Comput Geosci
DOI 10.1007/s10596-016-9581-4

ORIGINAL PAPER

Quantifying initial and wind forcing uncertainties in the Gulf of Mexico

Guotu Li[1] · Mohamed Iskandarani[2] · Matthieu Le Hénaff[3,4] · Justin Winokur[1] · Olivier P. Le Maître[5] · Omar M. Knio[1]

Received: 14 October 2015 / Accepted: 26 May 2016
© Springer International Publishing Switzerland 2016

Abstract This study aims at analyzing the combined impact of uncertainties in initial conditions and wind forcing fields in ocean general circulation models (OGCM) using polynomial chaos (PC) expansions. Empirical orthogonal functions (EOF) are used to formulate both spatial perturbations to initial conditions and space-time wind forcing perturbations, namely in the form of a superposition of modal components with uniformly distributed random amplitudes. The forward deterministic HYbrid Coordinate Ocean Model (HYCOM) is used to propagate input uncertainties in the Gulf of Mexico (GoM) in spring 2010, during the Deepwater Horizon oil spill, and to generate the ensemble of model realizations based on which PC surrogate models are constructed for both localized and field quantities of interest (QoIs), focusing specifically on sea surface height (SSH) and mixed layer depth (MLD). These PC surrogate models are constructed using basis pursuit denoising methodology, and their performance is assessed through various statistical measures. A global sensitivity analysis is then performed to quantify the impact of individual modes as well as their interactions. It shows that the local SSH at the edge of the GoM main current—the Loop Current—is mostly sensitive to perturbations of the initial conditions affecting the current front, whereas the local MLD in the area of the Deepwater Horizon oil spill is more sensitive to wind forcing perturbations. At the basin scale, the SSH in the deep GoM is mostly sensitive to initial condition perturbations, while over the shelf it is sensitive to wind forcing perturbations. On the other hand, the basin MLD is almost exclusively sensitive to wind perturbations. For both quantities, the two sources of uncertainty have limited interactions. Finally, the computations indicate that whereas local quantities can exhibit complex behavior that necessitates a large number of realizations, the modal analysis of field sensitivities can be suitably achieved with a moderate size ensemble.

Keywords Polynomial chaos expansion · Empirical orthogonal function · Sensitivity analysis · Basis pursuit denoising

✉ Omar M. Knio
omar.knio@duke.edu

1 Department of Mechanical Engineering and Material Science, Duke University, Durham, NC 27708, USA

2 Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL 33149, USA

3 Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, FL 33149, USA

4 NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, FL 33149, USA

5 LIMSI-CNRS, Orsay, France

1 Introduction

Polynomial chaos (PC) methods [7, 14, 21, 25, 39, 41] have been developed in recent years for uncertainty quantification in a variaty of scientific and engineering fields, including chemical reaction systems [24, 29], fluid/ocean dynamics [1, 21, 23, 25, 36]. The main idea of PC methods is to approximate physical model response to uncertain inputs in terms of a series expansion, which involves orthogonal polynomials of variables parameterizing the uncertain inputs. The format of this representation readily affords various statistical analyses, e.g., Bayesian calibration of model

Published online: 11 June 2016

parameters [27, 34, 35], as well as global/local sensitivity analysis [1, 9].

The key task in building a PC surrogate of a quantity of interest (QoI) is to determine the series coefficients. Two major categories of PC methods exist, namely the intrusive and non-intrusive approaches. The former requires reformulation of existing computational models into systems involving the unknown expansion coefficients [23]. Non-intrusive methods, on the other hand, rely on sampling existing computational models; the PC coefficients are consequently determined based on an ensemble of deterministic model realizations. Depending on sampling strategy, non-intrusive PC reconstructions are often achieved either by quadrature-based pseudo spectral projection (PSP) methods [7, 8, 13, 39] or regression/compressed sensing (CS) type methods [3, 10, 11, 30]. Quadrature-based methods have shown promising performance in mitigating the so-called curse of dimensionality, which manifests itself as a rapid increase in the number of model realizations required to evaluate the coefficients as the polynomial order and the number of uncertain (stochastic) dimensions increase [23]. Quadrature-based approaches, however, can face several challenges, for instance when model solutions are subject to noise or random forcing, or when the computation model fails at specific ("extreme") values of the random inputs. In contrast, both regression and CS techniques allow occurrences of simulation failures at "extreme" sample points by either treating simulation failures as missing data, or restricting the range of the uncertain inputs.

The primary objective of this work is to analyze the combined impact of field uncertainties originating from both initial condition and wind forcing perturbations in the Gulf of Mexico (GoM). The GoM is a semi-enclosed ocean basin, dominated by the intense Loop Current (LC). The LC has a pathway that varies with time, from a retracted position to an extended one, where it reaches the northern GoM before turning southeastward toward the Atlantic Ocean [17]. When extended, the LC finally sheds a large, anticyclonic eddy, called LC Eddy, which drifts westward in the GoM while the LC retracts to the south. This LC Eddy shedding sequence is influenced by smaller cyclonic eddies at the edge of the LC, which can trigger temporary or final detachments of a LC Eddy [31]. The Deepwater Horizon oil spill in the GoM in Spring 2010, which is to date the largest oil spill in US history, was affected by this shedding sequence, as a LC Eddy named Eddy Franklin detached from the LC, shutting an export pathway to the south, while winds favored the transport of oil toward the northern GoM coasts [22].

Unlike the most common situation in uncertainty quantification where one focuses on a small number of scalar parameters [1, 34], in the present case both the initial conditions and wind forcing are field quantities. A straightforward approach (in which field variable at each spatial grid point is considered as one QoI) would necessitates a prohibitely large number of uncertain parameters (stochastic dimensions), especially in a high-resolution OGCM. It is thus essential to formulate both initial and wind forcing fields in such a way that stochastic dimensions are significantly reduced. Due to the inherent dynamical correlations among field quantities at different spatial locations and in space–time, a suitable strategy is to employ empirical orthogonal functions (EOFs) [18, 19, 26] to decompose both initial and wind forcing fields into a small number of modes, and associate with each mode an uncertain amplitude. As a result, the number of stochastic dimensions is significantly reduced, while adequate representation of the variability of the stochastic fields is still maintained. A similar EOF decomposition approach was employed in [36] for propagating boundary uncertainties.

Our analysis of the combined impact of initial conditions and wind forcing uncertainties relies on two quantities of interest, namely the sea surface height (SSH) and mixed layer depth (MLD) (additional details will be given in Section 2). For the purpose of propagating uncertainties, we initially attempted to apply an adaptive PSP method [7, 39] in order to generate a sparse realization ensemble (referred to as PSP ensemble) and consequently build PC surrogates. However, quadrature-based PSP construction of PC surrogates was precluded by the fact that the model yielded nonphysical MLD predictions when extreme values of the random inputs were sampled. As mentioned earlier, it is possible to remedy this situation by treating the corresponding samples as missing data and to build PC surrogates relying on a CS approach . Meanwhile, we also considered generating an independent ensemble via Latin hypercube sampling (LHS) [28]. Although the symptoms of simulation failure persisted, the CS approach provided a robust mean of constructing PC surrogates. Both PSP and LHS simulation ensembles led to surrogates that are in close agreement with each other, as well as faithful representations of the simulation data. For brevity, the present discussion shall be limited to results obtained from the ensemble generated via LHS only.

This paper is organized as follows. Section 2 outlines the model setup and specifies the random inputs. The approach used to construct the PC surrogates is outlined in Section 3. In Section 4, we focus on the analysis of the variability in the selected QoIs. We first assess the validity of PC representations, and then exploit these representations to assess sensitivities to the uncertain inputs. In Section 5, we generalize the sensitivity analysis to field quantities. An EOF decomposition method is developed for this purpose. In addition to the analysis of sensitivity fields, we exploit the EOF decomposition to conduct a brief assessment of

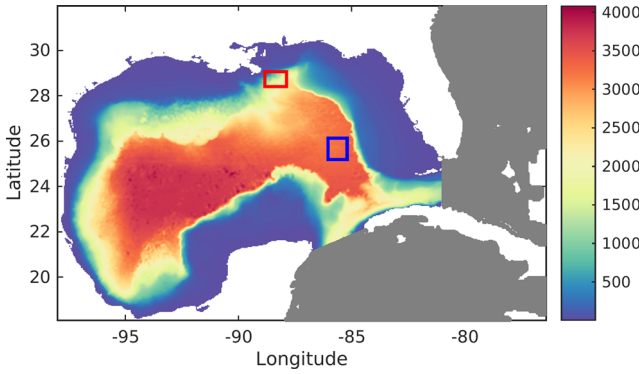

Fig. 1 Bathymetry of the Gulf of Mexico (meters), SSH (*blue box*) and MLD (*red box*) averaging domains.

the impact of ensemble size on the quality of stochastic predictions. Conclusions are discussed in Section 6.

2 HYCOM setup and uncertainties

The circulation in the GoM (Fig. 1) is simulated using the HYbrid Coordinate Ocean Model (HYCOM). HYCOM is an ocean general circulation model that uses a generalized vertical coordinate system to optimize the distribution of vertical computational layers; these layers are made to be isopycnic in stratified regions, terrain-following in shallow coastal regions, and isobaric in the unstratified mixed layer [4]. HYCOM serves a large user community who uses it for a wide variety of oceanic simulations[1]. The configuration adopted here is similar to the one used operationally by the US Navy for ocean prediction during the period 2003–2014. It has a 4-km horizontal resolution and 20 vertical levels. The computational domain is open along portions of its southern, eastern and northern boundaries, where values are provided by a lower-resolution (1/12° vs. 1/25°) simulation of HYCOM configured for the Atlantic Ocean (similar to [5]). The model is forced at the surface by three hourly outputs from the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMP [15]), which has a 27-km resolution. For the present study, model simulations with uncertain input parameters are run for 2 months, starting on May 1, 2010.

The PC methodology a functional relationship between the uncertain inputs and the quantities of interest, and two issues arise when contemplating uncertainties in field variables. The first concerns the identification of the random input variables that represent the uncertainties in the fields, and the second issue concerns the need to capture most of the uncertainty in the system while minimizing the number of uncertain variables that are needed to characterize this uncertainty. These issues can be effectively addressed by relying on EOF decompositions [18, 19, 26], that by construction capture the dominant variability modes of the system. The EOF modes are the eigenmodes of covariance matrices and are the finite-dimensional equivalent of a Karahunan-Loève modal decomposition. Furthermore, we identify the uncertain input variables as the amplitude of the EOF modes. In the following, we present the methodology adopted in constructing the covariance matrices for the initial conditions and wind forcing perturbations.

The selection of the optimal covariance matrices to characterize variability modes in geophysical systems remains an open question. Analysts, however, have considerable leeway in configuring the covariance matrices to target specific uncertainties. A primary interest in the present instance is in the local variability of frontal dynamics at the edge of the LC. The covariance matrices for the initial condition uncertainties were thus constructed empirically from 14 daily samples of the near-real-time HYCOM simulation of the GoM (performed at the Naval Research Lab in Stennis MS and predating the start of the present experiment). This relatively short period was found sufficient to capture localized variability in the LC front.

The EOF decomposition was performed simultaneously over a set of two model variables (see Appendix A for details). The first one is the 3D hydrostatic pressure increment within each HYCOM model layer, a quantity that is guaranteed to be defined at all times despite the hybrid nature of the HYCOM vertical coordinate system. The 3D pressure incorporates variability in the model vertical structure, which is associated with changes in heat and salt content, as well as in the dynamics (since horizontal gradients are associated with currents). The second variable is the model sea surface height which is a good proxy for the model surface dynamics. We use the principal components of each mode, which includes the temporal signature of the mode, to project the EOF modes obtained from the aforementioned decomposition to the actual model variables (temperature, salinity, layer thickness, velocity). In the present study, the four dominant EOF modes were retained in the initial conditions as they accounted for 68 % of the variability experienced by the model during this 14-day period (the percent of variance explained by each of the first four modes is 35, 15, 10, and 8 %). The amplitude of each EOF mode is a random variable defining a new dimension in the uncertain parameter space.

A similar procedure was followed to characterize the uncertainty in the wind forcing. The EOF analysis used the wind forcing fields from COAMP over a 60-day period, in May and June of 2010, i.e., during the study period. The EOF decomposition was performed over the wind vector in both horizontal directions, before being projected onto the wind amplitude and wind stress vectors, which are the

[1](See http://www.hycom.org for more information)

variables used to actually force the model. The reason for performing the EOF decomposition over the wind vector, and not the wind stress vector, is that the latter approach tends to emphasize events of large wind (or wind stress) amplitude, as the wind stress amplitude tends to evolve like the square of the wind amplitude. The EOF decomposition of the wind vector allows the identification of variability modes that are smoother in time (not shown). As for the perturbation of the initial conditions, we retain the first four space and time dependent dominant (scaled) modes. The first four modes account for 31.75, 23.22, 12.22, and 9.15 % of the total variance.

As a result, the initial condition and wind forcing fields are given as follows:

$$u(x, t = 0, \xi_a) = \bar{u}(x, \xi_a = 0) + \alpha_u[\sqrt{\lambda_1}\mathcal{U}_1, \sqrt{\lambda_2}\mathcal{U}_2, \sqrt{\lambda_3}\mathcal{U}_3, \sqrt{\lambda_4}\mathcal{U}_4]\xi_a^T \quad (1)$$

$$f(x, t, \xi_b) = \bar{f}(x, \xi_b = 0) + \alpha_f[\sqrt{\eta_1}\mathcal{F}_1, \sqrt{\eta_2}\mathcal{F}_2, \sqrt{\eta_3}\mathcal{F}_3, \sqrt{\eta_4}\mathcal{F}_4]\xi_b^T \quad (2)$$

where $\bar{u}$ and $\bar{f}$ are the unperturbed initial and wind forcing conditions; x denotes spatial coordinates; $\xi_a = [\xi_1, \xi_2, \xi_3, \xi_4]$ and $\xi_b = [\xi_5, \xi_6, \xi_7, \xi_8]$ are stochastic random vectors where all random variables ($\{\xi_i | i = 1, 2, ..., 8\}$) serve as amplitudes of corresponding perturbation modes and are assumed to be independent and uniformly distributed over $[-1, 1]$. (λ_i ,$\mathcal{U}_i$) and (η_i ,$\mathcal{F}_i$) are eigenvalue/eigenvector pairs of covariance matrices in initial and wind forcing EOF analysis respectively. Note that, for wind forcing perturbations, all EOF modes are time-dependent as well. Two scale factors ($\alpha_u = 0.8$ and $\alpha_f = 0.04$) are introduced to limit the amplitudes of perturbations in initial conditions and wind forcing respectively.

Two quantities of interest are the focus of the present study. The first is the SSH whose variations are a reflection of baroclinic and barotropic processes in the ocean, and whose slope is associated with ocean surface currents. SSH highs are usually associated with anticyclonic circulation, typically the LC and LC Eddies in the GoM, while SSH lows are associated with cyclonic circulation, in the GoM with LC frontal eddies. The second variable of interest is the MLD. The MLD defines the thickness (in meters) over which waters are homogeneous at the surface of the ocean. This homogeneity is mostly due to mixing by the surface winds, which tends to counteract the stratification generally caused by insolation at the surface of the ocean. In the present HYCOM simulation, the MLD is estimated as the thickness of the surface layer in which the water density is lighter than the surface density modified by a temperature decrease of 0.3 °C. Hence, the first variable of interest (the SSH) is essentially due to inherent oceanic dynamics, whereas the second one (the MLD) is essentially responding to local wind forcing. It is thus expected that the SSH will be sensitive to perturbations of the ocean state (typically of the initial conditions), whereas the MLD will be more sensitive to perturbations in atmospheric forcing.

3 Polynomial chaos surrogate

In this section, we detail the construction of the PC surrogate of a generic QoI, $Q(\xi)$, from the realizations $Q_{i=1,...,N_{\text{LHS}}}$ at the LHS points. For simplicity, we consider the case where Q is a real scalar. Since a HYCOM simulation depends on ξ, Q is a functional of ξ as well and is therefore a real-valued random variable. We shall assume that all QoIs considered are second-order random variables, that is

$$\mathbb{E}\left[Q^2\right] = \int_{\Xi} Q(\xi)^2 p(\xi) d\xi < +\infty. \quad (3)$$

Following the discussion in Section 2 above, ξ is an eight-dimensional random vector, whose components ξ_i are independent and identically distributed, with uniform distributions in $[-1, 1]$. Thus, the ξ_i's have joint density

$$p(\xi) = \begin{cases} 2^{-8} & \text{for } \xi \in \Xi, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Observe that the first four components of ξ are involved in the parameterization of the initial conditions, while the last four concern the wind forcing. We shall denote by $\Xi \doteq [-1, 1]^8$ the 8-dimensional hypercube, and by $L_2(\Xi, p)$ the space of second-order functionals in ξ, equipped with the inner product $\langle \cdot, \cdot \rangle$, $\forall Q, Q' \in L_2(\Xi, p)$,

$$\langle Q, Q' \rangle \doteq \mathbb{E}\left[QQ'\right] = \int_{\Xi} Q(\xi) Q'(\xi) p(\xi) d\xi, \quad (5)$$

and norm $\|Q\|_{L_2}^2 \doteq \langle Q, Q \rangle$.

3.1 Polynomial Chaos approximation

Because $Q \in L_2(\Xi, p)$, $Q(\xi)$ admits an infinite PC expansion of the form

$$Q(\xi) = \sum_{\alpha=0}^{\infty} q_\alpha \Psi_\alpha(\xi), \quad (6)$$

where the $q_\alpha \in \mathbb{R}$ are the expansion coefficients and the functions $\Psi_\alpha : \xi \in \Xi \mapsto \mathbb{R}$ are orthogonal multivariate polynomials in ξ. The orthogonality condition can be expressed as:

$$\langle \Psi_\alpha, \Psi_\beta \rangle = \begin{cases} \|\Psi_\alpha\|_{L_2}^2 & \text{if } \alpha = \beta, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Also, because the ξ_i's are independent and identically distributed in $[-1, 1]$, the Ψ_α are products of one-dimensional

Legendre polynomials [36]. We shall adopt the classical convention that $\Psi_0(\xi) = 1$. With this convention, we have:

$$\mathbb{E}[Q] = \sum_{\alpha=0}^{\infty} q_\alpha \langle \Psi_\alpha, 1 \rangle = q_0, \tag{8}$$

while

$$\mathbb{V}[Q] = \mathbb{E}\left[(Q - \mathbb{E}[Q])^2\right] = \sum_{\alpha,\beta=1}^{\infty} q_\alpha q_\beta \langle \Psi_\alpha, \Psi_\beta \rangle = \sum_{\alpha \geq 1} q_\alpha^2 \|\Psi_\alpha\|_{L_2}^2. \tag{9}$$

In practice, the PC expansion is truncated, yielding an approximation $\tilde{Q}$ of the quantity of interest:

$$Q(\xi) \approx \tilde{Q}(\xi) \doteq \sum_{\alpha=0}^{N_p} q_\alpha \Psi_\alpha(\xi). \tag{10}$$

We have denoted (N_p+1) the total number of terms retained in the truncated expansion. Classically, the truncation is made with respect to some polynomial order. However, due to the rapid increase in the basis size when increasing the polynomial order, the hyperbolic truncation introduced in [40] is employed in this study. Specifically, we use an isotropic polynomial basis which is the union set of three hyperbolic truncated basis given by maximum total order of $No = [6, 20, 35]$; the corresponding truncation parameters are $q = [1, 0.55, 0.42]$, respectively. The choice of this basis has been carefully analyzed for all the approximations constructed below. For this setting, the basis dimension is $N_p + 1 = 8695$. With the basis defined, the core task of the PC approximation is then determination of the expansion coefficients q_α's in Eq. 10.

3.2 Determination of PC expansions

Several methods are available for the determination of the PC expansion coefficients q_α in Eq. 10. As mentioned in Section 2, we generate an ensemble of realizations of Q at randomly selected values ξ_i ($\xi_i \in \mathcal{P}_{LHS}$, where $\mathcal{P}_{LHS}$ denotes the LHS set of random inputs). Let N_{LHS} denote the ensemble size (We choose $N_{LHS} = 798$ in this study to obtain faithful empirical estimations of various statistical measures), $Y = (Q_1 \cdots Q_{N_{\text{LHS}}})^T$ be the vector of realizations, $[\Psi] \in \mathcal{R}^{N_{\text{LHS}} \times (N_p+1)}$ denote the matrix with coefficients

$$[\Psi]_{i,\alpha} = \Psi_\alpha(\xi_i), \tag{11}$$

and $q \in \mathbb{R}^{N_p+1}$ denote the vector of unknown PC coefficients q_α. Thus, $[\Psi]q$ is the vector of predicted values for the QoI at the sample points; q is defined as the solution of the following basis pursuit denoising problem:

$$q = \arg \min_{c \in \mathbb{R}^{N_p+1}} \|c\|_{\ell_1} \quad \text{s.t.} \quad \|Y - [\Psi]c\|_{\ell_2} \leq \sigma ||Y||_{\ell_2}, \tag{12}$$

with ℓ_1 and ℓ_2 the classical 1 and 2 norms in $\mathbb{R}^m$. This optimization problem is solved using the SPGL1 algorithm [37, 38]. The ℓ_1-norm promotes the sparsity in the constructed PC model while $\sigma > 0$ is an error tolerance factor in the present context. The value of σ is adjusted in order to avoid over-fitting the realization data. The appropriate selection of σ is critical as we are considering a large PC basis while disposing of a limited number of realizations. To this end, we rely on K-fold cross validation procedure [10, 30], with $K = 5$, to determine the optimal value of σ (minimizing the estimated L_2-prediction error). The resulting predictions have been carefully analyzed and validated against alternative approaches and different basis selections.

3.3 Global sensitivity analysis

Though the uncertainty in initial conditions and in wind forcing are specified in terms of independent random variables, their effects may still combine in a complex manner because of the non-linear dynamics of the ocean model. Thus, we shall rely on variance-based sensitivity analysis methods [33] to investigate the global impact of uncertainties in initial condition and wind forcing. The approach selected is briefly outlined in this subsection.

Let i be a subset of $\{1, \cdots, 8\}$ and $i_\sim \doteq \{1, \cdots, 8\} \setminus i$ its complement. Following [16], we define the first-order and total-order sensitivity indices associated to i as

$$\mathbb{S}_i = \frac{\mathbb{V}[\mathbb{E}[Q|\xi_i]]}{\mathbb{V}[Q]}, \quad \mathbb{T}_i = 1 - \frac{\mathbb{V}\left[\mathbb{E}\left[Q|\xi_{i_\sim}\right]\right]}{\mathbb{V}[Q]}, \tag{13}$$

where $\mathbb{V}[\cdot]$ denotes the variance operator and $\mathbb{E}[Q|\xi_i]$ is the expectation of Q conditioned on the ξ_i for $i \in i$. Briefly, the interpretation of the indices is as follows [16]. The first-order index $\mathbb{S}_i$ measures the fraction of the variance of $Q(\xi)$ arising solely due to the random variables ξ_i with index $i \in i$. On the other hand, the total-order sensitivity index $\mathbb{T}_i$ is the fraction of variance due the variables ξ_i with index $i \in i$ as well as all their interactions with the others. When $\mathbb{S}_i$ is close to 1, this indicates that other variables with indices in $i_\sim$ have a low effect on Q. Additionally, one can conclude that variables with index in i are unimportant when $\mathbb{T}_i$ is close to zero. In the analysis below, we shall often report the global effects of the initial condition (that is sensitivity indices for $i = \{1, \ldots, 4\}$) or wind forcing ($i = \{5, \ldots, 8\}$), as the effect of individual variables is not as informative.

Note that the avalaibility of a PC representation of Q greatly simplifies the estimatation of the corresponding sensitivity indices. Specifically, the latter can be readily obtained from [1, 9]:

$$\mathbb{S}_i = \frac{\sum_{\alpha \in \mathcal{S}_i} q_\alpha^2 \|\Psi_\alpha\|_{L_2}^2}{\sum_{\alpha=1}^{\infty} q_\alpha^2 \|\Psi_\alpha\|_{L_2}^2}. \tag{14}$$

and

$$\mathbb{T}_i = \frac{\sum_{\alpha \in \mathcal{T}_i} q_\alpha^2 \|\Psi_\alpha\|_{L_2}^2}{\sum_{\alpha=1}^{\infty} q_\alpha^2 \|\Psi_\alpha\|_{L_2}^2}, \tag{15}$$

where $\mathcal{S}_i$ (resp. $\mathcal{T}_i$) is the set of polynomial indices $\alpha > 0$ such that Ψ_α has degree 0 in all the ξ_j with $j \notin i$ (resp. has degree > 0 in all ξ_j with $j \in i$). In other words, the indices are obtained from partial sums involving the coefficients of the PC expansions, and the norms of the basis elements.

4 Regionally averaged QoIs

In this section, we apply the PC methodology above in order to analyze the combined effects of initial condition and wind forcing uncertainties on two scalar QoIs, namely the SSH and MLD. Both quantities are regionally avaraged over Subdomains, as described in Section 4.1. We start in Section 4.2 by analyzing the performance of the PC representation, and exploit it in Section 4.3 to perform a global sensitivity analysis of the impact of the random inputs.

4.1 Definitions of QoIs

We refer to [4] and online HYCOM documentation[2] for a more detailed discussion of SSH and MLD. It is worthwhile to note that whereas SSH is a dynamic HYCOM variable, the MLD is a diagnostic quantity. The analysis in this section is based on two QoIs, obtained by (1) averaging SSH over the rectangular subdomain

$$\Omega_{\text{SSH}} = [-86.04°, -85.20°] \times [25.19°, 26.23°],$$

and (2) averaging MLD over the subdomain

$$\Omega_{\text{MLD}} = [-88.84°, -87.88°] \times [28.40°, 29.07°].$$

The two averaging domains are plotted in Fig. 1. As noted earlier, the former subdomain is located in an area associated with LC detachment, whereas the latter is near the DWH well.

4.2 Validation of the PC approximations

We start by illustrating the procedure of constructing PC approximations for our QoIs, focusing on the averaged SSH and MLD at day 30. Recall that the two QoIs are represented as:

$$\tilde{Q}^{\text{SSH,MLD}}(\xi) = \sum_{k=0}^{N_p} q_k^{\text{SSH,MLD}} \Psi_k(\xi), \tag{16}$$

where the respective PC coefficients are computed by solving the minimization problem in Eq. 12.

[2] http://www.hycom.org

Figure 2 reports the surface plots of the PC models of the regionally averaged SSH and MLD at day 30, in the plane (ξ_1, ξ_5) (all other random coordinates being set to 0). In addition to the PC model surfaces, several independent HYCOM simulations were performed for ξ in the considered plane and are reported using blue points. From these independent simulations, one can appreciate the fidelity of the PC approximations. It can be seen that most of HYCOM simulations of SSH closely agree with the PC model surface (top plot), demonstrating that $\tilde{Q}^{\text{SSH}}$ provides a reasonable prediction for ξ outside the construction set $\mathcal{P}_{\text{LHS}}$. The same conclusion holds for MLD surrogate model. We also observe that in the (ξ_1, ξ_5) plane the averaged SSH is essentially affine in ξ_1, whereas the averaged MLD appears to depend mostly on $|\xi_5|$. Such distinct behaviors highlight the need for global sensitivity analysis to properly characterize the various effects of the different random inputs.

The quality of PC models can be more quantitatively examined using empirical error measures. The empirical

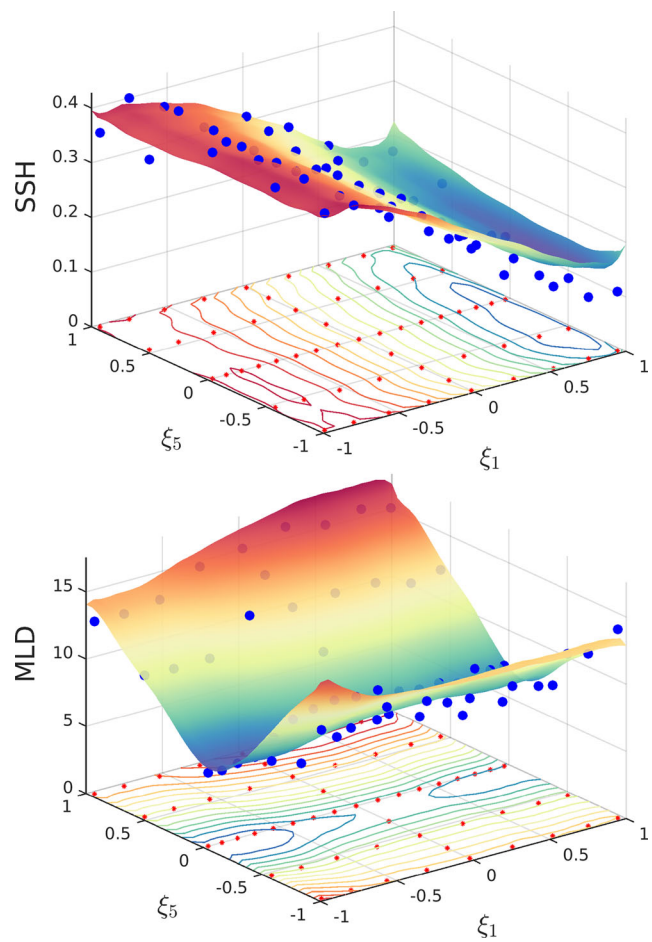

Fig. 2 Surface plots of PC models $\tilde{Q}(\xi)$ in the plan (ξ_1, ξ_5) all other coordinates being set to zero: (Top) SSH; (Bottom) MLD. Also shown using blue points are several HYCOM deterministic simulations in the same plan. Both SSH and MLD are in meters

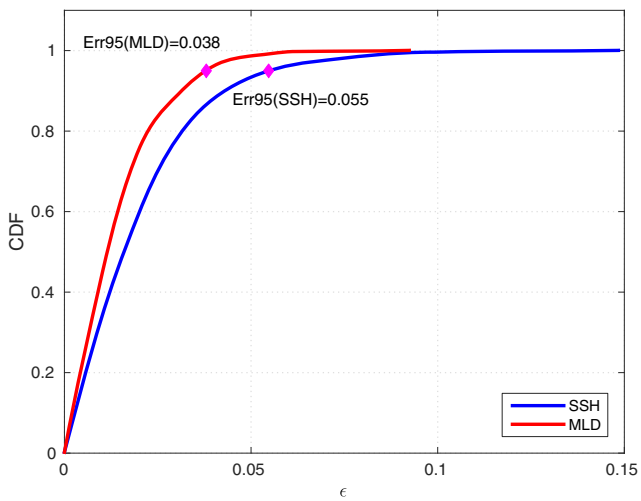

Fig. 3 Cumulative distribution functions of the relative local errors, see Eq. 17, for the regionally averaged SSH and MLD QoIs as indicated

errors for a QoI Q compare the PC predictions at the sample set points, $\tilde{Q}(\xi_i)$, to their corresponding HYCOM values Y_i. For $i = 1, \ldots, N_{\mathrm{LHS}}$ we then define the relative empirical error on Q according to:

$$\epsilon_i = \frac{|\tilde{Q}(\xi_i) - Y_i|}{\max_j(Y_j) - \min_j(Y_j)}. \tag{17}$$

Figure 3 shows the cumulative distribution functions of relative empirical errors ϵ_i for the two QoIs. It shows that in both cases the median relative empirical error is less than 2.5 % demonstrating the quality of the PC models. We further define the global accuracy measure (Err_{95}) using the relative empirical error level for which the CDF reaches 95 %, namely

$$\mathrm{Err}_{95} = \epsilon|_{CDF=0.95} \tag{18}$$

In other words, the criterion suggests that the relative error between $Q(\xi)$ and its PC approximation $\tilde{Q}(\xi)$ is less then Err_{95} with ≈ 0.95 probability. For SSH and MLD, the computed Err_{95} are 5.5 % and 3.8 %, respectively, confirming the quality of the constructed PC models. These findings are further confirmed by the results reported in Table 1 which compares the empirical (based on HYCOM realizations at $\xi_i \in \mathcal{P}_{LHS}$) and PC model standard deviations for the two QoIs. It is seen that the empirical and PC standard deviations again agree within approximately 4–5 %. Note the lower values for the PC standard deviations compared to the empirical ones, as expected from the model construction method which treats a fraction of the variability in $Q(\xi)$ as noise, and the fact that the number of realizations is insufficient to fully capture all the relevant features.

Table 1 Standard deviations of the regionally averaged SSH and MLD: empirical (from the N_{LHS} HYCOM simulations) and PC approximations

	Empirical	PC model
SSH (m)	0.1000	0.0956
MLD (m)	2.8873	2.7456

Further examination of the PC models for SSH and MLD is performed by plotting in Fig. 4 the probability density functions of $\tilde{Q}^{\mathrm{SSH}}(\xi)$ and $\tilde{Q}^{\mathrm{MLD}}(\xi)$. These densities are estimated by means of a classical kernel-density-estimation (KDE) method [32], from a large sample set of 10^5 ξ points drawn randomly in Ξ (blue curves). These PC densities are also contrasted with KDE estimates based on the HYCOM realizations Y_i (red curves) generated at the sample points $\xi_i \in \mathcal{P}_{\mathrm{LHS}}$. Overall the different densities qualitatively

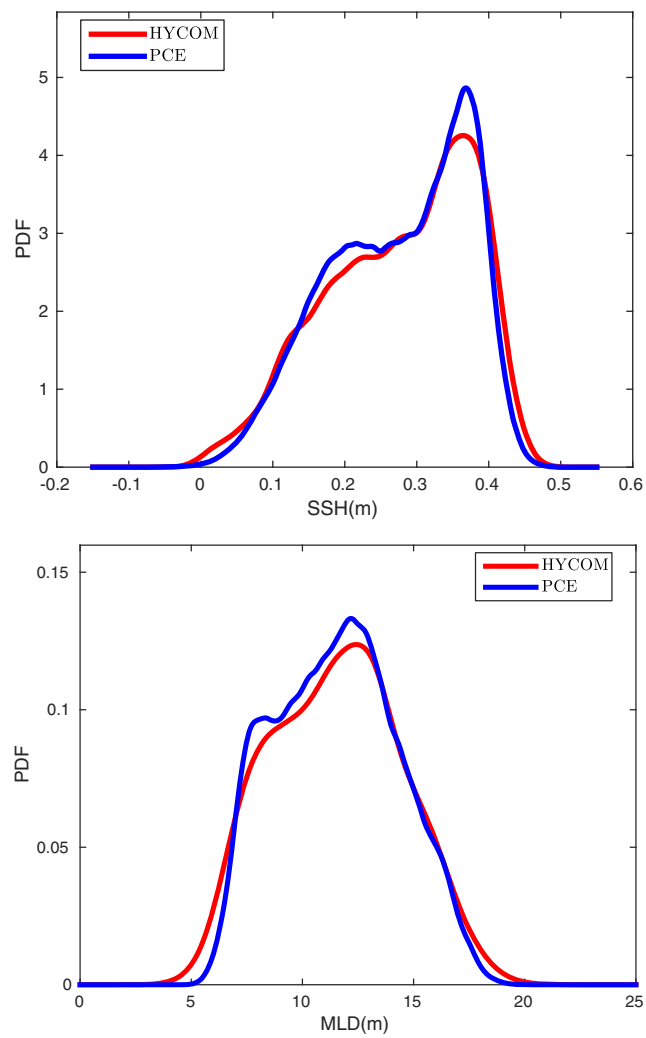

Fig. 4 Comparison of the SSH (*top*) and MLD (*bottom*) density functions estimated by KDE method. Empirical estimations from HYCOM realizations on $\xi_i \in \mathcal{P}_{\mathrm{LHS}}$ (*red curves*) and PC model predictions (*blue curves*) obtained by evaluating PC surrogates over a fine sampling of Ξ using 10^5 points

agree, with however shorter tails for the densities using the PC approximations $\tilde{Q}(\xi)$ (and consequently generally higher density peak(s)). Again, the shorter tails for the PC approximations, compared to the empirical densities, are expected due to the regularization which tends to smooth extreme values and features when they are not sufficiently observed by the realization ensemble.

All the analyses presented above were conducted using predictions obtained at day 30. These analyses were repeated for intermediate times, providing confidence in the ability of the PC approach to suitably model the essential trends in $Q^{\mathrm{SSH}}(\xi)$ and $Q^{\mathrm{MLD}}(\xi)$. A sample is provided in Fig. 5, which shows the evolution of Err_{95} (top plot) and compares the empirical and PC standard deviations for the two QoIs. Overall, the conclusions reached for the PC models at day 30 hold true at earlier times, though the error level can fluctuate significantly from one day to another. In particular, it is seen that for SSH Err_{95} seems to roughly increase in time, whereas for MLD it has a much noisy behavior. This is due to a more complex dynamics for MLD, which exhibits sharp variations in time; see for instance the time evolutions of the MLD-standard deviations in the bottom plot. Note that increasing the averaging domain Ω_{MLD} reduces the PC approximation error, as measured by Err_{95}, but sharp time variations can still be observed (not shown).

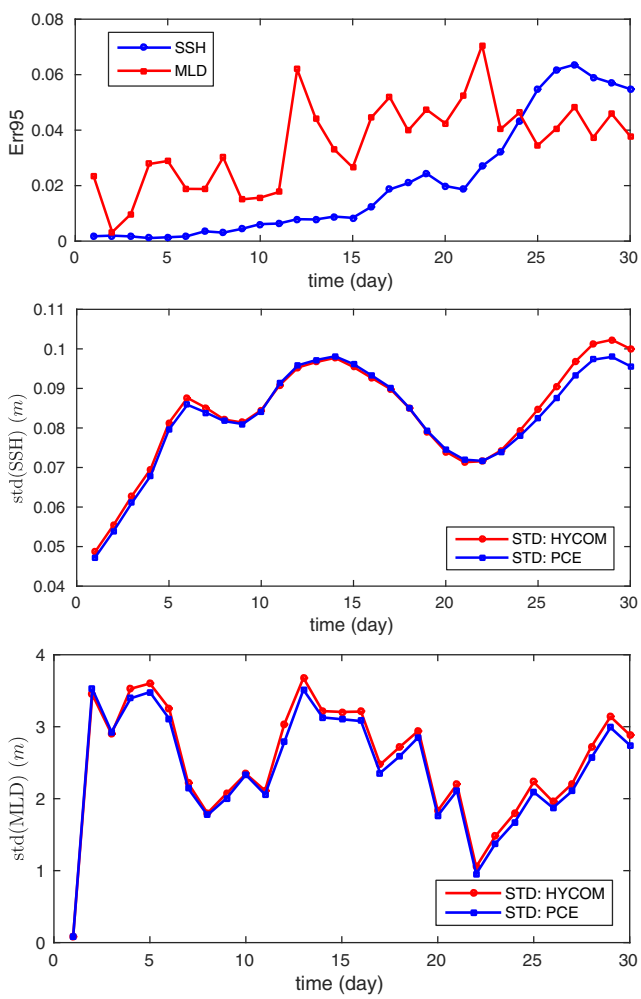

Fig. 5 Evolution in time of 95 % relative error estimates (Err_{95}, *top plot*) and comparison of the empirical and PC standard deviations for the SSH (*center plot*) and MLD (*bottom plot*)

4.3 Global sensitivity indices

The PC models for the two quantities of interest are now exploited to conduct a global sensitivity analysis. We start by discussing results obtained at day 30. The first and total sensitivity indices, computed from the PC expansion of $\tilde{Q}^{\mathrm{SSH}}$ and $\tilde{Q}^{\mathrm{MLD}}$ using Eqs. 14 and 15 with $i = \{i\}$ using $i = 1, \ldots, 8$, are reported in Fig. 6. For the regionally averaged SSH (top plot) we observe that all $\mathbb{T}_{\{i\}}$ are insignificant except for $i = 1$ and 2. It can then be concluded that the SSH is primarily influenced by the first two modes of the initial condition, while the wind forcing uncertainty has nearly no impact on the predicted SSH (at day 30). This finding has to be contrasted with the case of the regionally averaged MLD, for which $\mathbb{T}_{\{5\}}$ is clearly dominant, followed by much weaker effects reported for ξ_6 and ξ_1. Thus, the first wind forcing mode appears to be the main contributor to the MLD uncertainty. Further, the comparison of the total and first order indices reveals significant interactions between random sources within initial condition and wind forcing respectively, as one could have anticipated from the non-linearities of the model.

To simplify the sensitivity analysis, we now set $\mathrm{IC} = \{1, \ldots, 4\}$ and $\mathrm{WF} = \{5, \ldots, 8\}$, such that $\mathbb{S}_{\mathrm{IC}}$ and $\mathbb{T}_{\mathrm{IC}}$ (resp. $\mathbb{S}_{\mathrm{WF}}$ and $\mathbb{T}_{\mathrm{WF}}$) are the first and total order sensitivity indices associated with the uncertain initial condition

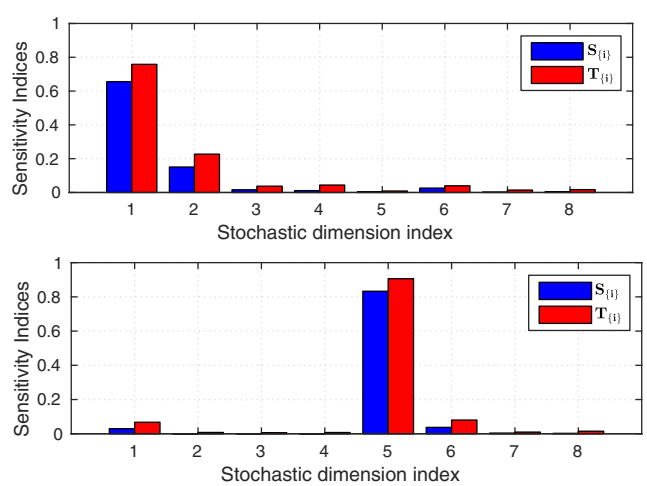

Fig. 6 First and Total order sensitivity indices associated to each input variables at day 30: (*Top*) SSH; (*Bottom*) MLD

(resp. uncertain wind forcing). Since no other source of uncertainty is considered, we have

$$\mathbb{T}_{\mathrm{IC}} + \mathbb{S}_{\mathrm{WF}} = \mathbb{T}_{\mathrm{WF}} + \mathbb{S}_{\mathrm{IC}} = 1, \tag{19}$$

and

$$\mathbb{T}_{\mathrm{IC}} - \mathbb{S}_{\mathrm{IC}} = \mathbb{T}_{\mathrm{WF}} - \mathbb{S}_{\mathrm{WF}} = \mathbb{I}_{\mathrm{IC}\times\mathrm{WF}}. \tag{20}$$

In the previous equation, we have denoted $\mathbb{I}_{\mathrm{IC}\times\mathrm{WF}}$ the sensitivity index, which measures the fraction of the variance due to the interaction between the uncertainties in the initial conditions and in the wind forcing. An alternative expression for this additional index is $\mathbb{I}_{\mathrm{IC}\times\mathrm{WF}} = 1 - \mathbb{S}_{\mathrm{IC}} - \mathbb{S}_{\mathrm{WF}}$.

With these definitions, we finally examine the evolutions in time of sensitivities of the regionally averaged SSH and MLD responses to the different sources of uncertainty. These evolutions are shown in Fig. 7. From the top plot, it is evident that essentially only the uncertain initial condition plays a role in the variability of the regionally averaged SSH, since both the corresponding $\mathbb{S}_{\mathrm{WF}}$ and $\mathbb{I}_{\mathrm{IC}\times\mathrm{WF}}$ remain very low over the whole time span reported. On the contrary, for the regionally averaged MLD response, we observe that at the beginning of HYCOM simulations, the initial condition is dominant (but the variance is then very low, see bottom plot of Fig. 5) as it take some time for the forcing to affect the flow. However, after day 2, the wind forcing is the clearly dominant source of uncertainty in the regionally averaged MLD prediction. Note that at day 22, the interaction between the two sources of uncertainty becomes

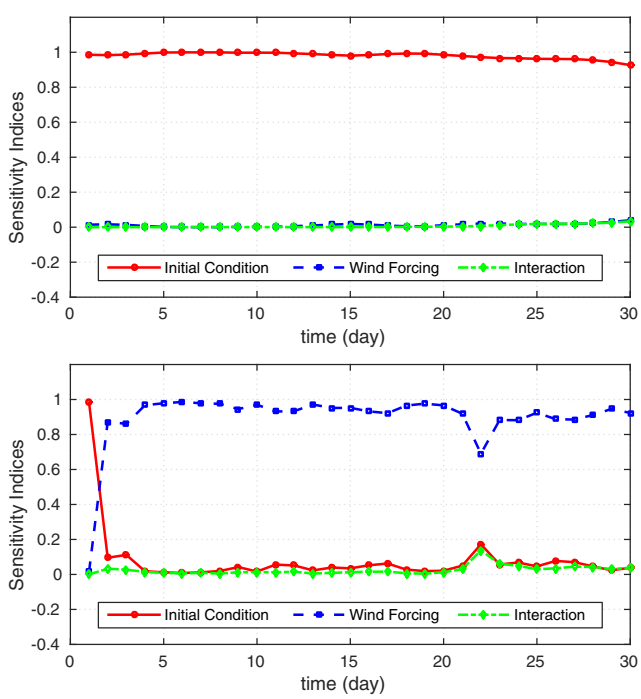

Fig. 7 Time evolutions of the sensitivity indices $\mathbb{S}_{\mathrm{IC}}$, $\mathbb{S}_{\mathrm{WF}}$ and $\mathbb{I}_{\mathrm{IC}\times\mathrm{WF}}$: (*Top*) SSH; (*Bottom*) MLD;

significant, but again this event occurs at a time when the variance in the averaged MLD achieves a low value.

It is worth pointing out that all sensitivity indices presented so far concern the SSH and MLD responses averaged over two spatial locations, namely the LC frontal area and the DWH region, respectively, as discussed in Section 4.1, and for a limited time horizon of 30 days. As a result, the conclusion on the weak interaction between initial condition and wind forcing uncertainties regarding the two QoIs may not hold for other averaging domains and for other times. In the specific configuration of the LC during our study period, it seems that local conditions in the DWH area during May–June 2010 are not sensitive to the uncertainties in LC frontal conditions on May 1st. Clearly, a more complete investigation is needed to better understand the dynamics of the uncertainties and the way they propagate in time through the domain. To this end, in the next section we investigate field uncertainties in SSH and MLD throughout the GoM.

5 Field sensitivities

Though it is possible to construct the whole SSH (or MLD) field over the GoM by building PC surrogates for field variables at each spatial grid independently, the computational cost would make this brute-force approach impractical due to the large number of grid points. As a result, the EOF-PC analysis is employed here to reduce the dimensionality of the field PC reconstruction problem. For clarity, we outline the EOF-PC approach below.

5.1 Decomposition and approximation

We first briefly discuss the EOF decomposition (also known as Proper Orthogonal Decomposition, Karhunen-Loeve Expansion, and Principal Component Analysis) [18, 19, 26]. Let $x \in \Omega$ and $U(x,\xi)$ be a real-valued second order stochastic process. We denote $(\cdot,\cdot)_\Omega$ the spatial inner product,

$$(u, v)_\Omega \doteq \int_\Omega u(x)v(x)\rho(x)dx, \tag{21}$$

with the spatial weighting function $\rho : \Omega \mapsto \mathbb{R}_+$, and define $\|\cdot\|_{L_2(\Omega)}$ the associated norm. From the N_{LHS} realizations of U at the sample points $\xi_j \in \mathcal{P}_{\mathrm{LHS}}$, we define the empirical average $\overline{U}(x)$ and the fluctuations $U'_j(x)$ as

$$\overline{U}(x) \doteq \frac{1}{N_{\mathrm{LHS}}}\sum_{j=1}^{N_{\mathrm{LHS}}} U(x,\xi_j), \quad U'_j(x) \doteq U(x,\xi_j) - \overline{U}(x),$$

and construct the matrix $[C] \in \mathbb{R}^{N_{\mathrm{LHS}}\times N_{\mathrm{LHS}}}$ of empirical (spatial) covariance, through

$$[C]_{i,j} = \left(U'_i, U'_j\right)_\Omega, \quad 1 \le i, j \le N_{\mathrm{LHS}}. \tag{22}$$

We observe that $[C]$ is symmetric, non-negative, and we denote $\lambda^k \in \mathbb{R}$ and $\phi^k \in \mathbb{R}^{N_{\text{LHS}}}$ its proper-elements such that

$$[C]\phi^k = \lambda^k \phi^k, \quad (\phi^k)^{\text{T}}(\phi^k) = 1. \tag{23}$$

The empirical spatial modes are subsequently defined as

$$u^k(x) = \sum_{j=1}^{N_{\text{LHS}}} U'_j(x)\phi^k_j. \tag{24}$$

It can be easily checked that the modes u^k are mutually orthogonal:

$$\left(u^k, u^l\right)_\Omega = \sum_{i,j=1}^{N_{\text{LHS}}} \phi^k_i \left(U'_i, U'_j\right)_\Omega \phi^l_j = (\phi^k)^{\text{T}}[C]\phi^l = \lambda^l \delta_{k,l}.$$

Ordering the eigenvalues as $\lambda^1 \geq \lambda^2 \geq \cdots \geq 0$, the truncated expansion of the stochastic field becomes

$$U(x,\xi) \approx \overline{U}(x) + \sum_{k=1}^{r} u^k(x)\phi^k(\xi), \tag{25}$$

where the $\phi^k(\xi)$ are uncorrelated, centered, second-order random variables and $r \leq N_{\text{LHS}}$ is the number of modes retained in the reduced representation. The number of modes is selected such that Eq. 25 retains a fraction $p \in [0, 1]$ of the empirical fluctuation energy. Specifically, we set r such that

$$\sum_{k=1}^{r} \lambda_k \geq p \sum_{k=1}^{N_{\text{LHS}}} \lambda_k. \tag{26}$$

It now remains to construct the PC approximations for the set of random coefficients $\phi^k(\xi)$. These random coefficients are treated as independent quantities of interest and each $\phi^k(\xi)$ is associated with a vector Y of realizations (see Eq. 12) defined as

$$Y_j = \phi^k(\xi_j) = \left(U'_j, u^k\right)_\Omega / \lambda^k, \quad 1 \leq j \leq N_{\text{LHS}}. \tag{27}$$

By using the method discussed in Section 3, namely the BPDN algorithm, each $\phi^k(\xi)$ can be approximated by a truncated PC expansion as follows:

$$\phi^k(\xi) \approx \tilde{\phi}^k(\xi) = \sum_{\alpha=0}^{N_p} c^k_\alpha \Psi_\alpha(\xi) \tag{28}$$

As a result, the EOF-PC approximation of the stochastic field can eventually be expressed as

$$U(x,\xi) \approx U^{PC}(x,\xi) \doteq \overline{U}(x) + \sum_{\alpha=0}^{N_p} u_\alpha(x)\Psi_\alpha(\xi), \tag{29}$$

where

$$u_\alpha(x) = \sum_{k=1}^{r} u^k(x) c^k_\alpha. \tag{30}$$

5.2 EOF-PC at day 30

In this section, we investigate the use of the EOF-PC approximation for the representation of the SSH and MLD fields. We used a constant spatial weighting $\rho(x) = 1$, but restrict the domain of interest to the inside of the GoM disregarding the grey-areas in the next figures. Similar to the regionally averaged quantities of interest, we restrict the analysis to the quality of the SSH and MLD EOF-PC approximations at day 30 only; similar trends were observed for earlier times (not shown).

5.2.1 SSH field

Starting with the SSH field at day 30, Fig. 8 presents the empirical average SSH field from realizations at the LHS set $\mathcal{P}_{LHS}$ (top plot). That figure illustrates the dominant dynamical features, with the LC associated with high SSH in the Eastern GoM, as well as a LC Eddy at (26° N, 94° W) , while small cyclonic frontal eddies are noticed at the edge of the LC, associated with low SSH. An isolated cyclonic eddy is also present at the southwest corner of the GoM. The decay of the spectrum (eigenvalues λ^k) in the empirical decomposition of the SSH fluctuation is then shown in the bottom plot of Fig. 8. In the present case, $r = 10$ empirical

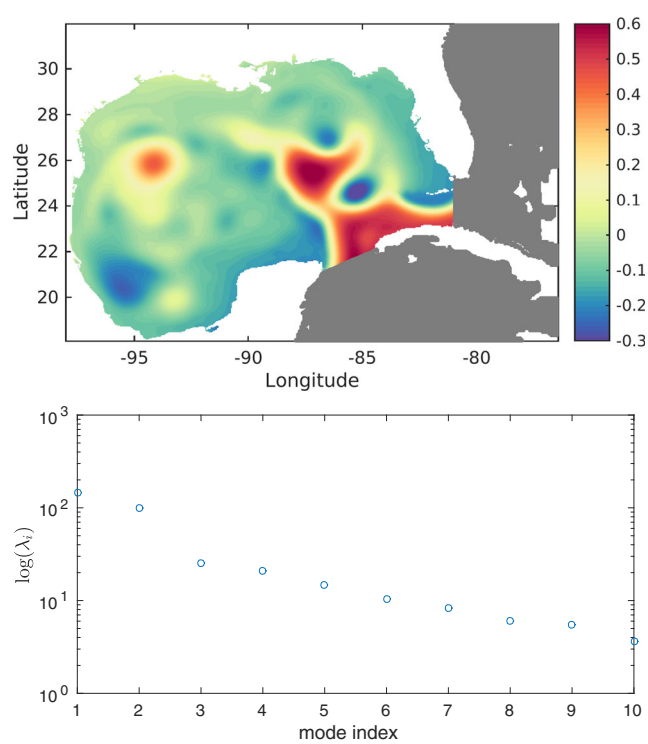

Fig. 8 SSH field at day 30. *Top*: empirical average using the LHS set of HYCOM realizations. *Bottom*: spectrum of the empirical spatial covariance

modes were found sufficient to retain 90 % of the energy fluctuation according to Eq. 26.

The first five spatial modes $u^k(x)/\sqrt{\lambda^k}$ of the decomposition of the SSH field fluctuation are shown in Fig. 9. Modes 1, 3, and 4 mostly resolve structures in the LC area, while mode 2 is dominated by coastal dynamics along the Northern GoM. Mode 5 mixes signatures in the LC area and along the coast. Following modes (not shown) exhibit features at smaller and smaller scales, with more and more homogeneous energy distribution in the domain analysed. In other words, the truncation principally affects small scale fluctuations with low energy levels.

The impact of the truncation on the field energy can be better appreciated in Fig. 10 where the standard deviation of the SSH field at day 30 is reported. The figure compares the empirical standard deviation estimated from the LHS ensemble of HYCOM simulations (left plot) with the standard deviation obtained from truncated EOF-PC approximation of the SSH field (right plot). These two plots demonstrate that with $r = 10$ modes only, the fluctuation of the field is very well approximated, particularly in areas of high energy fluctuation, namely the LC and north and west coastal areas. Such a low number of modes capable of reconstructing a field is quite common when EOF decompositions are used for analyzing ocean data (e.g.,[2]).

Finally, to briefly illustrate the effect of relying on a finite set of HYCOM realizations and using a PC approximation for the random coordinates, we present in Fig. 11 a comparison between the HYCOM solution and its

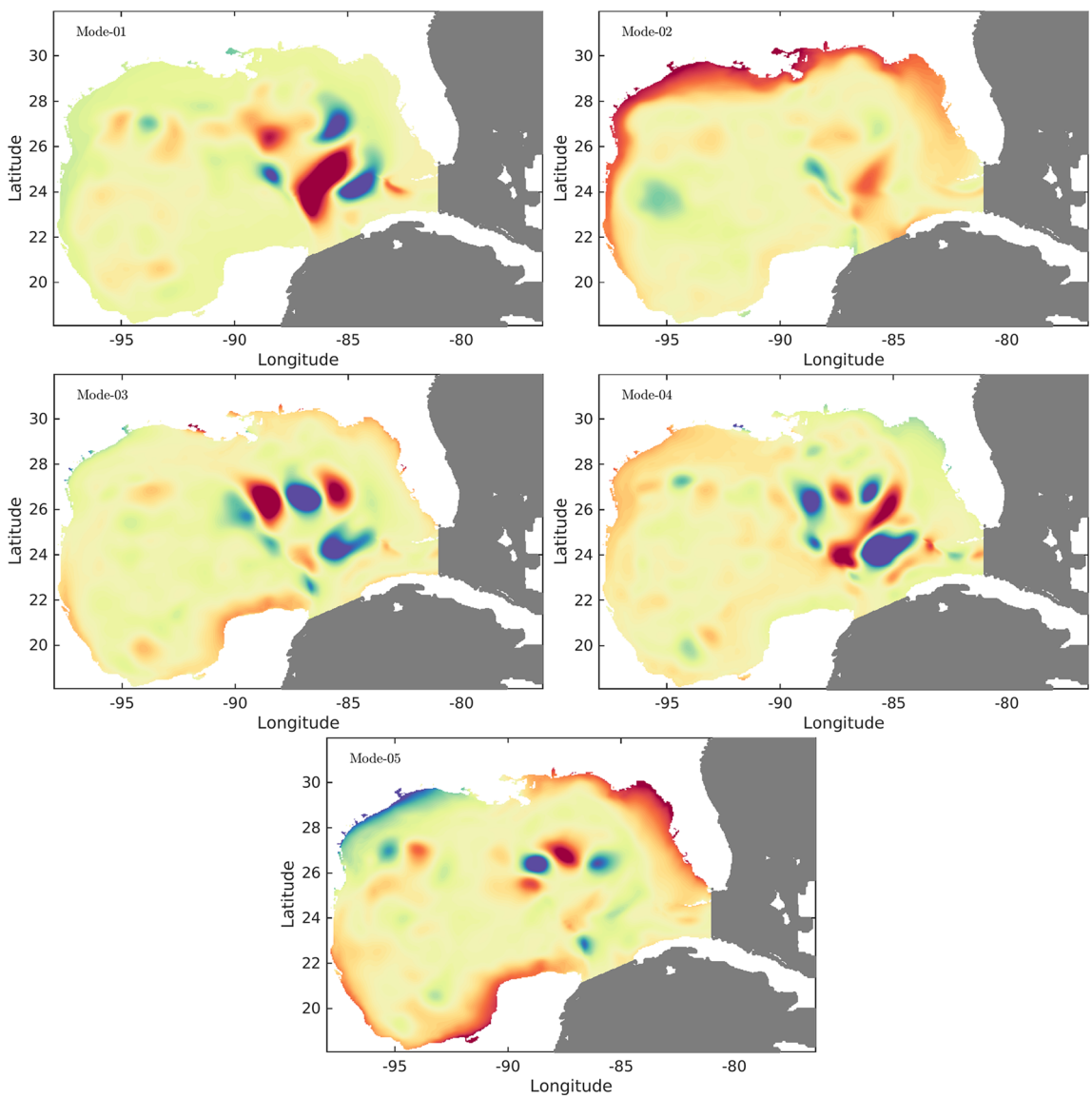

Fig. 9 First five spatial modes $u^k/\sqrt{\lambda^k}$ in the expansion of the SSH field at day 30

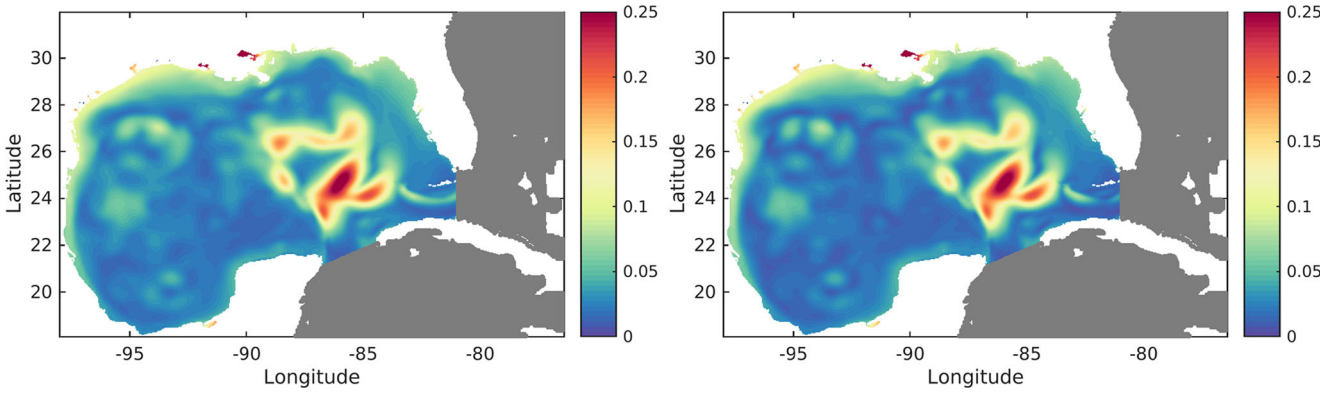

Fig. 10 Standard deviation of SSH fields at day 30. *Left*: empirical standard deviation calculated from the LHS set of HYCOM simulations. *Right*: standard deviation in the truncated expansion of the field using $r = 10$ modes

EOF-PC approximation for $\xi = 0$. Note that this particular realization of the HYCOM simulation is not part of the LHS sample set used in the reconstruction. In order to highlight the differences between both fields, the figures only show the fluctuations of SSH (with respect to the empirical average from the LHS simulation ensemble). As expected, the HYCOM realization (left plot) presents small-scale features of low-amplitude that are not present in the EOF-PC approximation (right plot); the latter field appears much smoother. However, it is seen that the principal structures, particularly the local SSH lows associated with LC frontal eddies, are well captured. Of course, this particular realization may not be representative of the approximation error for other realizations of ξ, but it nonetheless provides a clear illustration of the smoothing effect of truncation.

5.2.2 MLD field

We now repeat the analysis of the EOF-PC approximation of the previous section, but for the MLD field at day 30. The empirical MLD average on day 30, seen on Fig. 12 (top plot) shows that the MLD is deeper in the LC region than that in the rest of the GoM, and it tends to be shallower along the coast. As for the small-region average, it is found that the MLD field is significantly more complex and more demanding to approximate. Specifically, Fig. 12 shows that though the empirical average of the MLD field (top plot) is relatively smooth, the decay rate of the perturbation spectrum is significantly slower than that of the SSH field (bottom plot). Specifically, $r = 142$ modes are necessary to retain 90 % of the empirical variance. Because faithful PC recovery of higher-order modes would require a larger ensemble than is practical, in this analysis below the EOF expansion for MLD is limited to retain 80 % of field variability, which corresponds to $r = 27$ modes.

The first five dominant modes of the MLD covariance are plotted in Fig. 13. As for the SSH decomposition, Mode 1 is dominated by variability in the LC, which is also the case for Mode 4 to a large extent. Modes 2, 3, and 5 are mixed, with signals in the deep GoM as well as along the coasts. Compared with the dominant modes in SSH (Fig. 9), the dominant MLD modes involve shorter scale features, and tend to be less spatially localized.

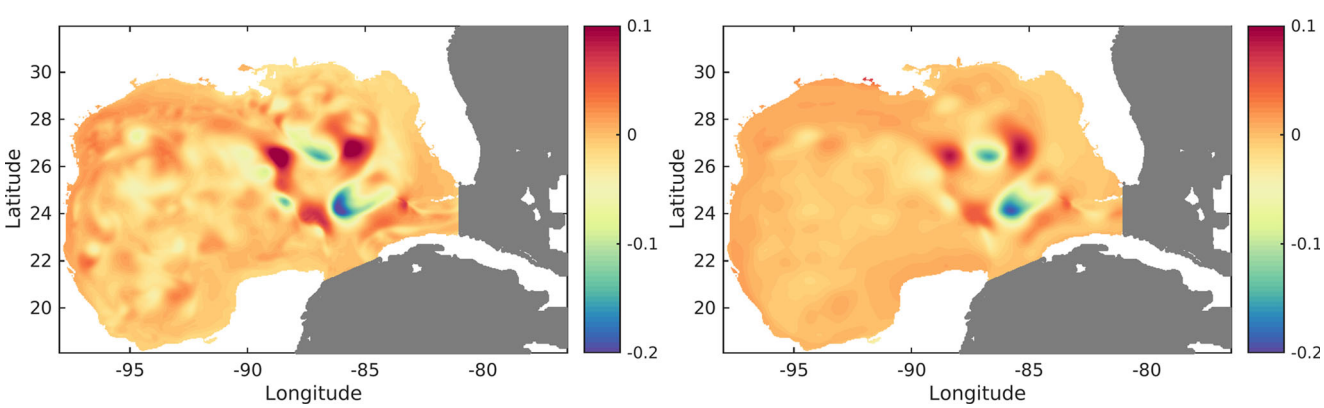

Fig. 11 SSH fluctuation fields for realization at $\xi = 0$ at day 30. *Left*: target fluctuation field. *Right*: EOF-PC reconstruction of the target fluctuation field

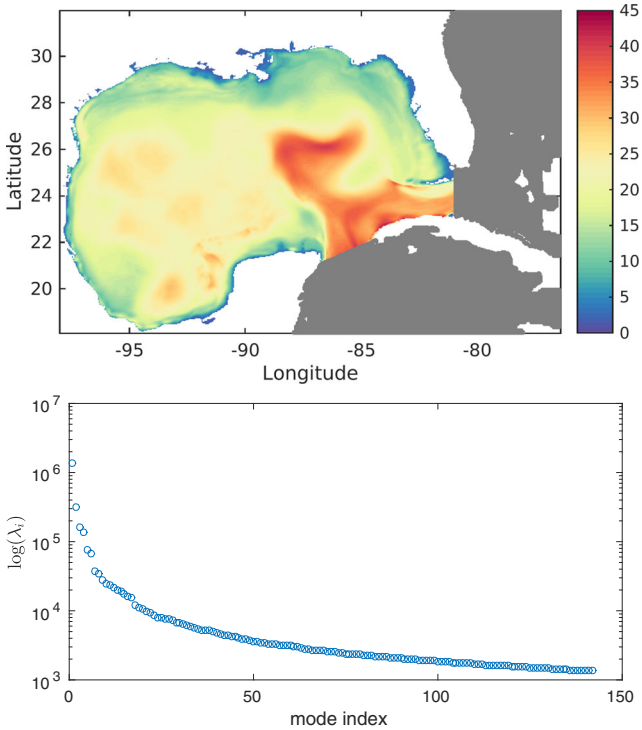

Fig. 12 MLD field at day 30. Empirical average using the LHS set of HYCOM realizations (*Left*) and spectrum of the empirical spatial covariance (*Right*)

In Fig. 14 we compare the empirical standard deviation of the MLD field (left plot) with the standard deviation of its EOF-PC approximation (right plot). As for the SSH, it is seen that the EOF-PC approximation is able to properly capture the main structures of the MLD uncertainty, in particular (1) near the LC region, (2) along the northern coastline, and (3) around the LC Eddy at (26° N, 94° W). However, we also observe that the differences between the two standard deviation fields are more significant than in the case of SSH (Fig. 10). In fact, the EOF-PC approximation contains only 80 % of the fluctuating energy in the set of HYCOM realizations. It is noted that inclusion of additional modes in the EOF-PC approximation would slowly improve the capture of the remaining field variability; again, this can be explained by the complex response of the local MLD to random inputs, which makes it difficult to approximate the reduced random coordinates $\phi^k(\xi)$. Indeed, the PC approximation of the ϕ^k yields an additional loss of variability, as some fluctuations are interpreted as realization noise by the PC construction procedure.

It should be stressed that though part of the variability in the MLD field is lost, the main structures in the empirical standard deviation field given by HYCOM realization ensemble are still present in the EOF-PC approximation and we shall see below that the information retained is enough to perform a suitable sensitivity analysis. However, the error in the EOF-PC approximation can be significant, in particular if it is used to predict specific realizations of the HYCOM simulation. This can be appreciated in Fig. 15 where the difference between the true HYCOM fluctuation at $\xi = 0$ (left plot) and its EOF-PC approximation (right plot) is much more significant than for the case of the SSH at the same parameter value (Fig. 11).

5.3 Global sensitivity analysis

From the EOF-PC approximation, one can easily compute field values for the sensitivity indices of SSH and MLD. Similar to the discussion in Section 4, we focus on the first order sensitivity indices with respect to the initial condition and wind forcing (as well as the contribution to the variance of the interaction between them).

Figure 16 reports the first order sensitivities at day 30 of the SSH (left column) and MLD (right column) fields to initial condition (top row), wind forcing (center row) and their interactions (bottom row). It can be seen from Fig. 16 that the SSH field is more sensitive to the initial condition almost everywhere except for the near shore area where wind forcing exhibits significant impact. This strong impact of wind forcing on the near shore SSH can be understood from the following facts: (1) the variability of the initial condition is concentrated in the LC region making initial SSH perturbation along coastline negligibly small; (2) The propagation over a 30 days time horizon of the initial uncertainty away from the LC area is limited; and (3) Wave amplitude is naturally more sensitive to wind forcing, when waves propagate from deep water to shallow water, and when the wave amplitude grows and thus amplifies the sensitivity to wind forcing near the coastline.

In contrast, the MLD field values at day 30 seem essentially sensitive to the wind forcing perturbations, which is expected since the MLD is primarily related to the turbulent mixing process in the near surface layer, where both momentum and heat fluxes are directly influenced by wind forcing perturbations. The initial condition impacts the MLD only in the LC region, where the perturbations were localized. Finally, one can notice that both SSH and MLD fields exhibit weak interactions between initial condition and wind forcing perturbations at the considered analysis day.

Finally, we quantify the time evolution of 1st order sensitivities of both SSH and MLD fields to initial condition, wind forcing and their interactions. Figure 17 depicts at days 10, 20, and 25 (from top to bottom) the first order sensitivity indices associated with the initial condition impact on the SSH (left) and to the wind forcing impact on the MLD. Because the interactions between initial condition and wind forcing remain relatively insignificant over the simulation span, the sensitivity indices associated with the

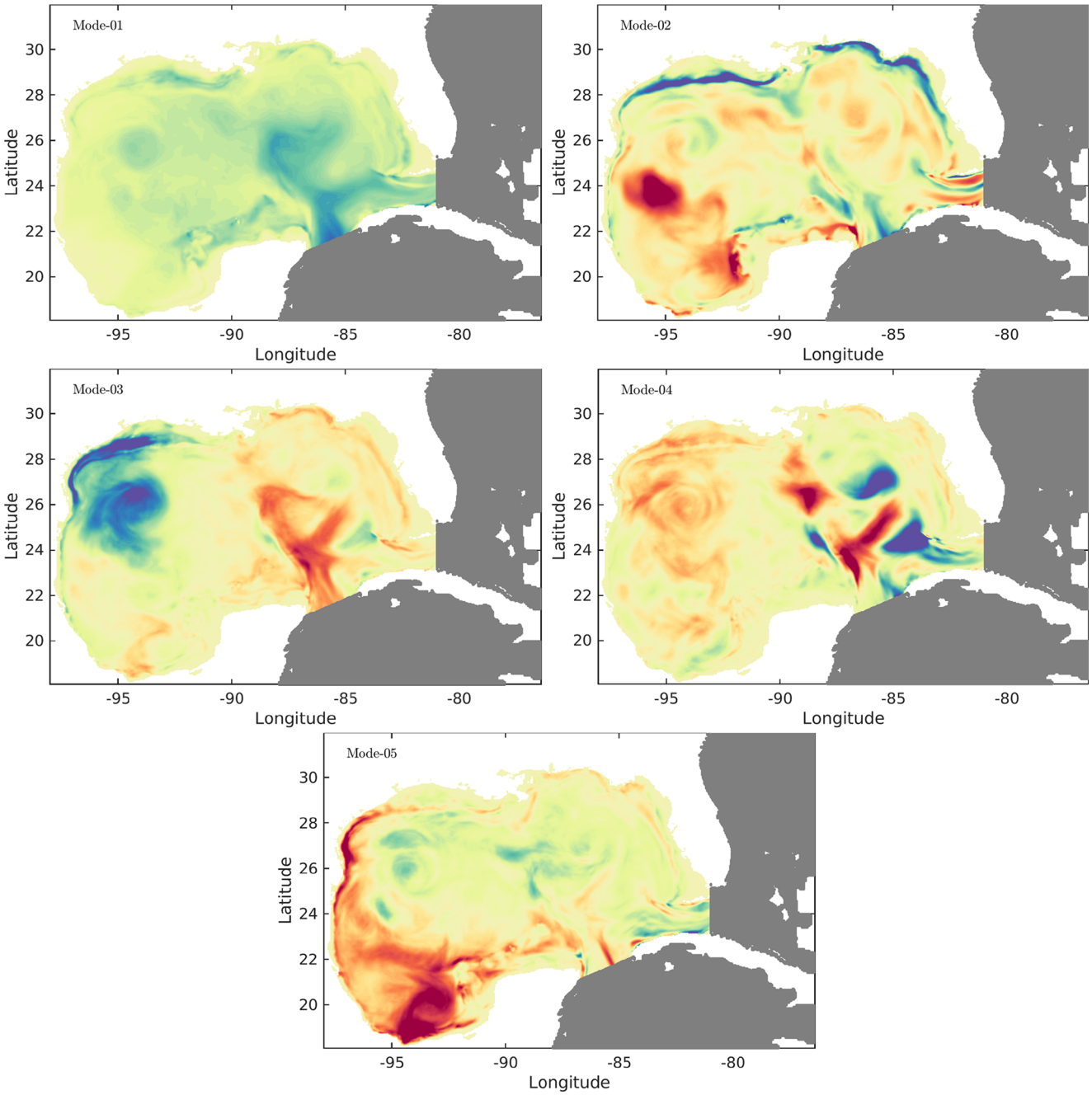

Fig. 13 First five spatial modes $u^k/\sqrt{\lambda^k}$ in the expansion of the MLD field at day 30

impact of wind forcing (resp. initial condition) on the SSH (resp. MLD) field can be estimated using $\mathbb{S}_{\mathrm{IC}} + \mathbb{S}_{\mathrm{WF}} \approx 1$. We remark that the analysis at different times are processed independently, so proposed EOF-PC approach offers flexibility of analyzing sensitivities at moments of interest only. This flexibility in turn provides means of investigating the dynamic of the uncertainty propagation. For instance, the plots for the SSH (left column) in Fig. 17 show how the effect of wind forcing becomes progressively important near the LC region as time advances. Similarly, observing the time evolution of the variability of the MLD fields (see right column of Fig. 17), one notices the fast evolution of MLD variance induced by uncertainties in the initial condition in the western half of the GoM.

5.4 Effect of sample set size

This section explores the robustness of the proposed EOF-PC approach with respect to the size of HYCOM realization

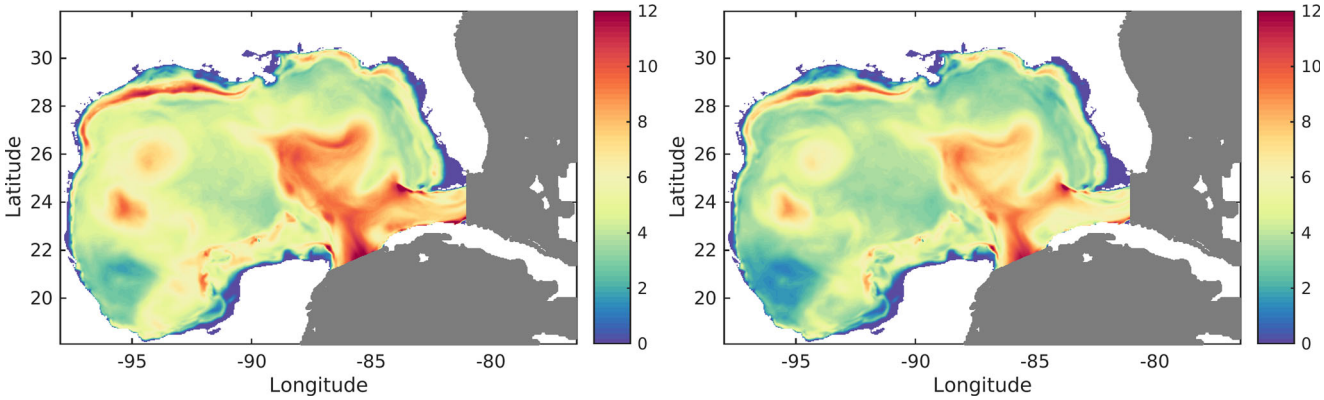

Fig. 14 Standard deviation of MLD fields at day 30. *Left*: empirical standard deviation calculated from the LHS set of HYCOM simulations. *Right*: standard deviation in the truncated expansion of the field using $r = 27$ modes

ensemble in global sensitivity analysis for the SSH and MLD fields. Our objective is to show that, for the present problem, a limited number of HYCOM realizations provides adequate estimates of the sensitivity indices. To this end, we repeat the previous procedure for the EOF-PC approximation of the SSH and MLD fields and subsequent sensitivity analysis, for different sizes of the realizations set. In order to avoid having to repeat multiple HYCOM simulations when the new sample sets are considered, a sub-sampling of the original LHS set considered so far is introduced. Figure 18 illustrates in two dimensions the sub-sampling strategy employed in this study, which is now briefly discussed. From the original set $\mathcal{P}_{\text{LHS}}$ with N_{LHS} samples, the objective is to select a subset $\mathcal{P}'_{\text{LHS}} \subset \mathcal{P}_{\text{LHS}}$ preserving the covering property of LHS schemes. To this end, we first draw at random a new set of N'_{LHS} sample points; for every element of this new sample set, we select its closest neighbor in $\mathcal{P}_{\text{LHS}}$ (with respect to the L_2 distance) that has not been previously drawn and complete the new sample set $\mathcal{P}'_{\text{LHS}}$ with this selected sample point. It is noted that the resulting $\mathcal{P}'_{\text{LHS}}$ may not necessarily be an LHS set.

The accuracy of the EOF-PC approximation of the SSH or MLD field depends on the error in the empirical covariance estimate, the number of modes r retained in the expansion, and finally the accuracy of PC approximation of the random coordinates $\phi^k(\xi)$. With the above quasi-uniform sub-sampling scheme, we choose different number of HYCOM realizations from the original LHS ensemble and reconstruct EOF-PC approximations U^{PC} for both SSH and MLD fields. The quality of the resulting approximations is first assessed using the (normalized) L_2-norm of the approximation error,

$$\epsilon_{L_2}^2 = \frac{\mathbb{E}\left[\|U - U^{PC}\|_{L_2(\Omega)}^2\right]}{\mathbb{E}\left[\|U\|_{L_2(\Omega)}^2\right]}. \tag{31}$$

This error can not be computed exactly and to avoid the need to perform additional HYCOM simulations, we here consider estimates based on the LHS sample set $\mathcal{P}_{\text{LHS}}$,

$$\epsilon_{L_2}^2 \approx \frac{\sum_{\xi_j \in \mathcal{P}_{\text{LHS}}} \|U(\cdot, \xi_j) - U^{PC}(\cdot, \xi_j)\|_{L_2(\Omega)}^2}{\sum_{\xi_j \in \mathcal{P}_{\text{LHS}}} \|U(\cdot, \xi_j)\|_{L_2(\Omega)}^2}. \tag{32}$$

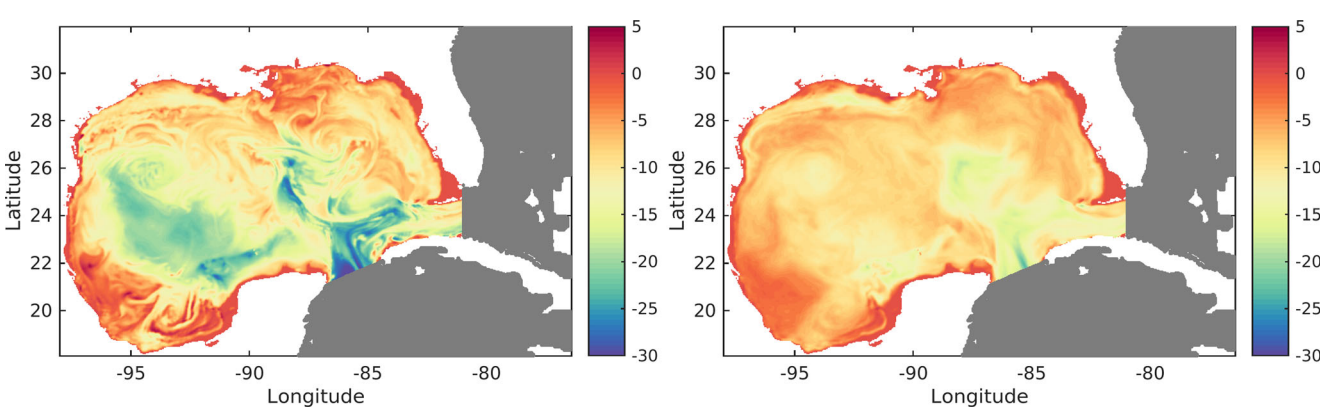

Fig. 15 MLD fluctuation fields for realization at $\xi = 0$ at day 30. *Left*: target fluctuation field. *Right*: EOF-PC reconstruction of the target fluctuation field

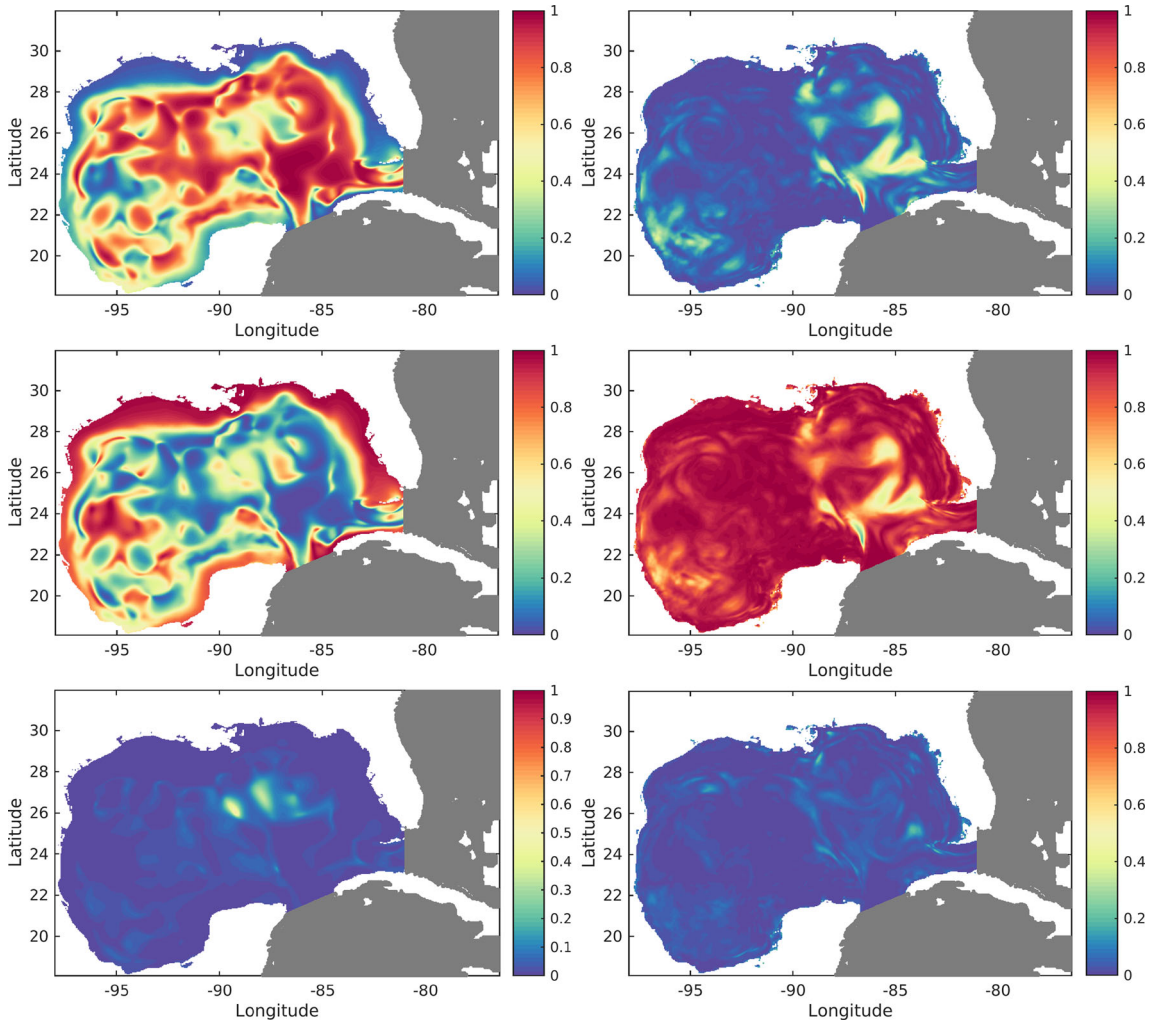

Fig. 16 Sensitivity analysis of the SSH (*left*) and MLD (*right*) fields at day 30. Plotted are the first order sensitivity indices related to the initial condition (*top row*), wind forcing (*center row*) and interaction between the two (*bottom row*)

Because this error estimate relies on the same sample set used for the construction, it is expected that it will underestimate the true L_2-error. However, this estimate remains useful to monitor the robustness of the approximation when different number of samples are used in the construction of the EOF-PC expansion. The evolutions of the estimated errors are shown in Fig. 19. The top plot concerns the SSH case at day 30, while the bottom plot corresponds to the MLD at the same date. In each plot, the blue curve corresponds to the estimate in Eq. 32 while the red curve corresponds to similar estimate obtained using the same subset as for the EOF-PC approximation. Finally, the dotted line corresponds to the a priori error estimate ($\epsilon^{priori} = 1 - \frac{\sum_{k=1}^{r} \lambda^k}{\sum_{i=k}^{N_{LHS}} \lambda^k}$) based on the truncation of the expansion retaining only r terms.

For the two fields, a similar behavior is reported. First, the lowest error estimation is the one based on the truncation criteria (dotted line), as one would have expected. This estimate is sightly smaller than the estimate based on the reduced sample set used for the construction of the EOF-PC approximation (red curve). The difference between these two estimate arises from the PC approximation of the stochastic coordinates ϕ^k, as discussed previously. Interestingly, the distance between the dotted and red curves remains essentially constant, suggesting that reducing the PC approximation error would require a significantly larger sample set. Finally, the two previous errors are seen to significantly under estimate the true error, as denoted by their distance to the blue curve which can be considered as a better approximation of the true L_2-error when the sample set size used for the construction is small compared to N_{LHS}. Obviously, the estimates for the original and reduced samples agree when the number of samples used for the construction goes to N_{LHS}. Overall, from the curves reported, it can be concluded that the EOF-PC error quickly decreases with the number of samples involved in the construction, but then stagnate as many more samples would be necessary to

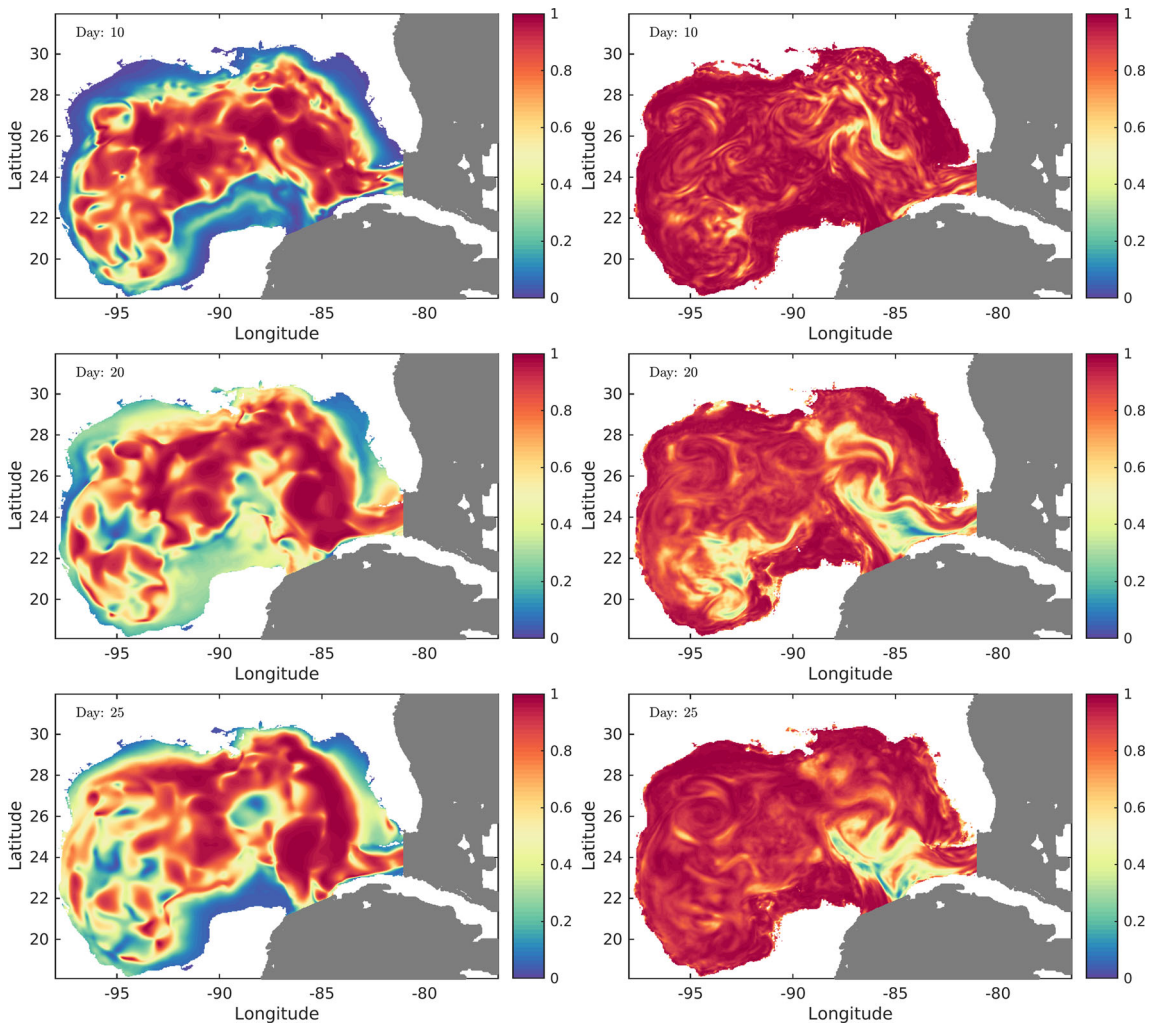

Fig. 17 1st order sensitivity of SSH to initial condition (*left column*) and MLD to wind forcing (*right column*) perturbations on selected days as indicated

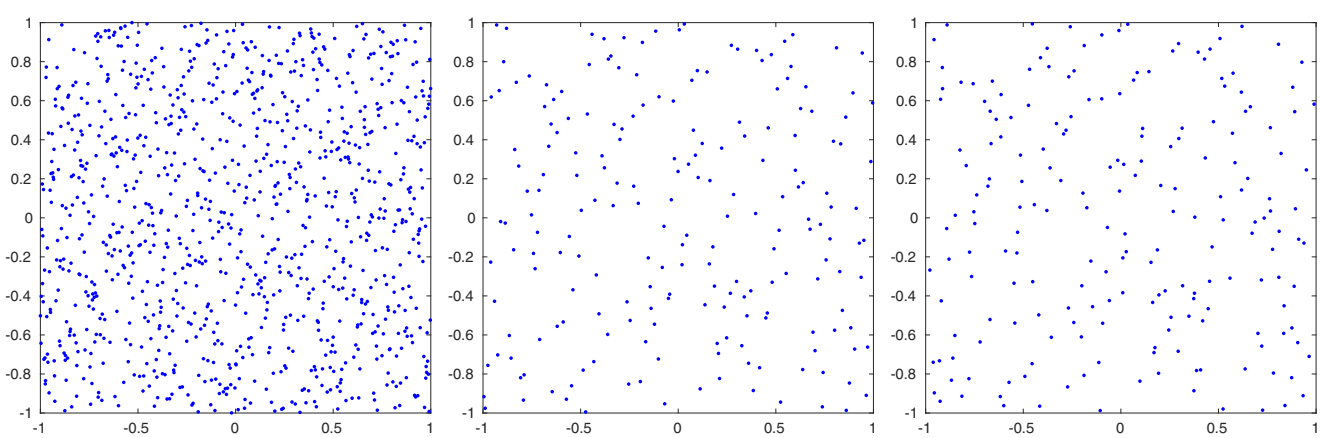

Fig. 18 Illustration of the sub-sampling strategy in 2D case. *Left*: original LHS set $\mathcal{P}_{\mathrm{LHS}}$. *Center*: independent LHS set $N'_{\mathrm{LHS}} < N_{\mathrm{LHS}}$ elements. *Right*: resulting sample set $\mathcal{P}' \subset \mathcal{P}_{\mathrm{LHS}}$

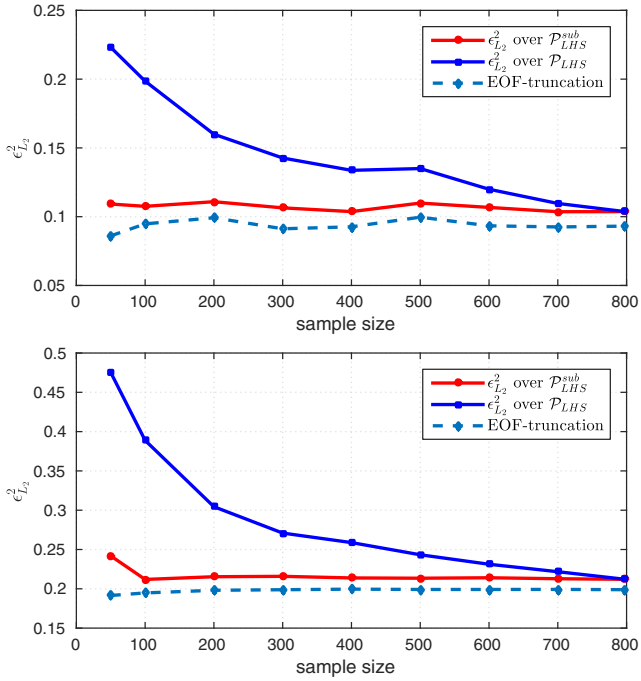

Fig. 19 Estimates of the mean squared error of the EOF-PC approximations for SSH (*top*) and MLD (*bottom*) fields as function the size of the subset of HYCOM realizations used in the construction

properly capture fine stochastic features. As discussed previously, this finding indicates that one should be cautious in using the EOF-PC approximation as a surrogate model to predict realizations of the HYCOM simulations at ξ points that do not belong to the sample, as significant (point-wise) errors can be expected.

However, despite the occurrence of appreciable point-wise errors and appreciable L_2 error estimates, we note that the EOF-PC is still able to capture how the variability is distributed as well as the impact of the different uncertain parameters. In fact, a closer inspection reveals that even for a sample set of only 50 HYCOM simulation the EOF-PC based sensitivity indices for both the SSH and MLD fields are well determined. This finding is illustrated in Fig. 20 which depicts in the left column (resp. right column) the standard deviation of the EOF-PC approximation for the SSH at day 30 (resp. MLD) and first order sensitivity indices associated to the initial condition (resp. wind forcing). It is seen that while using roughly 16 times less samples the approximations agree fairly well with the results reported previously (see corresponding plots in Fig. 10, 14, 16). It thus appears that a moderate number of HYCOM realizations would be sufficient to perform the global sensibility analysis of the SSH and MLD fields

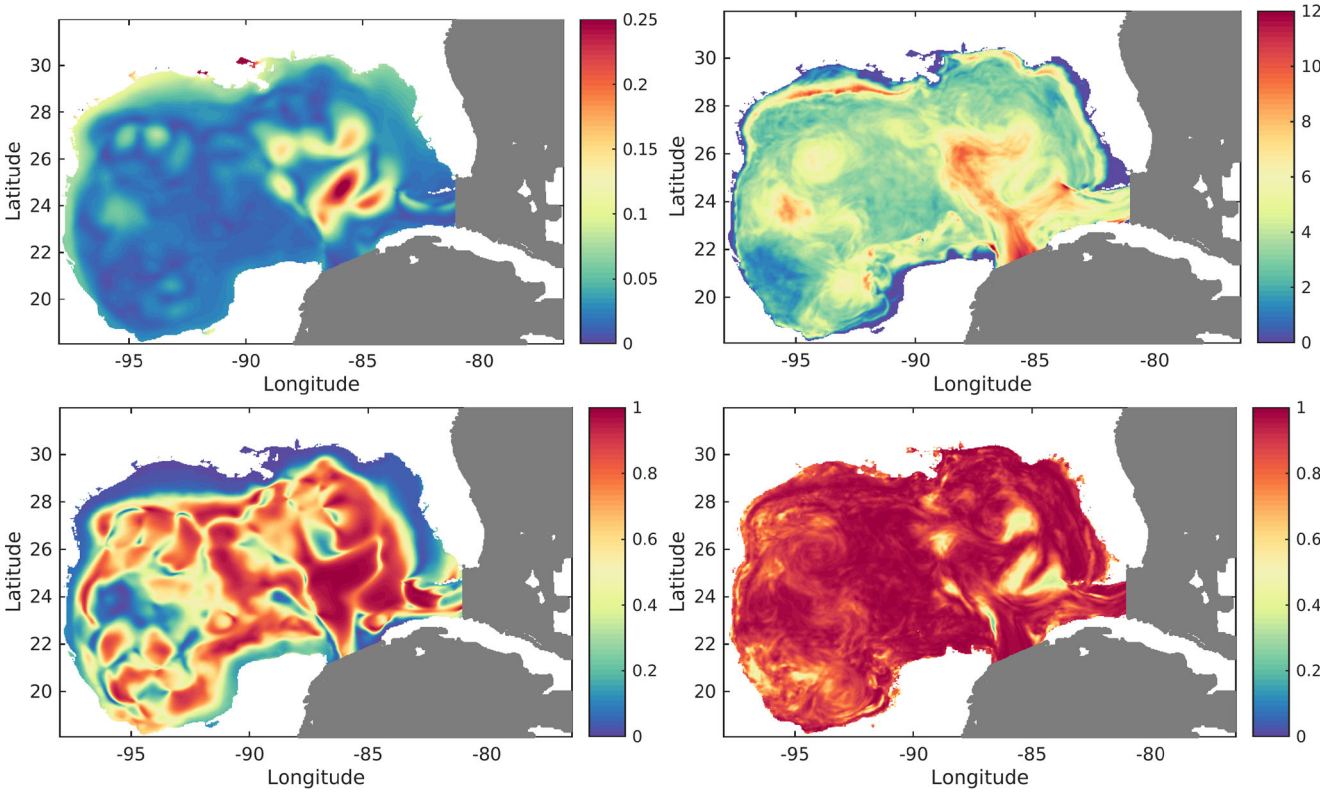

Fig. 20 Standard deviation fields (*top*) and first order sensitivity index (*bottom*) of the SSH (*left*) and MLD (*right*) fields at day 30. The first order sensitivity index is related to the initial condition for the the SSH (*left plot*) and to the wind forcing for the MLD case (*right plot*)

through EOF-PC reduction approach. This is an encouraging result for applying the EOF-PC methodology in the context of data assimilation, which requires quantifying model uncertainties. Because ocean models are computationally expensive and generate very large quantities of data, ocean data assimilation cannot be based on techniques requiring large ensembles. To date, ensemble data assimilation techniques implemented in oceanography typically use ensembles of size $\Omega(100)$ (e.g., [12, 20]), comparable to the ensemble size we find reasonable for implementing the EOC-PC approach.

6 Conclusion

A PC method was used to quantify the impact of initial condition and wind forcing uncertainties on the circulation in the GoM. Attention was focused on the sea surface height (SSH) and mixed layer depth (MLD). A LHS ensemble of oceanic forecasts covering a 30-day period were generated using HYCOM. The resulting database of realizations was then exploited to construct PC surrogates. A BPDN algorithm [37, 38] was used for this purpose.

For the purpose of propagating uncertainties, two quantities of interest were defined, obtained by averaging the SSH over a region located close to the zone of the LC detachment, and the MLD around the location of the DWH well. The adequacy of the PC representation was examined in light of discrete error metrics, and the analysis revealed that the reconstructions adequately capture the stochastic response of the QoIs, and provides robust estimates of statistical moments, including means and variances. A global sensitivity analysis was then performed in order to quantify the impact of the uncertain inputs. The analysis indicated that during the time of the simulation, the regionally averaged SSH response near the LC region is dominated by initial condition uncertainties, whereas the MLD around DWH is mostly sensitive to wind forcing.

A EOF-PC decomposition methodology was then developed in order to assess the field sensitivities of SSH and MLD. Due to the smooth response of SSH field in both spatial and stochastic domains, we were able to build faithful PC surrogates for SSH fields with as few as 10 EOF modes, and to capture the region of peak SSH uncertainty around the LC. On the other hand, the EOF-PC representation of MLD field was limited to include only 80 % of its variability due to the fact that capturing fine structures in MLD fields requires more HYCOM realizations. Nevertheless, despite the fact that fine spatial structures in the MLD field are smoothed out in the reconstruction, the dominant structures in variance and sensitivity fields were well captured.

A computational study was also conducted to assess the impact of the size of the realization ensemble on the performance of the EOF-PC representation. In order to avoid generating inpedendent ensemble, a simplified approach was adopted, based on coarsening the original LHS ensemble while maintaining its covering property. The analysis revealed the possibility of capturing the dominant features of the stochastic variability with a relatively modest ensemble size of 50 members, which offers perspective for applying the EOF-PC technique in an ocean data assimilation context.

We finally recall that the present analysis considers statistically independent uncertainties for the initial conditions and wind forcing fields, such that their respective EOF decompositions can be easily sampled simultaneously with a classical LHS ensemble construction. Though well motivated from the mathematical perspective, this simplified approach disregard the link between the ocean and atmosphere states. Similarly, passive OGCM simulations where wind forcing is not affected by the ocean states should be improved to a coupled atmosphere-wave-ocean model to better simulate the dynamics across the atmosphere-ocean interface. Based on the insight provided by the present experiences, work is underway to enhance both the construction and propagation of more realistic uncertain input fields, particularly by relying on the fully coupled atmosphere-wave-ocean model [6]. Results from this ongoing effort will be reported in a future study.

Acknowledgments This work was supported in part by a grant from BP/The Gulf of Mexico Research Initiative to CARTHE, and by the US Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research, under Award Number DE- SC0008789. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

M. Le Hénaff received partial support for this work from the NOAA Quantitative Observing System Assessment Program (QOSAP, grant NA15OAR4320064) and the base funds of the NOAA Atlantic Oceanographic and Meteorological Laboratory.

Appendix

EOF Calculation

The state vector in the EOF calculation consisted of the hydrostatic pressure increments in each layer and the sea surface pressure (which can be considered as an additional layer for the sake of accounting). The dimension of this state vector is then the product of the number of "wet" HYCOM points (150,684) times the number of layers plus one (21), i.e., 3,164,364; the data matrix, X, had 14 columns. The data were first standardized by removing each variables

mean and dividing by its standard deviation, so that all could be represented on a common scale (The components of the EOFs were multiplied by the corresponding standard deviations to restore their units prior to forming the perturbation).

The EOFs can be computed by a singular value decomposition: $X = USV^\top$ where U and V are unitary matrices whose columns are the left and right singular vectors, respectively, and S is a diagonal matrix containing the singular values. The EOFs are the left singular vectors and the principle components are the right singular vectors. It is more efficient for highly rectangular matrices to compute the principal components by explicitly forming the (14×14 in the initial conditions case) covariance matrix, $C = X^\top X$, and performing an eigenvalue decomposition using the matlab function $[V, \Lambda] = \mathbf{eig}(C)$. The principle components are then the eigenvectors of C and the eigenvalues are the squares of the singular values. The EOFs can then be computed as: $U = XVS^{-1}$. The principle components obtained from this procedure were used to identify the modes of the prognostic variables in the model: $\mathcal{U} = \mathcal{X}VS^{-1}$ where $\mathcal{X}$ is the state vector referring to temperature, salinity or velocity and $\mathcal{U}$ are the corresponding EOFs. This device to compute the EOFs of the model variable was needed because of the hybrid nature of HYCOM's vertical coordinate system which switches freely between isopycnal and isobaric coordinates, by the need to maintain non-negative layer thicknesses and by the need to maintain hydrostatic stability.

References

1. Alexanderian, A., Winokur, J., Sraj, I., Srinivasan, A., Iskandarani, M., Thacker, W.C., Knio, O.M.: Global sensitivity analysis in an ocean general circulation model: a sparse spectral projection approach. Comput. Geosci. **16**(3), 757–778 (2012)
2. Alvera-Azcárate, A., Barth, A., Rixen, M., Beckers, J.M.: Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the adriatic sea surface temperature. Ocean Model. **9**(4), 325–346 (2005)
3. Blatman, G., Sudret, B.: Adaptive sparse polynomial chaos expansion based on least angle regression. J. Comput. Phys. **230**(6), 2345–2367 (2011)
4. Bleck, R.: An oceanic general circulation model framed in hybrid isopycnic cartesian coordinates. Ocean Modell. **4**(1), 55–88 (2002). doi:10.1016/S1463-5003(01)00012-9
5. Chassignet, E.P., Hurlburt, H.E., Smedstad, O.M., Halliwell, G.R., Hogan, P.J., Wallcraft, A.J., Baraille, R., Bleck, R.: The hycom (hybrid coordinate ocean model) data assimilative system. J. Mar. Syst. **65**(1-4), 60–83 (2007). doi:10.1016/j.jmarsys.2005.09.016
6. Chen, S.S., Curcic, M.: Ocean surface waves in hurricane ike (2008) and superstorm sandy (2012): Coupled model predictions and observations Ocean Modelling (2015)
7. Conrad, P.R., Marzouk, Y.M.: Adaptive smolyak pseudospectral approximations. SIAM J. Sci. Comput. **35**(6), A2643–A2670 (2013)
8. Constantine, P.G., Eldred, M.S., Phipps, E.T.: Sparse pseudospectral approximation method. Comput. Methods Appl. Mech. Eng. **229**, 1–12 (2012)
9. Crestaux, T., Le Maître, O., Martinez, J.M.: Polynomial chaos expansion for sensitivity analysis. Reliab. Eng. Syst. Saf. **94**(7), 1161–1172 (2009)
10. Doostan, A., Owhadi, H.: A non-adapted sparse approximation of pdes with stochastic inputs. J. Comput. Phys. **230**(8), 3015–3034 (2011)
11. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
12. Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res **99**(C5), 10,143–10,162 (1994)
13. Gerstner, T., Griebel, M.: Dimension-adaptive tensor—product quadrature. Computing **71**(1), 65–87 (2003)
14. Ghanem, R.G., Spanos, P.D.: Stochastic finite elements: a spectral approach. Springer-Verlag, New York (1991)
15. Hodur, R.M.: The naval research laboratorys coupled ocean/atmosphere mesoscale prediction system (coamps). Mon. Weather Rev. **125**, 1414–1430 (1997). doi:10.1175/1520-0493(1997)125¡1414:TNRLSC¿2.0.CO;2
16. Homma, T., Saltelli, A.: Importance measures in global sensitivity analysis of nonlinear models. Reliab. Eng. Syst. Saf. **52**(1), 1–17 (1996)
17. Ichiye, T.: Circulation and water mass distribution in the gulf of Mexico. Geofis. Int. **2**(3), 47–76 (1962)
18. Kac, M., Siegert, A.: An explicit representation of a stationary gaussian process. Annals of Mathematical Statistics, 438–442 (1947)
19. Karhunen, K.: Über lineare Methoden in der Wahrscheinlichkeitsrechnung, vol. 37. Universitat Helsinki (1947)
20. Keppenne, C., Rienecker, M.: Initial testing of a massively parallel ensemble Kalman filter with the Poseidon isopycnal ocean general circulation model. Mon. Weather. Rev. **130**(12), 2951–2965 (2002)
21. Knio, O.M., Najm, H.N., Ghanem, R.G., et al.: A stochastic projection method for fluid flow: i. basic formulation. J. Comput. Phys. **173**(2), 481–511 (2001)
22. Le Hénaff, M., Kourafalou, V.H., Paris, C., Helgers, J., Aman, Z.M., Hogan, P.J., Srinivasan, A.: Surface evolution of the deepwater horizon oil spill patch: combined effects of circulation and wind-induced drift. Environ. Sci. Tech. **46**(13), 7267–7273 (2012)
23. Le Maître, O.P., Knio, O.M.: Spectral methods for uncertainty quantification: with applications to computational fluid dynamics. Springer Science & Business Media (2010)
24. Le Maître, O.P., Najm, H.N., Pébay, P.P., Ghanem, R.G., Knio, O.M.: Multi-resolution-analysis scheme for uncertainty quantification in chemical systems. SIAM J. Sci. Comput. **29**(2), 864–889 (2007)
25. Le Maître, O.P., Reagan, M.T., Najm, H.N., Ghanem, R.G., Knio, O.M.: A stochastic projection method for fluid flow: Ii. random process. J. Comput. Phys. **181**(1), 9–44 (2002)
26. Loève, P.: Fonctions Aléatoires Du Second Ordre, a Note in P. Lévy, Processus Stochastiques Et Mouvement Brownien. Gauthier-Villars, Paris (1948)
27. Marzouk, Y.M., Najm, H.N.: Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems. J. Comput. Phys. **228**(6), 1862–1902 (2009)
28. McKay, M.D., Beckman, R.J., Conover, W.J.: Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**(2), 239–245 (1979)

29. Najm, H.N., Debusschere, B.J., Marzouk, Y.M., Widmer, S., Le Maître, O.: Uncertainty quantification in chemical systems. Int. J. Numer. Methods Eng. **80**(6), 789 (2009)
30. Peng, J., Hampton, J., Doostan, A.: A weighted l1-minimization approach for sparse polynomial chaos expansions. J. Comput. Phys. **267**, 92–111 (2014)
31. Schmitz, W.J.: Cyclones and Westward Propagation in the Shedding of Anticyclonic Rings from the Loop Current. In: Sturges, W., Lugo-Fernandez, A. (eds.) Circulation in the Gulf of Mexico: Observations and Models, Geophysical Monograph Series, vol. 161, pp. 241–261. American Geophysical Union, Washington, D.C (2005)
32. Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. J. Royal Stat. Soc. Ser. B (Methodological) **53**(3), 683–690 (1991). http://www.jstor.org/stable/2345597
33. Sobol, I.: Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exp. **1**, 407–414 (1993)
34. Sraj, I., Iskandarani, M., Srinivasan, A., Thacker, W.C., Winokur, J., Alexanderian, A., Lee, C.Y., Chen, S.S., Knio, O.M.: Bayesian inference of drag parameters using axbt data from typhoon fanapi. Mon. Weather. Rev. **141**(7), 2347–2367 (2013)
35. Sraj, I., Maître, O.P.L., Knio, O.M., Hoteit, I.: Coordinates transformation and polynomial chaos for the bayesian inference of a gaussian process with parametrized prior covariance function. arXiv preprint arXiv:1501.03323 (2015)
36. Thacker, W.C., Srinivasan, A., Iskandarani, M., Knio, O.M., Le Henaff, M.: Propagating boundary uncertainties using polynomial expansions. Ocean Model. **43**, 52–63 (2012)
37. Van Den Berg, E., Friedlander, M.: Spgl1: a solver for large-scale sparse reconstruction (2007)
38. Van Den Berg, E., Friedlander, M.P.: Probing the Pareto frontier for basis pursuit solutions. SIAM J. Sci. Comput. **31**(2), 890–912 (2008)
39. Winokur, J., Conrad, P., Sraj, I., Knio, O., Srinivasan, A., Thacker, W.C., Marzouk, Y., Iskandarani, M.: A priori testing of sparse adaptive polynomial chaos expansions using an ocean general circulation model database. Comput. Geosci. **17**(6), 899–911 (2013)
40. Winokur, J.G.: Adaptive sparse grid approaches to polynomial chaos expansions for uncertainty quantification. Ph.D. thesis Duke University (2015)
41. Xiu, D., Karniadakis, G.E.: The wiener–askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002)