation prevents very low probabilities from having to

be estimated as is the case when the support of the integrand decreases, thus softening the requirement on the number of Monte-Carlo samples necessary to estimate the integrals with a given accuracy. Of course, the optima from the two formulations of the cost function are the same when the random components B_j, $1 \leq j \leq N_{\text{red}}$, are independent. This alternative definition of the likelihood will be used throughout the rest of the paper.

This kind of optimization problem is often solved with a simplex approach based on point estimations of the cost function for different sets of parameters $\mathbf{b}'$. Here we take advantage of the regularity of the cost function expression to derive an efficient gradient-based second order optimization technique.

A potential issue is that the solution of the optimization problem lies in a high-dimensional $\left(\mathbb{R}^{N^b_{\text{chaos}}}\right)$ space, bringing specific difficulties. In particular, the cost function response surface may exhibit local extrema and the optimization problem achieves only local convexity. The solution method relies on a combination of techniques. A good initial point candidate is given by matching the first statistical moments, Eggermont & LaRiccia (2001). However, this can only be done up to a limited order since it involves increasingly higher order inner products $E\left(H_{\boldsymbol{\beta}_i}\, H_{\boldsymbol{\beta}_j}\, H_{\boldsymbol{\beta}_k} \ldots\right)$ of the PC basis, which are costly to evaluate. Further, the accuracy of the experimental high order statistical moments estimation is poor since the number m_{obs} of observations is limited. The first four moments are considered and an initial point candidate is determined. Because the optimization space is of dimension N^b_{chaos}, which is much higher than four, we also rely on a collection of randomly chosen vectors on the constraint manifold $\mathcal{B}$. From the best, in the sense of $\mathcal{L}$, initial point found, a simulated annealing procedure is carried-out until it is reasonable to think that the basin of attraction of a good minimum (rather than a maximum, note the minus sign in Eq. (37)) has been reached. The optimization problem is then assumed locally strictly convex and a gradient-based approach is applied.

From Eqs. (34) and (37), the gradient of the cost function $\mathcal{L}$ can be computed as

$$\boldsymbol{\nabla}_{\left\{\mathbf{b}'_{\boldsymbol{\beta}}\right\}_j} \log \mathcal{L} = \sum_{m=1}^{m_{obs}} \sum_{j=1}^{N_{\text{red}}} \frac{2\, n_\delta\, a}{p_{\widehat{B}_j}\left(B_j(\theta_m); \mathbf{b}'\right)} \int_{\mathbb{R}^\nu} \frac{d_j^{2\, n_\delta - 1}(\boldsymbol{s})\, H_{\boldsymbol{\beta}}(\boldsymbol{s})}{\left(a^2 + d_j^{2\, n_\delta}(\boldsymbol{s})\right)^2}\, p_{\boldsymbol{\zeta}}(\boldsymbol{s})\, d\boldsymbol{s}, \quad (38)$$

with $|\boldsymbol{\beta}| \leq p$ and

$$d_j^2(\boldsymbol{s}) = \left(\widehat{B}_j - B_j(\theta_m)\right)^2. \quad (39)$$

To solve the optimization problem involving the non-linear constraint defining the manifold $\mathcal{B}$, a SQP technique achieving quadratic convergence is employed. It relies on a memory-limited second-order quasi-Newton algorithm (Gilbert & Lemaréchal, 1989) and estimates the Hessian matrix without computing it. The optimal point in $\mathbb{R}^{N^a_{\text{chaos}}}$ is found through a multi-grid-like procedure

as the PC order is initially set to a low value to find a coarse estimate of the stochastic source field. Once this estimate is found, a finer discretization is used from that initial point to determine the accurate source field. This approach reduces the risk of ending-up in a local, as opposed to a global, minimum of the Lagrangian and lightens the overall computational burden.

6.2 Algorithm

The global algorithm for solving the stochastic identification problem is summarized as follows:

(1) for each observation of the ocean surface field $w_{obs}^m(\boldsymbol{x}, t)$ after the m-th seismic event, solve the corresponding optimal stochastic problem, Eq. (18), to determine $\mathbf{a}^{m,\mathrm{opt}} = \mathbf{a}^{m,\mathrm{opt}}(t^m)$,
(2) from the collection of m_{obs} vectors $\mathbf{a}^{m,\mathrm{opt}}$, $1 \leq m \leq m_{obs}$, carry-out a statistical reduction and determine $\boldsymbol{B}(\theta_m)$, Eq. (27),
(3) compute the cost function from Eq. (37) for different randomly chosen initial vector-valued functions $\mathbf{b}'$ onto the constraint manifold $\mathcal{B}$ and retain the best (lowest $\mathcal{L}$) as the initial point,
(4) from this selected initial point, determine the optimal set $\mathbf{b}$, solution of Eq. (31), through a SQP-based solution method,
(5) use Eqs. (32)-(33) for constructing the optimal random field S^{opt}.

Remark The approach followed in this paper involves two main steps. In a first step, a random field is constructed, through an inverse procedure, in order to represent the observed surface displacement for each $1, \ldots, m_{obs}$ observations (section 4, Eq. (18), step (1) of the algorithm). This first inverse problem requires the use of a mechanical solver (solving the governing equations of the physical problem at hand). The second step consists in constructing a unique random field for which the set of realizations constructed in step (1) would maximize the likelihood (section 5, Eq. (31), steps (2)-(4)). In these last steps, the estimation of the likelihood does not require the use of the mechanical solver. The only difficulty lies on the high number of parameters required to represent the solution.

Alternatively, one could avoid this first step by directly searching for the parameters that would maximize the likelihood of the m_{obs} observations. As for the proposed methodology, one of the difficulties would lie on the high number of parameters required to represent the solution. But, in such method, additional difficulties arise due to the necessary intensive use of the mechanical solver involved in the evaluation of the likelihood at each iteration of the optimization procedure. This approach is hence not thought to be tractable in

the general case, unless the direct problem is very significantly cheaper than the inverse procedure and/or if the inverse step is overwhelmingly ill-posed.

7 Convergence analysis and results

7.1 Shallow Water numerical solver and solution method

To identify the seaquake source field, the inverse Shallow Water problem is solved. We rely on a 10×10-element mesh for the physical space, discretized with $p_x = p_y = 6$-th order Legendre cardinal polynomials for the fluid velocity. The unknowns are interpolated with these Legendre cardinal functions collocated at the Gauss-Lobatto points. The surface height, w, is discretized with a lower order polynomial to prevent spurious pressure modes from occurring ($Q_N - Q_{N-2}$ scheme), see Iskandarani *et al.* (1995). An additive Schwarz preconditioner is used to accelerate the solution process; see Douglas *et al.* (2003).

Solving the inverse SWE requires the solution of the adjoint equations which brings a potential issue since they involve the primal fields u, v and w. These fields may be large and their storage over the time horizon of the optimization procedure may introduce some difficulties. Different techniques exist to cope with this issue such as the check-pointing strategy. The interested reader may refer to Biros & Ghattas (2005) for further details. However, such an issue is not encountered in the present work due to its moderate size and the primal fields can be stored without specific problems.

As mentioned in section 3.1, $m_H = 1$ was retained for describing the uncertain ocean depth field. The random vector $\widetilde{\boldsymbol{\xi}}$ thus reduces to a random variable $\widetilde{\xi}$ and the random vector $\boldsymbol{\xi}$ then reduces to $\left(\widetilde{\xi}\, \xi\right)^T$. Other retained parameters include $c_{sat} = 10^{-3}$ (Eq. (13)), $q = 2$ (Eq. (16)) and $m_{obs} = 1514$.

7.2 Inverse problem verification for $\boldsymbol{S}^{m,\mathrm{opt}}$

For the purpose of verification, let us first consider the case $c_0 = 0$ in Eq. (9) and $c_{sat} = 0$ in Eq. (13). Determining $\boldsymbol{S}^{m,\mathrm{opt}}$ thus reduces to identifying the sea bottom displacement field giving rise to the observed surface perturbation: this is an inverse problem (IP) which solution is given by Eq. (18). This verification step is necessary since the IP is an essential ingredient of the whole statistic identification problem considered in this paper.

7.2.1 Identification of $\boldsymbol{S}^{m,\mathrm{opt}}$ for $c_{sat} = 0$ and $c_0 = 0$

In this case, no uncertainty affects the physical model or the observation sensor and the inverse problem is deterministic. The surface is initially at rest and a time-varying source is applied at point $\boldsymbol{x}_0 = \{x_0, y_0\}$:

$$\underline{S}(\boldsymbol{x}, t) = e^{-\left[\frac{1}{\sigma_S^2}\,\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2\right]} \times \left(\frac{t^2}{1+t^4}\,\left(1 - e^{-20\,t}\right)\right), \tag{40}$$

with $\sigma_S^2 = 0.005$ and $\boldsymbol{x}_0 = \{0.4, 0.9\}$. All variables are non-dimensional and we are interested in the source field at the time of its maximum amplitude, corresponding to $t = t^m \simeq 4$. Using the deterministic SWE model, the corresponding synthetic observation field w_{obs} may be derived and mimics an observation w_{obs}^m that would have risen from an actual seismic event.

The inverse problem is initialized with an arbitrary source field at all time, as shown in Fig. 1. In general, one often has some expertise about the most likely location of the seaquake but the goal here is to investigate the efficiency of the inverse method and its robustness. The solution field of the IP is plotted in Fig. 2 (right) together with the true solution (left) and the agreement is seen to be very satisfying. No regularization terms were found necessary to achieve good identification results, even not a simple Tikhonov approach. This mainly comes from the fact that the problem at hand is only weakly non-linear and exhibits a one-to-one correspondence between the surface fluctuation w and the source S for a whole basin of considered parameters set. The problem is then bijective in a subdomain of these parameters' space. Further, the whole solution domain is observed and the null space of the inverse SWE operator is thus essentially not in the primal space. The inverse problem is then well-posed and essentially does not require regularization.

7.2.2 Identification of $\boldsymbol{S}^{m,\mathrm{opt}}$ for $c_{sat} \neq 0$ and $c_0 \neq 0$

A more complex, and somehow more realistic, case is now considered for verifying the determination of the source field through a stochastic inverse problem (SIP): the ocean depth H is considered uncertain and the problem is then to determine a stochastic source field which, through the (uncertain) SWE equations, and given the uncertainty in the measure, would lead to surface data as close as possible to the observed (deterministic) data.

Relying on a similar synthetic observed field as for the previous test but centered around $\boldsymbol{x}_0 = \{0.3, 0.9\}$, the mean of the identified source field (PC mode 0) at $t = 4$ is plotted in Fig. 3 together with the first, Fig. 4 (left), and second, Fig. 4 (right), stochastic modes of the identified stochastic source field decomposition. The initial field was the same as that plotted in Fig. 1 (deterministic).

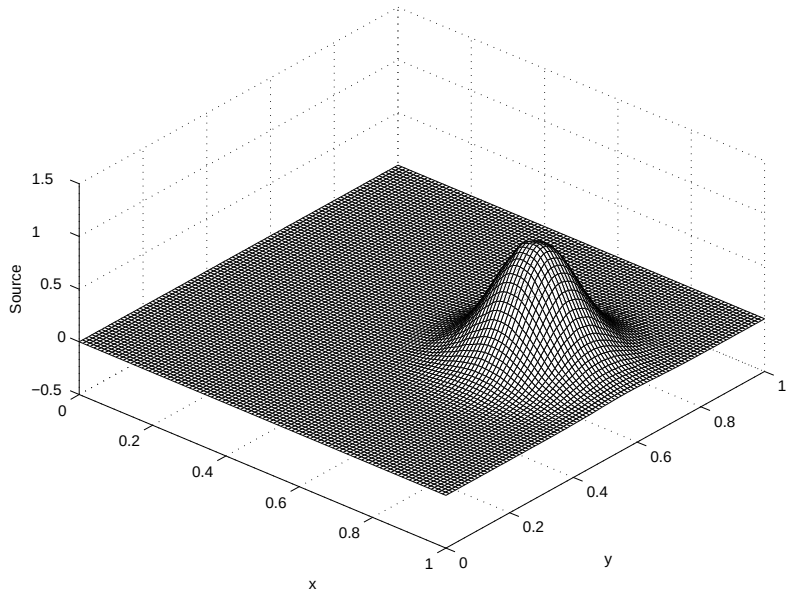

Fig. 1. Chosen initial source field $\underline{S}(\boldsymbol{x}, \cdot)$.

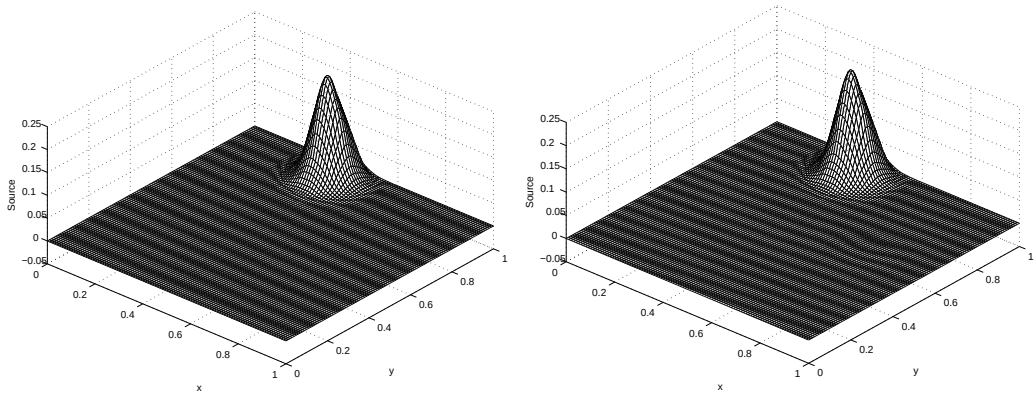

Fig. 2. Plot of $\underline{S}(\boldsymbol{x}, t)$ (left) and $S^{\text{opt}}(\boldsymbol{x}, t)$ (right) for $c_{sat} = 0$, $c_0 = 0$ and $t = 4$.

Again, the identified source is centered around the true solution. In this case, non-zero PC modes are also present as a consequence of the uncertainty of the problem.

The identified source field is stochastic instead of being deterministic, since, to lead to a given (deterministic quantity) surface observation at all times within the observation horizon T despite an uncertain depth field and observation device, the source term in the SWE equations must itself be uncertain. It represents the information one can derive on the source field from a given observation w_{obs}^m provided the uncertainty in the sensor is well characterized as well as the uncertain model describing the effects of the source field onto the ocean surface field.

The validity of the implementation is confirmed in Figs. 5, 6 and 7 where

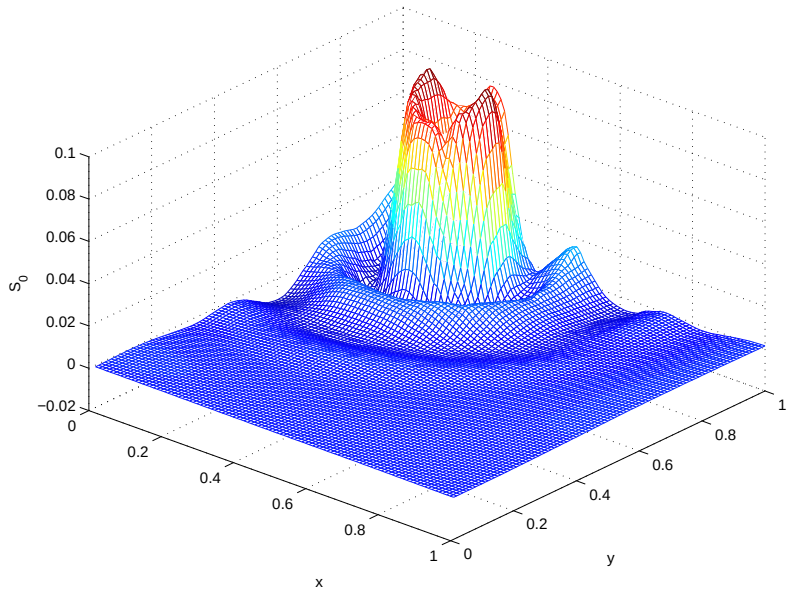

Fig. 3. Plot of the mean identified source field (PC mode 0): $S_{\boldsymbol{\alpha}}^{m,\mathrm{opt}}(\boldsymbol{x},t) = \mathbf{e}(\boldsymbol{x})^T \mathbf{a}_{\boldsymbol{\alpha}}^{m,\mathrm{opt}}(t)$ with $\boldsymbol{\alpha} = (0\,0)^T$. $t = 4$.

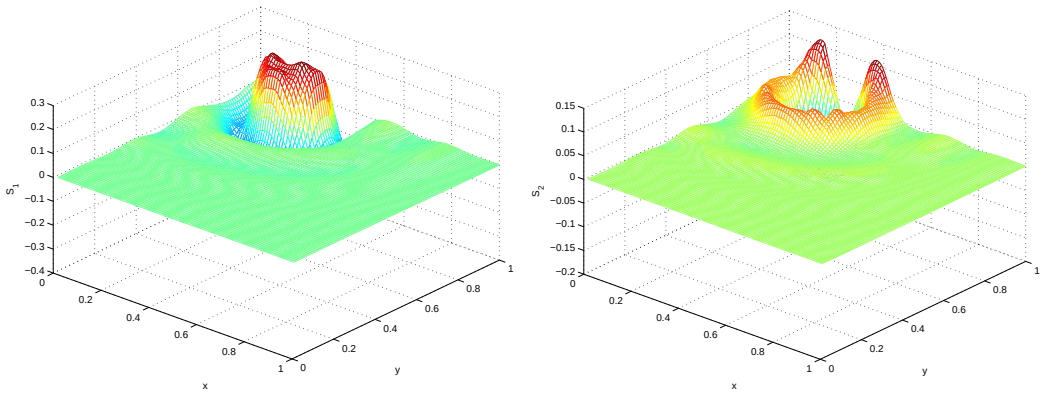

Fig. 4. Plot of the first two PC modes of the identified source field: $S_{\boldsymbol{\alpha}}^{m,\mathrm{opt}}(\boldsymbol{x},t) = \mathbf{e}(\boldsymbol{x})^T \mathbf{a}_{\boldsymbol{\alpha}}^{m,\mathrm{opt}}(t)$ with $\boldsymbol{\alpha} = (1\,0)^T$ (PC modes 1), (left) and $\boldsymbol{\alpha} = (0\,1)^T$ (PC mode 2), (right). $t = 4$.

the resulting surface motion at a particular time ($t = 300$) is computed from the SWE equations with the actual, *i.e.*, deterministic and given by Eq. (40), source field $\underline{S}(\boldsymbol{x},t)$ (left plots) and the identified source field issued from the SIP (right plots). Both surfaces are issued from the same stochastic code and only the source terms differ. In the left plots, the source term is deterministic and corresponds to the true, synthetic, source. On the right plots, the source term is that identified by the SIP. The resulting sea surface motion is seen to be similar in the statistical average sense (Fig. 5). However, the actual (deterministic) source term leads to an uncertain surface field (non-zero higher stochastic modes) while the identified (uncertain) source leads to a quasi deterministic surface (stochastic modes higher than 0 are essentially null), as shown in Figs. 6 and 7. This clearly shows that a deterministic source field

cannot induce a deterministic surface displacement field through the stochastic SWE model.

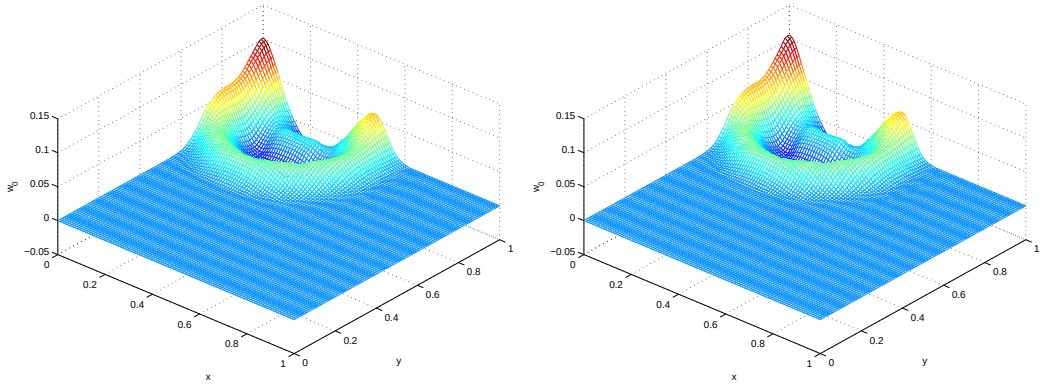

Fig. 5. Mean surface field at intermediate time ($t = 300$) for the actual (deterministic) source field (left) and the IP-computed source field (right). Both surfaces are issued from the same stochastic code and only the source terms differ.

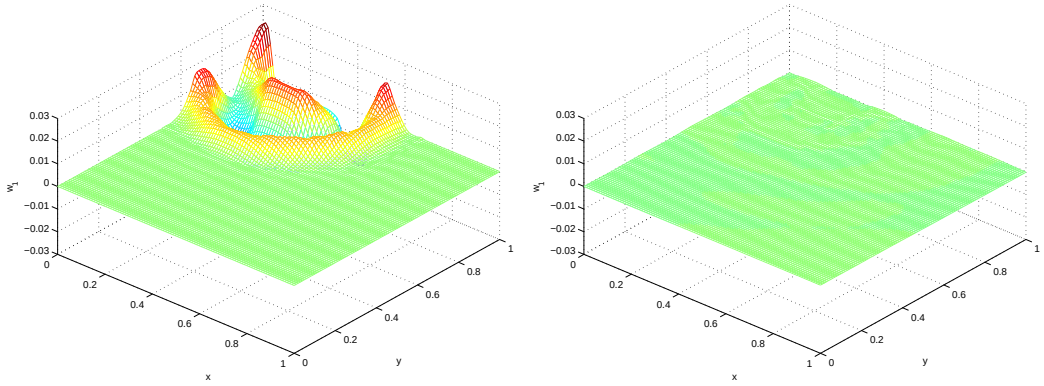

Fig. 6. Stochastic mode 1 surface field at intermediate time ($t = 300$) for the actual source field (left) and the SIP-computed source field (right).

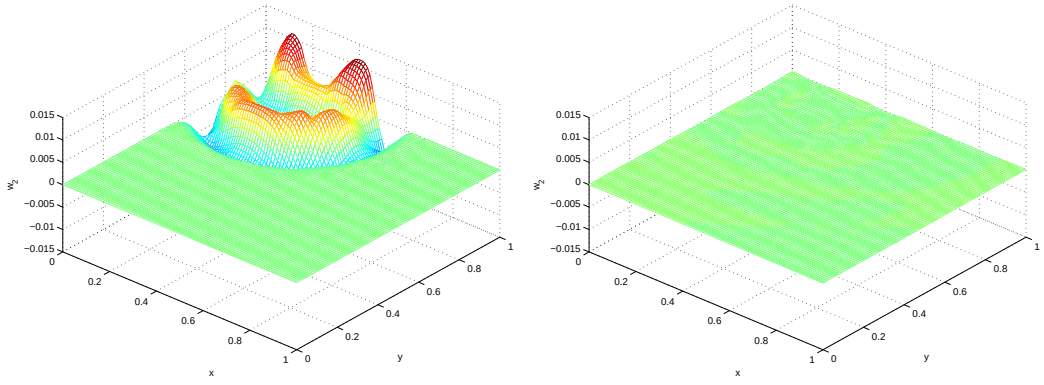

Fig. 7. Stochastic mode 2 surface field at intermediate time ($t = 300$) for the actual source field (left) and the SIP-computed source field (right).

This example and the essentially null uncertainty (quantified by the magnitude of the stochastic modes of non-zero index) in the predicted surface is a strong argument supporting the validity of the stochastic inverse problem strategy.

The subsequent global statistical identification of the seaquake random process can then reliably rely on the source estimation step. In particular, it shows that the uncertain source field, with first modes plotted in Figs. 3 and 4, is indeed a versatile description of the source field which leads to a deterministic surface field given the uncertainty in the sensor and the physical model at hand.

7.3 Convergence analysis

7.3.1 Convergence analysis of the random vector $\boldsymbol{B}$

In this section, the convergence of the stochastic identification process in terms of the Polynomial Chaos order p of the PC expansion of $\boldsymbol{B}$, see Eq. (28), and stochastic germ dimension ν of $\boldsymbol{\zeta}$ is investigated. For sake of clarity, we focus on the first random vector component B_1. The fourth statistical moment for different PC orders p and a 2-D stochastic germ ($\nu = 2$) is given in Table 1. The moment converges to an asymptotic value as the approximation relies on a finer discretization. The corresponding probability density function of the vector component is plotted in Fig. 8 for $p = 5$ and already exhibits a reasonable agreement with the pdf estimated from the data.

p	2	3	4	5	6	7	8	9
m_4^1	13.09	13.23	13.19	13.26	13.25	13.31	13.32	13.32

Table 1
Fourth statistical moment of the first random vector component $m_4^1 \equiv E\left((B_1)^4\right)$ for different PC orders p. $\nu = 2$, $N_{\text{red}} = 4$.

A similar study is carried-out for different germ dimensions ν at a given PC order ($p = 5$). Results are gathered in Table 2 in terms of the fourth statistical moment of the first component B_1. Again, the statistical moment is seen to converge when the stochastic discretization is enriched enough. A germ dimension $\nu \gtrsim 4$ leads to a reasonably converged estimation. Similar findings (not shown) hold for the other components of $\boldsymbol{B}$.

ν	1	2	3	4	5	6
m_4^1	13.42	13.26	13.34	13.44	13.45	13.45

Table 2
Fourth statistical moment of the first random vector component $m_4^1 \equiv E\left((B_1)^4\right)$ for different PC dimension ν. $p = 5$, $N_{\text{red}} = 4$.

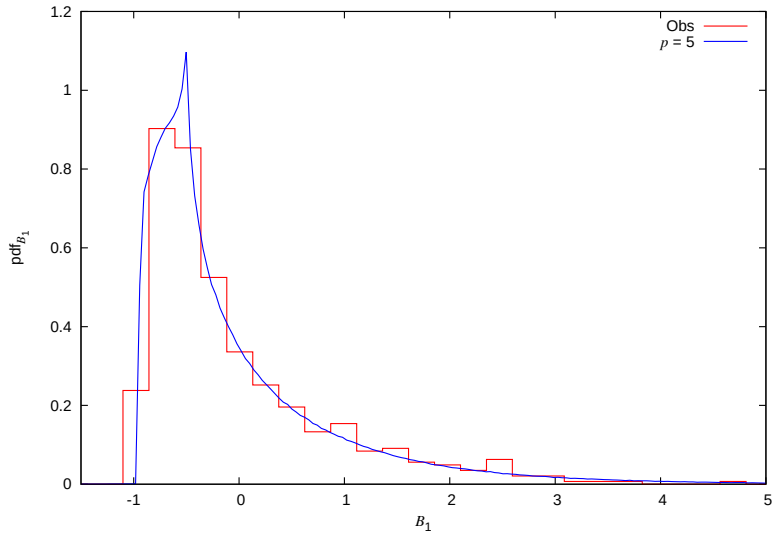

Fig. 8. Probability density function of the first random vector component $\boldsymbol{B}$ for a $p = 5$-PC order. $\nu = 2$. The pdf estimated from data is plotted for comparison (histogram).

7.3.2 Identified source statistics

The mean-square convergence of the statistically reduced decomposition (Eq. 27) resulting from the identification procedure can be appreciated from Fig. 9 in terms of the eigenvalues spectrum. The spectrum decays quickly as higher index eigenvalues are considered. Since the variance of the approximation error σ_ε^2 can be expressed as

$$\sigma_\varepsilon^2 = \sum_{i=N_{\text{red}}+1}^{N_{\text{chaos}}^a} \lambda_i, \tag{41}$$

with N_{red} being the number of modes retained, the approximation of the identified sources rapidly converges in the mean-square sense as more modes are accounted for in the reduction. The approximation series can then be truncated to a limited number of terms. From the eigenvalue spectrum, it is seen that retaining only the first four modes ($N_{\text{red}} = 4$) already accounts for most of the variance:

$$\frac{\sum_{i=1}^{4} \lambda_i}{\sum_{i=1}^{N_{\text{chaos}}^a} \lambda_i} \simeq 94\,\%. \tag{42}$$

Once the coefficients $\mathbf{b}_{\boldsymbol{\beta}}$ are found, we focus on the statistics of the identified source field, which is the main quantity of interest in this work. To this end, the auto-covariance function of the identified source random field is plotted in Fig. 10 together with that directly estimated from the collection of observations. The auto-covariance function writes

$$C_S(\boldsymbol{x}, \boldsymbol{x}') \equiv E\left(\left(S^{\text{opt}}(\boldsymbol{x}) - E\left(S^{\text{opt}}(\boldsymbol{x})\right)\right)\left(S^{\text{opt}}(\boldsymbol{x}') - E\left(S^{\text{opt}}(\boldsymbol{x}')\right)\right)\right). \tag{43}$$

It is a 4-D object and it is thus here plotted as a function of x only for the

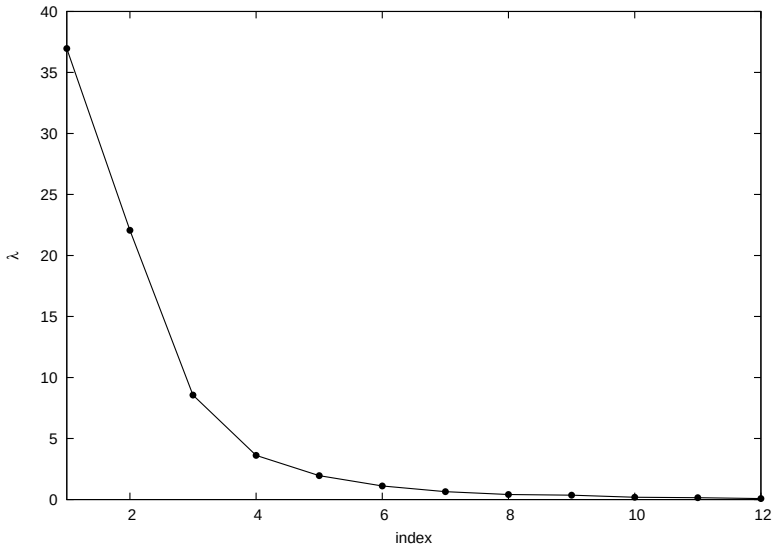

Fig. 9. Spectrum of the identified sources covariance kernel. The eigenvalues decay quickly and most of the statistical variance is accounted for by the first few modes.

sake of clarity. As expected from Fig. 9, as more modes are accounted for, the agreement between the experimental and the identified random field quickly improves as a few modes are sufficient to reach a decent approximation.

The retained approximation parameters then include $N_{\text{red}} = 4, \nu = 4, p = 5$ and the identified stochastic source field can now be fully characterized. It is finally represented in Fig. 11 in terms of its 20% and 80% quantiles (nested plots). It is seen to be essentially located in a particular region of space, with a slight elongation along the x-axis, and exhibits a large uncertainty.

Since the identification procedure is subject to uncertainty (due to sensor and direct model uncertainty), the accuracy one can identify the underlying stochastic source field with is intrinsically limited. The method employed in this work allows one to distinguish between the intrinsic uncertainty of the physical system at hand (seaquakes) and that introduced by the identification procedure relying on limited knowledge of the physical system. As an illustration of the confidence one can have in the identified source description, the variance of S^{opt} solely arising from the uncertainty introduced by the finite accuracy of the sensor and the poor knowledge of the ocean depth field, *i.e.*, in the case $E\left(\boldsymbol{\zeta}^2\right) = 0$, is plotted in Fig. 12. The uncertainty brought to the identified source field is seen to be reasonably low in the present case. Its spatial distribution resembles that of the source field, while exhibiting a more pronounced elongation along the x-axis. This plot shows how accurate and quantitative the estimation can be over the whole physical domain.

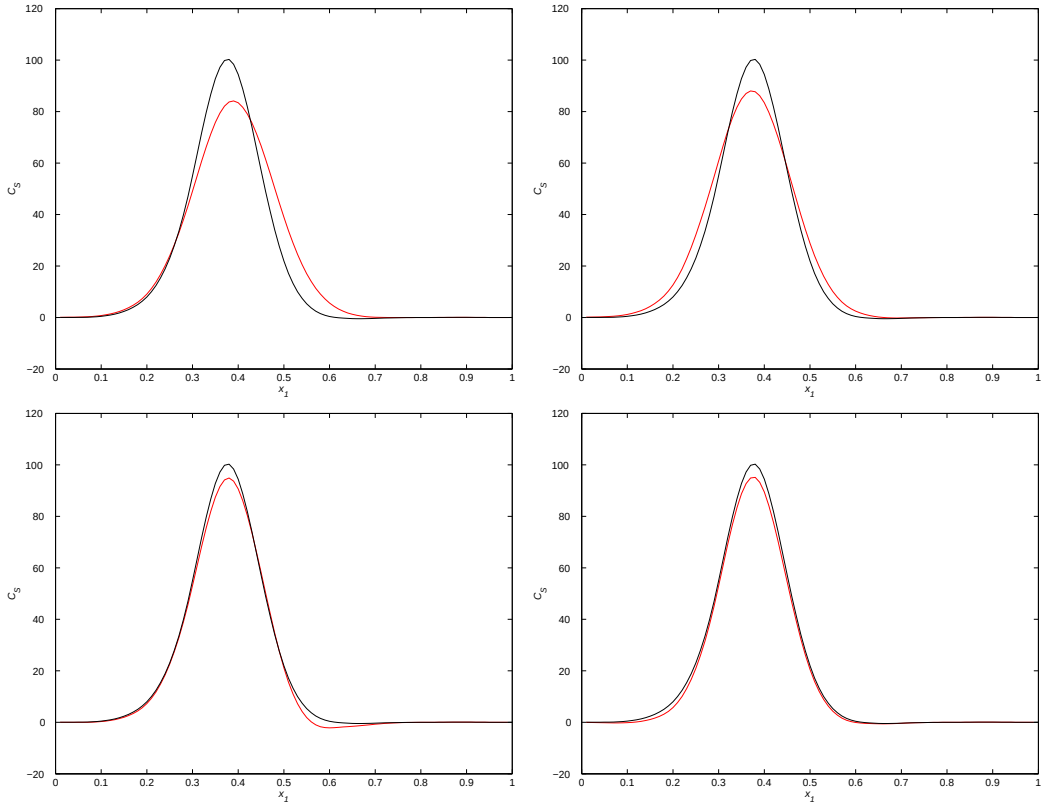

Fig. 10. Auto-covariance $C_S\,(x_1, x_2, y_1, y_2)$ 1-D field for different numbers N_{red} of modes in the reduced representation. From top left to bottom right, $N_{\text{red}} = 1, \ldots, 4$; $\nu = 2$; $p = 3$. $x_2 \simeq 0.38$, $y_1 = y_2 \simeq 0.62$. The auto-covariance estimated from the experimental data is plotted with a solid black line.

8 Concluding remarks

An efficient statistical estimation procedure has been proposed for use with realistic, uncertainty affected, data obtained from the indirect observations of statistically independent realizations of a random process that needs to be identified. The measurements are done on some quantity affected by the random process at hand and are subject to uncertainty arising from the experimental observation device and a poor knowledge of the physical model. For each statistical realization of the random process, a stochastic inverse problem is solved to evaluate the source field giving rise to that observation. A statistical description of the resulting set of source fields is then sought. A statistical reduction is performed to lower the dimensionality of the solution space and the random process to be identified is approximated under the form of a truncated Principal Component series. The resulting random vector is approximated with a Polynomial Chaos decomposition and an optimality principle, reminiscent of a maximum likelihood approach, is invoked to derive a good estimation of the PC development of the random vector components. To find the maximum of the manifold-constrained regularized cost function,

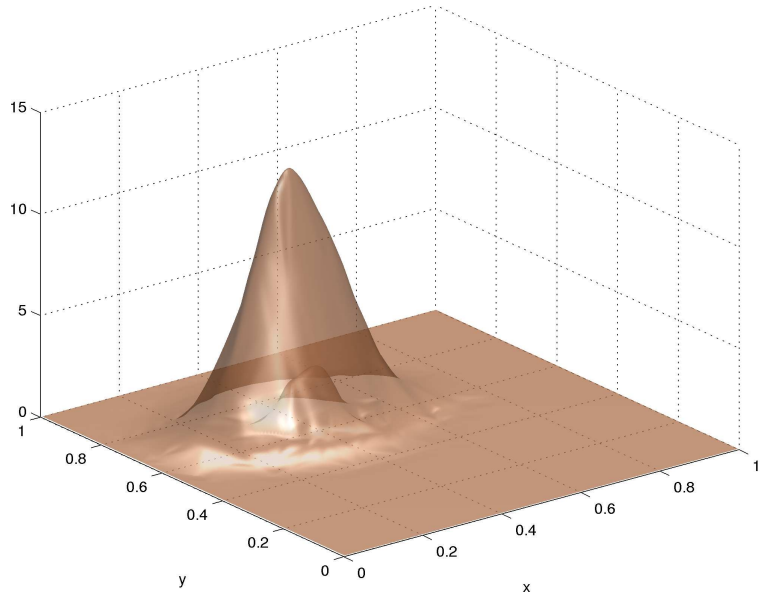

Fig. 11. Confidence interval for $S^{\mathrm{opt}}(\boldsymbol{x})$. The two nested surfaces plotted correspond to the 20% and 80% quantiles.

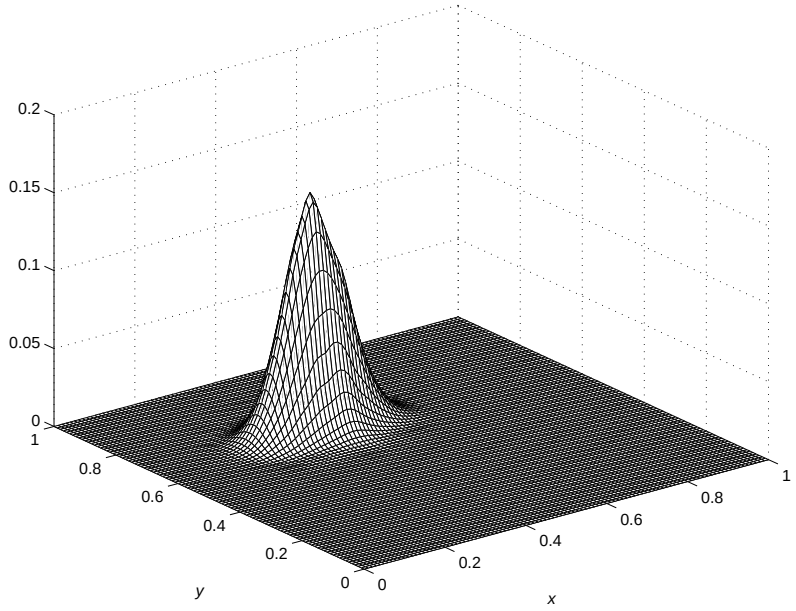

Fig. 12. Plot of the variance of $S^{\mathrm{opt}}(\boldsymbol{x})$ in the case $E\left(\boldsymbol{\zeta}^2\right) = 0$.

a Sequential Quadratic Programming algorithm was used, providing a second order convergence rate and allowing for non-linear, equality and/or inequality, constraints on the parameters to identify.

The methodology derived in this work was applied to the estimation of the stochastic properties of a seaquakes source. Input data were available from uncertainty subjected satellite observations of the ocean surface after each

seismic event and a statistical description of the ocean displacement field was determined. Results show that the statistical behavior of the identified random field (auto-covariance, probability density function) reproduces that given by empirical estimators from the experimental data.

Compared to the usual maximum likelihood-based method, and at the price of an additional effort in the development, the present inverse problem-based approach avoids the need for multiple resolutions of the forward model, a step that is CPU challenging for many problems of practical interest. Further, in addition to the estimation of the source field statistical properties, the precise uncertainty associated with this estimation is available and allows one to define, say, confidence bounds. Future developments of this methodology include improvements in the search of a global optimal estimation and an alternative, numerically more efficient, optimality principle.

Acknowledgement

The first two authors would like to thank Christian Soize for fruitful discussions. LM also gratefully acknowledges the financial support of the Office of the Provost during his stay at Florida State University, the US NSF and the French National Agency for Research under projects ASRMEI JC08#375619 and TYCHE (ANR- 2010-BLAN-0904).

References

Biros G. & Ghattas O., Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization. Part I: The Krylov-Schur solver, *J. Sci. Comput.*, **27** (2), p. 687–713, 2005.

Desceliers C., Ghanem R.G. & Soize C., Maximum likelihood estimation of stochastic chaos representations from experimental data, *Int. J. Numer. Meth. Engng.*, **66**, p. 978–1001, 2006.

Desceliers C., Soize C. & Ghanem R.G., Identification of chaos representations of elastic properties of random media using experimental vibration tests, *Comput. Mech.*, **39** (6), p. 831–838, 2007.

Douglas C.C., Haase G. & Iskandarani M., An additive Schwarz preconditioner for the spectral element ocean model formulation of the shallow water equations, *Elec. Trans. Numer. Anal.*, **15**, p. 18–28, 2003.

Eggermont P.P.B. & LaRiccia V.N., *Maximum Penalized Likelihood Estimation. Volume I: Density estimation*, Springer, 532 p., 2001.

Ghanem R.G. & Spanos P.D., Polynomial chaos in stochastic finite elements, *J. Appl. Mech.*, **57** (1), p. 197–202, 1990.

Ghanem R.G. & Spanos P.D., *Stochastic finite elements. A spectral approach*, rev. edn., Dover Publications, Inc., 222 p., 2003.

Gilbert J.C. & Lemaréchal C., Some numerical experiments with variable-storage quasi-Newton algorithms, *Math. Program.*, **45**, p. 407–435, 1989.

Hyvärinen A., Survey on Independent Component Analysis, *Neural Comput. Survey*, **2**, p. 94–128, 1999.

Iskandarani M., Haidvogel D.B. & Boyd J.P., A staggered spectral element model with application to the oceanic shallow water equations, *Int. J. Num. Meth. Fluids*, **20** (5), p. 393–414, 1995.

Koutsourelakis P.S., A multi-resolution, non-parametric, Bayesian framework for identification of spatially-varying model parameters, *J. Comput. Phys.*, **228** (17), p. 6184–6211, 2009.

Marzouk Y.M. & Najm H.N., Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems, *J. Comput. Phys.*, **228** (6), p. 1862–1902, 2009.

Marzouk Y.M., Najm H.N. & Rahn L.A., Stochastic spectral methods for efficient Bayesian solution of inverse problems, *J. Comput. Phys.*, **224** (2), p. 560–586, 2007.

Owen A. & Zhou Y., Adaptive importance sampling by mixtures of products of beta distributions, Stanford university, *Tech. Rep.*, 24 p., 1998.

Soize C. & Ghanem R.G., Reduced chaos decomposition with random coefficients of vector-valued random variables and random fields, *Comput. Meth. Appl. Mech. Eng.*, **198**, p. 21–26, 2009.

Wang J. & Zabaras N., A Bayesian inference approach to the inverse heat conduction problem, *Int. J. Heat Mass Trans.*, **47** (17-18), p. 3927–3941, 2004.

Webb A., *Statistical pattern recognition*, 2nd edn., Wiley, 496 p., 2002.

Zabaras N. & Ganapathysubramanian B., A scalable framework for the solution of stochastic inverse problems using a sparse grid collocation approach, *J. Comput. Phys.*, **227** (9), p. 4697–4735, 2008.