2017-08-02

The Application of Uncertainty Quantification Techniques and Information Theory to Oil Spill and Ocean Forecasting

Shitao Wang

University of Miami, swang@rsmas.miami.edu

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation


https://scholarlyrepository.miami.edu/oa_dissertations/1931
UNIVERSITY OF MIAMI

THE APPLICATION OF UNCERTAINTY QUANTIFICATION TECHNIQUES AND INFORMATION THEORY TO OIL SPILL AND OCEAN FORECASTING

By
Shitao Wang

A DISSERTATION

Submitted to the Faculty of the University of Miami in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Coral Gables, Florida
August 2017
THE APPLICATION OF UNCERTAINTY QUANTIFICATION TECHNIQUES AND INFORMATION THEORY TO OIL SPILL AND OCEAN FORECASTING

Shitao Wang

Approved:

Mohamed Iskandarani, Ph.D.
Professor of Ocean Sciences

Tamay Özgökmen, Ph.D.
Professor of Ocean Sciences

Arthur Mariano, Ph.D.
Professor of Ocean Sciences

Matthieu Le Hénaff, Ph.D.
Assistant Scientist
Cooperative Institute for Marine and Atmospheric Studies

Omar M. Knio, Ph.D.
Professor of Mechanical Engineering and Materials Science
Duke University

Guillermo Prado, Ph.D.
Dean of the Graduate School
WANG, SHITAO (Ph.D., Meteorology and Physical Oceanography)
The Application of Uncertainty Quantification Techniques and Information Theory to Oil Spill and Ocean Forecasting (August 2017)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Mohamed Iskandarani.
No. of pages in text. (99)

Quantifying uncertainties in ocean current forecasts is an important component of formulating a response to an oil spill, e.g. to compute the anticipated oil trajectories. Polynomial Chaos (PC) methods have recently been used to quantify uncertainties in the circulation forecast of the Gulf of Mexico caused by uncertain initial conditions and wind forcing data. The input uncertainties consisted of the amplitudes of perturbation modes whose space-time structure was obtained from Empirical Orthogonal Functions (EOF) decompositions. These efforts were the first to rely on a PC approach to efficiently quantify uncertainties in an ocean model, and as such have raised a number of issues that we wish to address, namely the realism of the perturbations, the effective choices in choosing the uncertain variables, the information trade-offs of the different uncertain input choices, and the ability to reduce these uncertainties if observational data is available.

We explore whether these EOF-based perturbations lead to realistic representation of the uncertainties in the circulation forecast of the Gulf of Mexico. We also use information theoretic metrics to quantify the information gain and the computational trade-offs between different wind forcing and initial condition EOF modes. Surface and subsurface model data comparisons show that the observational data falls within the envelope of the ensemble simulations and that the EOF decompositions deliver “realistic” perturbations in the Loop Current region. The result of the computational
trade-offs indicate that two initial condition EOF modes are enough to represent the uncertainties in the Loop Current region; while wind forcing EOF modes are necessary in order to capture uncertainties in the coastal zone. This result is consistent with the global sensitivity analysis.

The ensemble statistics are then explored using the PC approach and the newly developed contour boxplot method. Specifically, the contour boxplot is used to identify the most representative ensemble member and the outliers. The full probability density functions of sea surface height are estimated using the PC method. With 20 years of satellite observations, the predictability in the circulation forecast of the Gulf of Mexico is investigated using information theory.

Finally, we update our knowledge about the uncertain inputs using along track satellite observations. The best initial perturbations are found using the Bayesian optimization approach and the full posterior distributions of the uncertain inputs are estimated using the Bayesian inference framework.
To my family and friends.
Acknowledgments

First, I would like to acknowledge my advisor, Dr. Mohamed Iskandarani. He encouraged me to pursue the field of uncertainty quantification at the beginning of my graduate study. During the five years as his student, he is not only an advisor but also a close friend. He is always there whenever I need any help. There are many obstacles in my Ph.D. journey. It is him who connects me to the big picture when the details drive me nuts; it is him who connects me to the why when the what and how are insurmountable. It is his constant guidance, support and encouragement that made this work possible.

I would also like to thank all my committee members, Arthur Mariano, Omar M. Knio, Tamay Özgökmen, Matthieu Le Hénaff for all their support and help during the countless committee meetings and seminars. I want to thank the UQ group at RSMAS, Carlisle Thacker, Rafael Gonçalves, Ashwanth Srinivasan and Matthieu Hénaff. I learned a lot from them through individual discussions and weekly group meetings. Also to the whole class of fall 2012, they are my fellow classmates as well as my family, making the time of my graduate study more enjoyable. Lastly, I would like to thank my parents and my wife for their constant love and support.
This work was supported by a grant from BP/The Gulf of Mexico Research Initiative to the Consortium for Advanced Research on Transport of Hydrocarbon in the Environment (CARTHE).
Contents

List of Figures ix

List of Tables xv

1 Introduction 1

1.1 Background ............................................. 3
  1.1.1 PC-based surrogate ........................... 4
  1.1.2 Characterizing uncertainty using relative entropy ........ 6
  1.1.3 Updating input uncertainties using observational data .. 8
  1.1.4 HYCOM Experiments ........................... 10

2 Methodology 12

  2.1 Polynomial Chaos .................................. 13
    2.1.1 The projection approach ...................... 16
    2.1.2 The regression approach ...................... 17
    2.1.3 Sensitivity analysis .......................... 20

  2.2 Gaussian Process ................................. 21
    2.2.1 Hyperparameter selection ..................... 25

  2.3 Updating uncertain inputs using satellite observations .... 25
    2.3.1 Bayesian optimization ......................... 26
    2.3.2 Bayesian inference ............................ 29

  2.4 Remarks ............................................. 30
# Model and ensemble prediction

3.1 HYCOM setup ................................................................. 33
3.2 PC ensemble ............................................................... 34
3.3 Model-data comparison .................................................. 39
   3.3.1 Comparison against satellite SSH ................................. 39
   3.3.2 Comparison against mooring data ............................... 41

# Sensitivity analysis and information trade-offs

4.1 Sensitivity analysis ...................................................... 46
4.2 Quantifying information trade-offs using relative entropy ........ 47

# Statistical analysis and Bayesian inference

5.1 Ensemble statistics ...................................................... 56
   5.1.1 Contour boxplot .................................................. 56
   5.1.2 An exploration of SSH PDF .................................... 60
   5.1.3 Ensemble covariance .......................................... 62
   5.1.4 Predictability .................................................. 65
5.2 Updating the initial guess with observations ....................... 71
   5.2.1 AVISO along track observations ................................ 71
   5.2.2 Bayesian optimization approach ............................... 73
   5.2.3 Bayesian inference approach ................................. 77

# Summary and conclusions

6 Summary and conclusions ................................................. 82

A Cross-Validation .......................................................... 86

B Adaptive Metropolis Algorithm .......................................... 87

C Sensitivity Analysis ...................................................... 89
C.1 Sobol-Hoeffding decomposition: ........................................ 89
C.2 Sensitivity indices ............................................................ 90
C.3 Polynomial Chaos estimation of sensitivity indices ................. 92

References ................................. 94
List of Figures

2.1 First 6 Legendre polynomials. ..................................................... 14
2.2 Sampled realizations from Gaussian Process using different covariance functions. ................................................................. 23
2.3 An example of Bayesian optimization on an 1D toy problem. (picture from (Brochu et al., 2010)) The objective function is approximated by a Gaussian Process model where the posterior mean and variance of the Gaussian Process are shown in the figure. The solid line is the Gaussian Process mean, the shaded area is the Gaussian Process variance and the dash line represents the true underlying objective function. Black dots are the current observations and the red dot is the new observation. The bottom green curve is the acquisition function which determines where to evaluate next. ................................................................. 28
3.1 SSH patterns for initial condition EOF mode1-mode4. ............ 37
3.2 Wind speed patterns for wind forcing EOF mode1-mode4 and wind stress vectors are superimposed on the wind speed field. ........... 38
3.3 The cumulative variance explained by the first 10 EOF modes, the red labels specify the cumulative variance explained by the first 4 EOF modes. ................................................................. 39
3.4 SSH 17 cm contours from HYCOM ensemble (black) and AVISO SSH (white) at indicated forecast time. The background is AVISO SSH data. Top: 49-member HYCOM ensemble; Bottom: 798-member HYCOM ensemble.

3.5 Mean velocity and standard deviation ellipses at different depths built from 9 Bureau of Ocean Energy Management (BOEM)/Science Applications International Corporation (SAIC) full-depth mooring data (blue) and HYCOM ensemble (black and red). The period is from 05/01/2010 to 05/30/2010. Left: 49-member HYCOM ensemble; Right: 798-member HYCOM ensemble.

4.1 Time evolution of SSH mean (left) and standard deviation (right) estimated using PC surrogate.

4.2 Time evolution of the first order sensitivity indices for the initial condition uncertainties (left) and the wind forcing uncertainties (right).

4.3 Time evolution of the interaction terms in the sensitivity analysis for the initial condition uncertainties (left) and the wind forcing uncertainties (right). The interaction terms are defined by the difference between the total order sensitivity indices and the first order sensitivity indices.

4.4 SSH variability loss by omitting the variability of the input uncertainties (scenario 1 in table 4.1) measured by the relative entropy between the ensemble that perturbs all 8 EOF modes and ensemble that perturbs 2 EOF modes of each uncertain input sources.
4.5 SSH variability loss by omitting the wind forcing input uncertainties (scenario 2 in table 4.1) measured by the relative entropy between the fully perturbed ensemble and the ensemble that perturbs only its first 2 initial condition EOF modes. . . . . . . . . . . . . . . . . . . . . . 53

4.6 SSH variability loss by omitting the initial condition input uncertainties (scenario 3 in table 4.1) measured by the relative entropy between the fully perturbed ensemble and the ensemble that perturbs only its first 2 wind forcing EOF modes. . . . . . . . . . . . . . . . . . . . . . . . . . 53

4.7 SSH variability loss by retaining only the 1st initial condition EOF mode (scenario 4 in table 4.1) measured by the relative entropy between the fully perturbed ensemble and the ensemble that perturbs only its 1st initial condition EOF mode. . . . . . . . . . . . . . . . . . . . . . 54

4.8 SSH variability loss by retaining only the 2nd initial condition EOF mode (scenario 5 in table 4.1) measured by the relative entropy between the fully perturbed ensemble and the ensemble that perturbs only its 2nd initial condition EOF mode. . . . . . . . . . . . . . . . . . . . . . 54

5.1 Contour boxplot of the Loop Current edge derived from the 49-member HYCOM ensemble. All HYCOM ensemble contours are color coded according to the legend and their associated uncertain input perturbations are shown in the bottom left panel. The satellite SSH Loop Current edge is shown in black for reference. . . . . . . . . . . . . . . . 58

5.2 SSH initial condition perturbation patterns. The SSH difference between the perturbed and the unperturbed simulations one day after the start of the simulation. The 17cm SSH contour of the unperturbed simulation is added for reference. . . . . . . . . . . . . . . . . . . . . . . 60
5.3 The SSH PDFs of selected points at day 30. The yellow contours on the map shows the Loop Current edges in the ensemble simulations. The red contour is the Loop Current edge for the unperturbed simulation.

5.4 The relative entropy of SSH PDFs to their Gaussian counterparts at day 30 (the color shows only the region where the relative entropy is greater than 0.4). The black contours are HYCOM Loop Current edges for reference.

5.5 Covariance between SSH at one point (indicated by the black asterisk) and the SSH in the entire Gulf of Mexico. Empirical estimate is shown on the left, PC estimate is shown in the middle and the difference between the empirical estimate and the PC estimate is shown on the right.

5.6 Covariance between SSH at the same point shown in Figure 5.5 and the temperature along a vertical section (25.98°N). Empirical estimate is shown on the left, PC estimate is shown in the middle and the difference between the empirical estimate and the PC estimate is shown on the right.

5.7 The SSH predictability measured by the ratio between the standard deviations between the HYCOM ensemble simulations and the climatological data. The red contours identify the region where the ratio is greater than 1.

5.8 Comparison of area-averaged (2 degree boxes) SSH PDFs at day 10. The AVISO climatological SSH PDFs is in red and the HYCOM SSH PDFs is in blue.
5.9 Comparison of area-averaged (2 degree boxes) SSH PDFs at day 20.
The AVISO climatological SSH PDFs is in red and the HYCOM SSH
PDFs is in blue. ................................................................. 70
5.10 Comparison of area-averaged (2 degree boxes) SSH PDFs at day 30.
The AVISO climatological SSH PDFs is in red and the HYCOM SSH
PDFs is in blue. ................................................................. 70
5.11 AVISO along track data used in the experiment starting from May 1
to May 30, 2010. Three satellite missions are used here: Envisat(red),
Jason-2(blue) and Jason-1 on its new orbit (black) . . . . . . . . . . 72
5.12 Preprocessed along track SSH for different time segments. . . . . . 72
5.13 Modeled objective function with Gaussian Process at the end of the
optimization cycle. All previous evaluation points are marked with
blue dots. The next proposed evaluation point is shown in black. The
surface is the mean function of the Gaussian Process model. . . . . . 75
5.14 The acquisition function at the end of the optimization cycle. The
black dot represents the next proposed evaluation point. . . . . . . 76
5.15 The minimum observed objective function (RMS error) values versus
the number of function evaluations. 10 random restarts are performed
in the experiment. Colored curves represent each individual optimization
cycle while the thick black curve is the mean of all optimization
cycles. ................................................................. 76
5.16 MCMC chain samples for $\xi_1$ (left), $\xi_2$ (middle) and $\sigma^2$ (right). . . . 79
5.17 Joint posterior distribution of $\xi_1$ and $\xi_2$. ......................... 80
5.18 Root mean square error between model simulations and along track
satellite observations with different initial condition perturbation $\xi_1$
and $\xi_2$. ................................................................. 80
5.19 SSH snapshots at day 10 (top), day 20 (middle) and day 30 (bottom).

The AVISO observations are located on the right, and the model simulations are located on the left. The left three columns represent simulations with most negative perturbation on $\xi_1$ (left), zero perturbation on $\xi_1$ (middle) and most positive perturbation on $\xi_1$ (right). . . . . . . 81

B.1 2D example of Metropolis algorithm. . . . . . . . . . . . . . . . . 87

C.1 Illustration of different sensitivity indices. . . . . . . . . . . . . . . 91
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Some common input PDFs and their associated orthogonal polynomial basis. Figure 2.1 shows the first 6 Legendre polynomials.</td>
<td>14</td>
</tr>
<tr>
<td>4.1</td>
<td>Left: Different uncertain input perturbation scenarios. Right: The uncertain input distributions in the relative entropy function form, where the shown variables will be perturbed uniformly from -1 to 1 and 0 means no perturbation. The perturbation amplitude is described by the standardized random variable $\xi = \frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}}$ where $\theta$ represents the actual uncertain input, the standard random variable, $\xi$, ranges from -1 to 1.</td>
<td>51</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Numerical simulations have become an essential tool for engineers and scientists thanks to the continuous improvement in numerical algorithms and rapid growth of computer powers. Numerical simulations allow engineers to reduce the number of costly real experiments, provide scientists the necessary tools for studying the complex systems and enable the forecasting of oceanic and atmospheric circulations. However, numerical simulations need to be carefully performed in order to ensure good fidelity between the simulation results and the system they are meant to represent. In many cases, numerical simulations inherently involve uncertainties. The sources of these uncertainties can be grouped into (1) model errors (e.g. oversimplification of the mathematical equations describing the system being studied); (2) numerical errors caused by discretization of the continuous model equations; and (3) data errors, such as uncertainties in specifying the initial condition, boundary condition, forcing, etc. Only the data errors will be considered in this work. Uncertainty Quantifications (UQ) aims to estimate the impact of model uncertain inputs on the model output. UQ enhances the model’s usefulness by presenting its output in a probabilistic framework, in which the probability density function (PDF) of the model output is estimated given the PDFs of the uncertain inputs. This forward
propagation of uncertainty can be used to gauge the confidence in the model prediction. Decision makers can then consider a range of possible scenarios that reflect the uncertainties of the model inputs. Another UQ objective is to determine the principle contributor to the uncertainties of the model output. This sensitivity analysis can be used to guide the response or experimental design team to focus on the most influential uncertain input given the limited resources. When observational data is available, inference or optimization techniques can be leveraged to blend the model output and observational data to refine the model input uncertainties and thereby reduce the uncertainties of the model output.

Our interest in UQ is motivated by an environmental catastrophe, the 2010 Deepwater Horizon (DWH) oil spill, the worst marine oil spill in US history (McNutt et al., 2012). One of the critical question during the DWH concerned the path of the spilled oil. A major concern at the beginning of the DWH was whether the oil would be entrained into the Loop Current and affect the ecosystems in the Florida keys. Oil fate model could potentially answer this question and help the response team to allocate their limited resources. An oil fate model usually mimics the evolution of the hydrocarbon (oil and natural gases) released from the wellhead through a subsurface buoyant plume to the sea surface advecting by the ocean currents (Özgökmen et al., 2016). More sophisticated oil fate model considers other important factors, such as emulsification, biodegradation, weathering, etc. The reliability of the oil fate model depends critically on the quality of its input data, such as the initial release flow rate, droplet size distribution, and ocean currents. Quantifying uncertainties in these upstream input data, such as oceanic and atmospheric model predictions, will enhance the forecasting abilities of the oil fate model.
The objective of the thesis is to quantify the uncertainties in the ocean general circulation model, which focuses on uncertainties connected to the long term evolution of the oil spill. The specific scientific objectives are as follows:

- To compare the model ensemble simulations against the observational data.
- To investigate the information trade-offs between the number of model simulations and the amount of uncertainties captured.
- To explore the statistics of the model ensemble simulations.
- To determine whether the ensemble oceanic prediction system still has predictability given climatological data.
- To refine the initial uncertain inputs given observational data.

1.1 Background

The Gulf of Mexico, where the DHW oil spill took place, is a suitable test bed for uncertainties studies. It is a well-observed regional sea that presents many dynamical features typical of the deep ocean such as currents and eddying jets. The Loop Current (LC) is a particularly dominant feature of the circulation in the Gulf of Mexico as it flows from the Yucatan Channel between Mexico and Cuba, to the Straits of Florida between Cuba and the Southeastern US. The LC presents a time varying extension, from a retracted path at the south of the basin, to an extended one reaching the edge of the continental shelf in the northeastern Gulf of Mexico. When it is extended, the LC sheds a large, anticyclonic eddy, called LC eddy, which then drifts westward, and the current retracts to the south. This shedding sequence often undergoes temporary detachments of the LC eddy from the current, before final separation. Small, cyclonic eddies at the edge of the LC play an active role in necking down and chopping the
extended LC, leading to the LC eddy detachments or separation. The DWH oil spill took place during such a LC eddy shedding sequence, and the fate of the spilled oil was partly influenced by the LC evolution and its frontal dynamics (Le Hénaff et al., 2012; Walker et al., 2011).

1.1.1 PC-based surrogate

The conventional approach to UQ relies on Monte Carlo methods, which estimate the PDF of the model output by randomly perturbing the model uncertain input parameters and repeating the model forecast many times. The ensemble is then analyzed for statistical information such as mean and variance of model output. It usually requires a large number of model runs and becomes inefficient when the model is computationally expensive. More advanced UQ techniques rely on a different sampling strategy: build a surrogate using a carefully chosen small ensemble and leverage the surrogate to generate a much larger ensemble for reliable statistics. One of these advanced UQ techniques is Polynomial Chaos (PC) expansions (Ghanem and Spanos, 1991; Xiu and Karniadakis, 2002; Le Maître and Knio, 2010; Iskandarani et al., 2016b), where the surrogate takes the form of a spectral series:

\[ u(x, t, \xi) \approx u_P(x, t, \xi) = \sum_{n=0}^{P} \hat{u}_n(x, t)\psi_n(\xi) \]  \hspace{1cm} (1.1)

where \( u(x, t, \xi) \) is the model output that depends on space \( x \), time \( t \), and uncertain inputs \( \xi \); \( u_P \) denotes the PC series representation of the model output, \( \hat{u}_n(x, t) \) are the PC coefficients and \( \psi_n(\xi) \) are suitably chosen basis functions. The PC basis function is selected according to the distribution of the uncertain inputs, e.g. Gaussian distribution associated with Hermite polynomial, uniform distribution associated with Legendre polynomial (Xiu and Karniadakis, 2002). Different variants of PC method
can be derived depending on the basis functions and the approach to determine the series coefficients, see Iskandarani et al. (2016a) for an overview of different methods with applications to oil spill and ocean forecasting. All these PC methods rely on samples from the original model, which are then used to determine the series coefficients. The more basis functions in the series, the more accurate the surrogate is to the original model; however more model samples are needed to compute the extra PC coefficients. The PC method has several benefits: (1) The PC method can be implemented non-intrusively in an ensemble fashion such that the model can be treated as a black-box without any code modification. (2) The constructed PC surrogate can be used as a cheap forward non-linear propagator, which can be leveraged to generate a much larger sample set for reliable statistics. (3) The PC method provides a straightforward sensitivity analysis, which does not require the adjoint code. (4) The PC method makes the full Bayesian analysis possible by constructing an efficient surrogate. The PC method is certainly of interest to oceanographers and meteorologists, as the oceanic and atmospheric simulations are computationally expensive and have many uncertainty sources. Recently, the PC method has been successfully applied to the oceanic and oil spill simulations, including propagation of boundary condition uncertainties (Thacker et al., 2012; Li et al., 2016), propagation of initial condition uncertainties (Li et al., 2016; Iskandarani et al., 2016a), quantification of model parametric uncertainties (Winokur et al., 2013; Wang et al., 2015; Gonçalves et al., 2016), sensitivity analysis (Alexanderian et al., 2012), Bayesian inverse problem (Sraj et al., 2013, 2014, 2016). It should be noted that PC is not the only approach for UQ. Gaussian Process is also one of the available UQ techniques, see Iskandarani et al. (2016b) for more details.
1.1.2 Characterizing uncertainty using relative entropy

The PC method provides us a means to robustly and efficiently compute the output PDF of a computationally expensive model. The information content of a PDF can then be quantified using measures from information theory. The fundamental measures in information theory are entropy, mutual information and relative entropy, which are defined as functionals of probability distribution (Cover and Thomas, 1991). The entropy, a measure of the averaged uncertainty of a random variable, is a fundamental measure of uncertainty and arguably the best measure suited for a general evaluation of uncertainty (DelSole and Tippett, 2007). The entropy of a random variable $X$ can be defined as:

$$H(X) = - \int p(x) \log_2 p(x) \, dx$$ (1.2)

where $p(x)$ is the PDF of a random variable and logarithms to base 2 is usually used.

When we have two random variables $X$ and $Y$, the conditional entropy can be defined as $H(X|Y)$. The mutual information is a measure of the dependency of two random variables. The mutual information can also be considered a measure of the uncertainty reduction of one random variable due to another random variable, which can be expressed as:

$$I(X,Y) = H(X) - H(X|Y) = \int p(x, y) \log_2 \left[ \frac{p(x, y)}{p(x)p(y)} \right] \, dx \, dy$$ (1.3)

where $p(x, y)$ is the joint distribution of $X$ and $Y$.

Relative entropy measures the “distance” between two random variables, which can be expressed as (Kleeman, 2002):

$$D(p||q) = \int p(x) \log_2 \left[ \frac{p(x)}{q(x)} \right] \, dx$$ (1.4)
where \( q, p \) denote the PDF of two distributions (following the notation in Cover and Thomas (1991), we use \( p \) and \( q \) to distinguish two different distributions). For vector quantities, \( p \) and \( q \) are high dimensional joint probability distributions. For scalar quantities, \( p \) and \( q \) are univariate probability distributions. Mutual information is a special case of the relative entropy and their relationship can be explained by \( I(X,Y) = D(p(x,y)||p(x)p(y)) \).

Information theory based measures have been used in a variety of disciplines, such as statistical physics, computer science, probability and statistics (Cover and Thomas, 1991). In the last decade, information theory based measures have been applied in atmospheric science and ocean science, see Kleeman (2011) for an overview. The relative entropy, also called Kullback-Leibler divergence, is a natural measure for quantifying the efficiency of an ensemble perturbation scheme and the predictability of an ensemble system. Relative entropy can be used for different purposes depending on the definition of \( p \) and \( q \) shown in Equation 1.4. In order to examine the potential values of a dynamical forecasts relative to the best available initial oceanic state estimate, Kim et al. (2009) quantified the predictive information content in the forecast ensemble using relative entropy, in which \( p \) and \( q \) are selected to be initial ensemble distribution and forecast ensemble distribution. When \( p \) and \( q \) are chosen to be the forecast ensemble distribution and climatological distribution, the relative entropy can be considered as a measure of model predictability (Kleeman, 2002). In a predictability study, the climatological distribution is expected to contain all possible scenarios and thus has a “broader” distribution than that of the forecast. The forecast distribution usually broadens as the forecast time increases because of the forecast error growth. Finally, the forecast will lose its predictability when the difference between the two distributions vanishes, which means the forecast is no better than a randomly drawn state from the climatology. The predictability limit is then defined as
the time when the forecast PDF no longer differs from the climatological PDF (DelSole and Tippett, 2007). There are challenges when applying relative entropy to practical predictability problems. One is the definition of climatology. The ideal climatological PDF should be obtained from long time observational data. If the observational data is not available, climatological PDF may be constructed from long time model simulation after invoking the perfect model assumption, in which the observed state need to be captured by the model predictions. When the climatological data is built using real observations, model bias need to be taken into account. The relative entropy will be uncomputable when the probability supports of the climatological PDF $q$ and the forecast PDF $p$ are not overlap. Another challenge is the calculation of the high dimensional joint probability since probabilities are inherently exponentially-sized objects. When we treat the PDF of the field quantity as a joint distribution, the dimensionality of the problem may require us to manipulate the probability by making simplifying assumptions. Using the model-based climatology, the information theory based predictability measure has been recently applied in geophysical fluid dynamics, especially in the study of atmospheric and ENSO predictability (Schneider and Griffies, 1999; Kleeman, 2002; Kleeman and Majda, 2005; Tang et al., 2008; Cheng et al., 2011). We will use the relative entropy measure for two purposes: 1) quantify the difference of information content in the forecast PDF generated by different EOF perturbations; 2) quantify the predictability of the HYCOM Gulf of Mexico forecast ensemble using observation-based climatology.

1.1.3 Updating input uncertainties using observational data

When observational data is available, forecasters can update their estimates of the input uncertainties. The most general procedure revolves around Bayes theorem.
The practical implementation of this theorem can take many form, such as Bayesian optimization and Bayesian inference.

Bayesian optimization (Brochu et al., 2010; Snoek et al., 2012) is used to find the best uncertain input values that minimize the objective function defined by the Root Mean Square (RMS) error between the satellite observational data and the associated model simulations. Bayesian optimization is one of the most efficient approach in terms of the number of model evaluations required. The main idea of Bayesian optimization is to maintain a Gaussian Process (GP) model which approximates the underlying objective function and update this GP model at each optimization step. The GP model delivers not only the mean of the estimated objective function but also the variance or uncertainty associated with the estimated mean. The details of the GP model will be presented in Chapter 2. Bayesian optimization finds the extrema in the objective function in an adaptive manner, in which the evaluation point at each optimization step is determined adaptively by exploring the high uncertainty area and exploiting the most likely area for the minimum or maximum value. Bayesian optimization is suitable for our problem since the evaluation of the objective function, which involves model simulations, is computationally expensive. Additionally, Bayesian optimization does not require the convex properties and the derivatives of the objective function.

Bayesian inference is used to infer the full posterior distribution of the uncertain inputs. Specifically, an adaptive Metropolis algorithm is used to sample the posterior distribution which potentially involves tens of thousands of model simulations. To accelerate this process, we first construct a PC surrogate and then sample the surrogate instead of the original model. Bayesian inference provides us the whole posterior distributions of the uncertain inputs which include much more information compared with the single optimal values obtained from the Bayesian optimization
approach. The trade-off is that Bayesian inference usually requires much more model simulations even if the PC surrogate is used to accelerate this process.

1.1.4 HYCOM Experiments

Two recent efforts (Iskandarani et al., 2016a; Li et al., 2016) have relied on PC methods to quantify the uncertainties in forecasting the circulation in the Gulf of Mexico stemming from uncertainties in the initial conditions alone (Iskandarani et al., 2016a) or in combination with wind forcing uncertainties (Li et al., 2016). Both studies were based on perturbing the model fields (initial conditions and wind forcing) with space-time patterns obtained from an Empirical Orthogonal Function (EOF) decomposition where the amplitudes of these patterns were considered as uncertain input parameters. The initial condition experiment (Iskandarani et al., 2016a) retained only 2 initial condition EOF modes resulting in a two-dimensional parameter space that could be explored well with a 49-member ensemble. The parameter space of the second experiment was much larger (8-dimensional): it included the amplitudes of two additional initial condition EOF modes (to explore the impact of additional uncertainties in these fields), as well as four amplitudes of wind forcing EOF modes that constitute an additional source of uncertainty; the enlarged parameter space called for a much larger ensemble to compute the output uncertainties.

Both experiments relied on the PC methodology to generate the ensemble and to propagate the uncertainties forward efficiently. The choices made during the course of the uncertainty analysis have raised a number of issues that we wish to address here concerning the “realism” of the uncertainty analysis, the computational and information trade-offs in choosing different uncertain inputs, and the exploration of the statistical information conveyed by the PC approach. Specifically, we 1) validate the perturbation of the PC ensemble by comparing the ensemble against observational
data both at the surface and at depth; 2) leverage the ability of PC methods to deliver output PDF to quantify, using information theoretical measures, the variability loss by omitting sources or variability of input uncertainty. A second aim is to explore the statistics of the ensemble. In order to obtain the most representative ensemble member and to identify the outliers, contour boxplot (Whitaker et al., 2013), a generalization of the conventional boxplot, is applied to the ensemble. Furthermore, the output PDF delivered by the PC method are used to explore the non-Gaussian statistics in the vicinity of the LC region.

The layout of this dissertation is as follows. Chapter 2 introduces methods used in the dissertation. Chapter 3 describes the HYCOM model configuration and ensemble generation procedure. Chapter 4 investigates the information trade-offs between the number of ensemble members and the amount of uncertainties included. Chapter 5 explores the statistics using the PC approach and presents the refinement of the uncertain inputs using observational data. Chapter 6 summarizes the dissertation.
Chapter 2

Methodology

The first step of Uncertainty Quantification (UQ) is to identify the uncertainty sources. For an ocean general circulation model, these uncertainty sources can be model initial conditions, boundary conditions (such as bottom boundary, lateral boundary in regional model, fluxes) and model parametrizations of small scale processes. The uncertainty sources will generate an uncertainty space and the dimension of the space depends on the number of the uncertain inputs. All UQ tasks mentioned in the introduction require us to explore the uncertain space through sampling.

One approach belongs to the class of Monte Carlo (MC) (Gilks et al., 1995) methods, which rely on direct sampling of the model output to construct the PDF and their drawback concerns the large number of simulations required to build robust estimates of the output statistics. The computational cost of a large ensemble size is impractical whenever the forward model is computationally expensive and/or the uncertainty quantification analysis must be performed in a relatively short period. Other approaches to UQ rely on what could be described as indirect sampling. A functional relationship between the model output and the uncertain inputs is constructed using a relatively small ensemble and reliable statistics are estimated based on a much larger ensemble generated by sampling this functional relationship. The reli-

12
ability and efficiency of this approach hinges on producing an efficient and accurate functional relationship for the forward model. And this functional relationship is often called “surrogate”, “emulator”, “proxy” and “meta-model” in the literature. We will use the word “surrogate” hereafter. Polynomial Chaos (PC) expansion and Gaussian Process (GP) regression are two such techniques to construct the surrogate. They both rely on the same philosophy of building a functional representation for the model output that is both computationally efficient and accurate. In the following section we present a brief overview of these techniques.

2.1 Polynomial Chaos

The PC paradigm is based on describing the functional relationship of the model output and the uncertain inputs in the form:

\[ u(x, t, \xi) \approx u_P(x, t, \xi) = \sum_{n=0}^{P} \hat{u}_n(x, t)\psi_n(\xi) \]  

(2.1)

Here \( u(x, t, \xi) \) is the model output that depends on space \( x \), time \( t \) and uncertain input variables \( \xi \) encapsulating the uncertainty in the input parameters with PDF \( p(\xi_i) \); \( u_P \) is its truncated series representation and \( P \) is the number of terms in the truncation series; \( \hat{u}_n(x, t) \) are the PC coefficients; and \( \psi_n(\xi) \) are basis functions in the \( \xi \)-space. The basis functions are chosen to form an orthogonal basis in the weighed inner product defined by

\[ \langle u, v \rangle = \int u(x, t, \xi) v(x, t, \xi) p(\xi) \, d\xi \]  

(2.2)

(see Tab 2.1 for some common examples).
Table 2.1: Some common input PDFs and their associated orthogonal polynomial basis. Figure 2.1 shows the first 6 Legendre polynomials.

<table>
<thead>
<tr>
<th>PDF</th>
<th>Orthogonal Polynomial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>Hermite</td>
</tr>
<tr>
<td>Gamma</td>
<td>Laguerre</td>
</tr>
<tr>
<td>Beta</td>
<td>Jacobi</td>
</tr>
<tr>
<td>Uniform</td>
<td>Legendre</td>
</tr>
</tbody>
</table>

The appearance of the input PDF in the inner product definition simplifies the calculation of the statistical moments of the output. The statistics of the uncertain output can be efficiently described by the polynomial series that are orthogonal with respect to their PDF. The expectation or the mean of the model output is:

\[
E[u] = \int u(\xi) p(\xi) d\xi = \langle u, \psi_0 \rangle = \hat{u}_0
\]

(2.3)
The mean of the model output is unaffected by truncation when the inner product is evaluated exactly; it does however include aliasing errors when quadrature rules are used to approximate the inner products. The variance of \( u \) can also be obtained easily via the polynomial expansion:

\[
E[(u - E[u])^2] = \sum_{m,n=1}^{\infty} \hat{u}_m \hat{u}_n \int \psi_m \psi_n p(\xi) d\xi = \sum_{m,n=1}^{\infty} \hat{u}_m \hat{u}_n \langle \psi_m, \psi_n \rangle
\]

\[
= \sum_{n=1}^{P} \hat{u}_n^2 \| \Psi_n \|^2 + \sum_{m=P+1}^{\infty} \hat{u}_n^2 \| \Psi_n \|^2 \tag{2.4}
\]

The truncated series underestimates the variance but the error remains small if the amplitude of the higher order terms is small. Monitoring the variance is hence a good indicator whether the series includes enough terms or not.

The series truncation at the \( P \)-th (Equation 2.1) term controls its accuracy and the computational cost required to calculate the \( (P + 1) \) coefficients (the constant term). Once the \( \hat{u}_n \) are available the series can be used as a surrogate for the model and sampled much more efficiently. The bulk of the computational burden is then associated with the determination of these coefficients.

Two classes of methods are usually used to determine the PC coefficients - intrusive and non-intrusive. In the intrusive approach, the PC expansion will directly be plugged into the governing equations and the PC coefficients will be determined using a Galerkin projection. It requires the user to deal with the legacy code which may not be efficient. The so-called non-intrusive approach is of interest in this thesis since the coefficients can be determined without any model modification. Several non-intrusive approaches are available to determine these coefficients, namely projection approach, and regression approach.
2.1.1 The projection approach

The PC coefficients can be obtained by a Galerkin-type projection, which projects random variables onto an orthonormal polynomial space. The projection method aims to exploit the orthogonality of the truncated basis to determine the PC coefficients.

The projection procedure is based on minimizing the norm of the error between the true solution and its spectral representation: $\epsilon = u - u_P$. The error norm is the one induced by the inner product so that the problem becomes one of minimizing $\langle \epsilon, \epsilon \rangle = \langle u - u_P, u - u_P \rangle$. The minimum occurs for those values of $\hat{u}_m$ where

$$\frac{\partial}{\partial \hat{u}_m} \langle u - u_P, u - u_P \rangle = -2 \left( \frac{\partial u_P}{\partial \hat{u}_m} , u - u_P \right) = 0 \tag{2.5}$$

And hence the requirement that

$$\langle \psi_m, u_P \rangle = \langle \psi_m, u \rangle, \ m = 0, 1, 2, \ldots, P \tag{2.6}$$

follows from Equation 2.1 since $\frac{\partial u_P}{\partial \hat{u}_m} = \psi_m$. Inserting the expression for $u_P$ and invoking the orthogonality of the basis function, we can obtain an explicit expression for the coefficients, namely:

$$\hat{u}_m = \frac{\langle u, \psi_m \rangle}{\langle \psi_m, \psi_m \rangle}, \ m = 0, 1, 2, \ldots, P \tag{2.7}$$

The inner product in Equation 2.7 can be expressed in the form:

$$\langle u, \psi_m \rangle = \int \psi_m(\xi) u(\xi) p(\xi) \, d\xi \approx \sum_{i=1}^{S} \psi_m(\xi_i) u(\xi_i) \omega_i \tag{2.8}$$

$$\langle \psi_m, \psi_m \rangle = \int \psi_m(\xi) \psi_m(\xi) p(\xi) \, d\xi = D_{m,m} \tag{2.9}$$
Equation 2.8 approximates the integral numerically using quadrature. The $\omega_i$ refer to the quadrature weights at the quadrature roots $\xi_i$. For generalized PC basis, and classical PDFs, $\langle \psi_m, \psi_m \rangle$ are known analytically.

Again, the main computational expense is the evaluation of $u(\xi_i)$ since it requires a realization of the forward model for each quadrature point. A number of quadrature rules are available that maintain the orthogonality of the basis functions such as tensorized Gauss quadrature rules and Smolyak quadrature. Additionally, adaptive integration rules can be used to sample efficiently by adding samples along specific directions based on user selected criteria, such as the variance of the quantities of interest. Detailed discussion of numerical integration is beyond the scope of this dissertation, and we refer the interest reader to the following papers (Gerstner and Griebel, 1998, 2003; Huan, 2010; Constantine et al., 2012; Conrad and Marzouk, 2013; Winokur et al., 2013).

2.1.2 The regression approach

The projection approach provides an efficient way to estimate the PC coefficients by exploiting the orthogonality properties of PC basis functions. However, the projection approach requires model realizations on structured quadrature points based on specific quadrature rules. Many times, model instability will lead to model crashes at a few quadrature points (Li et al., 2016), which will make the projection approach inappropriate. Another way to estimate the PC coefficients is to use regression based approach (Blatman, 2009). The regression approach provides more flexibility in terms of the model evaluation points.

The evaluation of the Equation 2.1 at the $S$ sampling points $\xi_i$ leads to the following equation (for simplicity, time and space dimensions are dropped):
\[ u(\xi_i) = \sum_{k=0}^{P} \hat{u}_n \psi_n(\xi_i) \] or in matrix form \( \mathbf{\Psi} \hat{\mathbf{u}} = \mathbf{u} \) \hspace{1cm} (2.10)

Here \( \mathbf{\Psi} \) is a \( S \times (P + 1) \) matrix whose entries are given by \( \mathbf{\Psi}_{i,n} = \psi_n(\xi_i) \); \( \hat{\mathbf{u}}^T = (\hat{u}_0, \ldots, \hat{u}_P) \) is the vector of PC coefficients and \( \mathbf{u}^T = (u(\xi_1), \ldots, u(\xi_S)) \) is the \( S \times 1 \) vector of model samples. The computation of the vector \( \mathbf{u} \) constitutes the bulk of the cost for computationally intensive models because each sample requires a model realization; the remaining cost is associated with the solution of the linear system in Equation 2.10.

If the number of sampling points \( S = (P + 1) \) the matrix \( \mathbf{\Psi} \) is square and admits a solution of the form:

\[ \hat{\mathbf{u}} = \mathbf{\Psi}^{-1} \mathbf{u} \]

This is equivalent to requiring the series to \textbf{interpolate} the forward model run at the sampling points, thus \( u(\xi_i) = u_P(\xi_i) \) for \( 1 \leq i \leq S \). The choice of sampling points is a delicate matter since an ill-conditioned matrix \( \mathbf{\Psi} \) may lead to an unstable calculation and to large oscillations known as the “Runge phenomenon”.

If the number of sampling points \( S > (P + 1) \) then the system is over-determined and a \textbf{least square} approach must be adopted to determine the coefficients. This is achieved by minimizing the L2 norm of the residual \( \mathbf{r} = \mathbf{u} - \mathbf{\Psi} \hat{\mathbf{u}} \), i.e. find \( \hat{\mathbf{u}} \) that minimizes \( ||\mathbf{r}||_2^2 = \sum_{i=1}^{S} r_i^2 \). The solution can be easily shown to take the form:

\[ \hat{\mathbf{u}} = \left( \mathbf{\Psi}^T \mathbf{\Psi} \right)^{-1} \mathbf{\Psi}^T \mathbf{u} \]

\hspace{1cm} (2.11)

(Blatman, 2009).

Interpolation can be seen as a special case of least-square where enough degrees of freedom are retained in the series to drive the residual \( ||\mathbf{r}||_2 \) to zero. In both
instances the sampling points must be chosen carefully to avoid ill-conditioning of $\Psi$. Empirically, satisfactory results may be obtained with a number of model evaluations given by $S = n \times (P + 1)$ with $n = 2 - 3$ (Blatman, 2009).

If the number of sampling points $S < (P + 1)$ then the system is under-determined and a proper method need to be introduced to calculate the coefficients. This is an ill-posed problem with an infinite number of solutions. However, a sparsity requirement for the series may be invoked to determine $\hat{u}$. The solution of $\hat{u}$ is sparse when a small number of coefficients is sufficient to represent the function accurately. The idea is to find the sparsest solution which can be expressed as follows:

$$\min ||\hat{u}||_0 \text{ subject to } \Psi \hat{u} = u$$

(2.12)

where $||\hat{u}||_0$ is the L0 norm of $\hat{u}$ which represents the number of non-zero coefficients.

However, this solution is not always unique due to the combinatorial search (Muthukrishnan et al., 2005). An alternative is to relax the problem using the L1 norm instead of L0 norm. The minimum one-norm solution of this under-determined problem can be obtained using the Basis Pursuit method (Chen et al., 1998) which can be expressed as:

$$\min ||\hat{u}||_1 \text{ subject to } \Psi \hat{u} = u$$

(2.13)

where $||\hat{u}||_1 = \sum_{k=o}^{P} |u_k|$

If the model response includes noise that we would like to exclude from the proxy, the constraint in the basis pursuit can be relaxed to:

$$\min ||\hat{u}||_1 \text{ subject to } ||\Psi \hat{u} - u||_2 \leq \sigma$$

(2.14)

This form is called Basis Pursuit Denoising (BPDN) method (Van Den Berg and Friedlander, 2008). The choice of the noise level $\sigma$ is important. If $\sigma$ is too small,
we may suffer from over-fitting the data, while if \( \sigma \) is too large, we may obtain a surrogate with low accuracy. In order to systematically select \( \sigma \), we adopt the cross validation approach proposed in Doostan and Owhadi (2011); Peng et al. (2014). The details of determining \( \sigma \) using cross validation algorithm can be found in Algorithm 1. The BPDN algorithm used in this dissertation is a Matlab solver called SPGL1 (Van Den Berg and Friedlander, 2007).

### 2.1.3 Sensitivity analysis

The variance-based sensitivity analysis, often referred to as the Sobol method (Sobol, 2001), provides a powerful framework for global sensitivity analysis. In this approach, the contributions of the model inputs to the variance of the model output are decomposed into individual model input effects and interaction effects with other model inputs. The Sobol method is based on an orthogonal Sobol-Hoeffding decomposition (Hoeffding, 1948), which partitions the model output into terms of increasing dimensionality. It can be expressed as:

\[
 u(\xi_1, \ldots, \xi_k) = u_0 + \sum_{i=1}^{k} u_i(\xi_i) + \sum_{1 \leq i < j \leq k} u_{i,j}(\xi_i, \xi_j) + \cdots + u_{1,2,\ldots,k}(\xi_1, \ldots, \xi_k) \quad (2.15)
\]

Thus, the variance of each part in this decomposition can be related to the contributions of different individual variables. The Sobol sensitivity indices are defined as:

\[
 S_{i_1\ldots i_s}(u) = \frac{V[u_{i_1\ldots i_s}]}{V[u]} \quad (2.16)
\]

which gives a fraction of the variance due to the combination of a set of parameters \((\xi_{i_1}, \ldots, \xi_{i_s})\). However, there are \(2^d-1\) terms in Equation 2.16 which accounts for the individual effects and complex interactions of all variables, simpler characterizations are thus needed.
The first order sensitivity indices and total order sensitivity indices are defined in order to simplify the sensitivity measures. The former one describes the effect of individual model inputs on the model output without any interactions with others; while the latter one measures the “total effect” of a specific model input which accounts for the variance of the model output due to its individual effect and all its interactions with other parameters. Furthermore, the PC approach provides a straightforward way to calculate the first order and total order sensitivity indices. The details of estimating sensitivity indices using PC approach can be found in Appendix C.

2.2 Gaussian Process

Gaussian Process (GP), also referred to as kriging in geostastics (Matheron, 1967), is another popular approach for UQ. By definition, GP is an extension of multivariate Gaussian distribution to an infinite dimension stochastic process, any finite number of which have a multivariate Gaussian distribution (Rasmussen and Williams, 2006). Like a Gaussian distribution is a distribution over a random variable, GP is a distribution over a function which can be expressed in the following form:

\[ f(\xi) \sim \mathcal{GP}(m(\xi), k(\xi, \xi')) \] (2.17)

where a GP can be described by a mean function \( m(\xi) \) and a covariance function \( k(\xi, \xi') \) denoting the input as \( \xi \). We can simply consider GP as a function and it returns a Gaussian distribution instead of a scalar value at each given input point. As a result, the prediction of a GP provides not only the most likely estimate, the mean, but also the uncertainty associated with the estimate, the variance. For simplicity, the mean function \( m(\xi) \) of GP is very often assumed to be zero or constant everywhere. What relates all the data is the covariance function \( k(\xi, \xi') \). The covariance
function can be specified in different forms, we first restrict our attention on a popular covariance function choice - squared exponential covariance function which has the form (1D):

\[ k(\xi, \xi') = \sigma_f^2 \exp\left(-\frac{(\xi - \xi')^2}{2l^2}\right) \]  

(2.18)

Here \( \sigma_f^2 \) is the maximum allowable variance and \( l \) is the length scale. These two parameters are often referred to as hyperparameters in GP literature. When \( \xi \) and \( \xi' \) are close to each other, \( \xi \approx \xi' \), the covariance function \( k(\xi, \xi') \) achieves its maximum, which means \( f(\xi) \) and \( f(\xi') \) have perfect correlation. When \( \xi \) and \( \xi' \) are far apart, the covariance function \( k(\xi, \xi') \) approaches zero, which means \( f(\xi) \) and \( f(\xi') \) have no correlation. The hyperparameter, \( l \), defines how far \( \xi' \) needs to moved from \( \xi \) before \( f(\xi') \) loses correlation with \( f(\xi) \). The squared exponential covariance function is infinitely differentiable and is thus very smooth. Such strong smoothness assumption is unrealistic for modeling many physical processes as argued by Stein (1999), and Matérn class covariance function was suggested. Fig 2.2 shows the shape of different squared exponential covariance functions and Matérn covariance functions. The squared exponential is probably the most widely used covariance function within the kernel machines field (Rasmussen and Williams, 2006). There are also other different options to select covariance functions in GP literature, we refer the interested reader to Chapter 4 in Rasmussen and Williams (2006).
The covariance functions used in GP have flexible options. For example, we can easily fold a noise term into the covariance function in order to account the model noise:

\[ k(\xi, \xi') = \sigma_f^2 \exp\left(-\frac{(\xi - \xi')^2}{2l^2}\right) + \sigma_n^2 \delta(\xi, \xi') \]  

(2.19)

where \( \delta(\xi, \xi') \) is the Kronecker delta function and \( \sigma_n^2 \) is the maximum allowable variance in the error.

For problems with more than one dimension, reasonable GP covariance function needs at least one characteristic length scale to describe the surface structure in each dimension. The squared exponential covariance function with noise term in high dimensions can be expressed as:

\[ k(\xi, \xi') = \sigma_f^2 \exp\left(-\frac{1}{2}(\xi - \xi')^T M (\xi - \xi')\right) + \sigma_n^2 \delta(\xi, \xi') \]  

(2.20)
where $M = \text{diag}(l)^{-2}$ denotes the characteristic length scales in different dimensions and $l = (l_1, l_2, ..., l_d)$ is a vector of positive values.

One important feature of GP is the existence of analytical solutions. GP assumes the model output can be represented as samples from a joint Gaussian distribution which can be expressed as (for notational simplicity, we take the mean function to be zero):

$$
\begin{bmatrix}
  \mathbf{u} \\
  \mathbf{u}_* 
\end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix}
  \mathbf{0} \\
  K + \sigma_n^2 I \\
  K_*^T \\
  K_* \\
  K_{**}
\end{bmatrix} \right)
$$

(2.21)

Here $\mathbf{u}$ is the actual model realization (also called training set) and $\mathbf{u}_*$ is the surrogate solution according to the prior knowledge. The covariance function can be calculated following Equation 2.22 & 2.23 (only one sample is illustrated, more samples can be added by simply enlarging the $K_*$ and $K_{**}$ matrix).

$$
K = 
\begin{bmatrix}
  k(\xi_1, \xi_1) & k(\xi_1, \xi_2) & \cdots & k(\xi_1, \xi_n) \\
  k(\xi_2, \xi_1) & k(\xi_2, \xi_2) & \cdots & k(\xi_2, \xi_n) \\
  \vdots & \vdots & \ddots & \vdots \\
  k(\xi_n, \xi_1) & k(\xi_n, \xi_2) & \cdots & k(\xi_n, \xi_n)
\end{bmatrix}
$$

(2.22)

$$
K_* = [k(\xi_*, \xi_1)k(\xi_*, \xi_2)\cdots k(\xi_*, \xi_n)] \quad K_{**} = k(\xi_*, \xi_*)
$$

(2.23)

According to the Bayes theorem, we can obtain the conditional probability of the model response to specific model inputs:

$$
p(\mathbf{u}_*|\mathbf{X}, \mathbf{u}, \mathbf{X}_*) \sim \mathcal{N}(\overline{\mathbf{u}}_*, \text{cov}(\mathbf{u}_*))
$$

(2.24)

where:

$$
\overline{\mathbf{u}}_* = K_*^T (K + \sigma_n^2 I)^{-1} \mathbf{u}
$$

(2.25)
\[ \text{var}(u_*) = K_{**} - K_*[K + \sigma_n^2 I]^{-1}K_*^T \] 

(2.26)

2.2.1 Hyperparameter selection

The GP model is determined once we specify the mean and covariance functions. However, the hyperparameters in the covariance function can change the GP model structure completely. Thus, the reliability of the GP model depends on how well we select the hyperparameters once the form of the covariance function is chosen. The basic idea of hyperparameter selection is to generate an ensemble of model simulations to update the prior using the Bayes theorem; thus leads to a revised posterior of the associated prior distribution. This idea is similar to the PC construction: run a small ensemble of model simulations to determine the PC coefficients. Here, we run a small ensemble of model simulations to determine the GP hyperparameters.

2.3 Updating uncertain inputs using satellite observations

Observational data allows us to update our knowledge of the uncertain inputs. In order to balance the trade-offs between the amount of information obtained for the uncertain inputs and the computational cost, we apply two different approaches in the dissertation. Specifically, we first use an optimization approach which aims to find the best uncertain inputs that minimizes the root mean squared error between the model simulations and the observations. Then, we apply the Bayesian inference framework to infer the complete probability distribution of the uncertain inputs. The latter approach requires much more model simulations.
2.3.1 Bayesian optimization

Bayesian optimization is a method to find the extrema of a computationally expensive objective function which can be expressed as:

\[ \mathbf{x} = \arg \max_{\mathbf{x}} f(\mathbf{x}) \]  

(2.27)

Here is the form for maximization and the minimization can be performed with the same form with a transformed function \( g(\mathbf{x}) = -f(\mathbf{x}) \).

The benefits of using Bayesian optimization is that it does not require the derivatives and the convexity properties to be available for the objective function. Bayesian optimization maintains a GP model to approximate the objective function at each optimization step. The GP model provides not only the estimated mean of the underlying objective function but also the variance associated with the estimated mean. This feature allows adaptive sampling of the next objective function evaluation point. The exploration and exploitation trade-offs need to be balanced since we need to sample both from the area of high uncertainty indicated by the high GP posterior variance and from the area of high likelihood indicated by the GP posterior mean.

The key innovation of Bayesian optimization is the use of acquisition function which combines the need to explore the high uncertainty area and to exploit the high likely area. For a maximization problem, the acquisition function is defined to find the area where the prediction mean of the GP model is high and the prediction uncertainty of the GP model is high. Among many different choices of acquisition functions, we adopt the confidence bound criteria in the following form:

\[ LCB(\xi; \{\xi_n, u_n\}) = \mu(\xi; \{\xi_n, u_n\}) - \kappa \sigma(\xi; \{\xi_n, u_n\}) \]  

(2.28)
where $\xi$ is the future evaluation point, $\{\xi_n, u_n\}$ are the previous evaluation points and their associated objective function values. $\mu$ and $\sigma$ are the mean and variance from the GP posterior. $\kappa$ is a tunable parameter which balances the exploration and exploitation trade-offs. Details of other acquisition functions, such as probability of improvement, expectation of improvement, can be found in Brochu et al. (2010); Snoek et al. (2012).

Figure 2.3 demonstrates the procedures of finding the maximum in the objective function using Bayesian optimization. The whole process starts with two observations. Then, a GP model is built based on these two observations and used to approximate the underlying objective function. An acquisition function is defined to determine the next model evaluation point. The acquisition function accounts for both the mean and the variance of the constructed GP model: mean provides the most likely value for the maximum; variance identifies the high uncertain area. After two iterations, the Bayesian optimization finds a reasonable maximum for the underlying objective function as shown in Figure 2.3.
Figure 2.3: An example of Bayesian optimization on an 1D toy problem. (picture from (Brochu et al., 2010)) The objective function is approximated by a Gaussian Process model where the posterior mean and variance of the Gaussian Process are shown in the figure. The solid line is the Gaussian Process mean, the shaded area is the Gaussian Process variance and the dash line represents the true underlying objective function. Black dots are the current observations and the red dot is the new observation. The bottom green curve is the acquisition function which determines where to evaluate next.
2.3.2 Bayesian inference

Bayesian optimization approach can efficiently find the uncertain inputs that minimize the root mean squared error between model simulations and associated observations. However, the point estimate can only partially reflect the underlying distribution of the uncertain inputs. The point estimate may not be appropriate when the distribution of the uncertain inputs is nearly uniform. Bayesian inference technique infers the whole distribution of the uncertain inputs which provides the complete information of the updated knowledge of the uncertain inputs given observations.

Let the observational data be arranged in a vector $H = (H_1, H_2, ..., H_N)$; their associated model predictions are denoted by $M = (M_1, M_2, ..., M_N)$ and let $\Theta = \theta_1, \theta_2, ... \theta_N$ be the vector of uncertain inputs. Bayes’ theorem can be invoked here to provide the posterior distributions of uncertain input distributions given observational data:

$$p(\Theta|H) \propto p(H|\Theta)p(\Theta) \quad (2.29)$$

where $p(\Theta)$ is the prior probability distributions of the uncertain inputs, which represent our prior knowledge about the uncertain inputs; $p(H|\Theta)$ is the likelihood that represents the probability of obtaining the data given a set of uncertain inputs; and $p(\Theta|H)$ is the posterior distribution of the uncertain inputs, which represent our knowledge given observational data.

Let’s assume the errors between the model simulations and the observations are uncorrelated normally distributed with zero mean and $\sigma^2$ variance. The likelihood function can then be expressed in a Gaussian form as:

$$p(H|\Theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(H_i - M_i)^2}{2\sigma^2}\right] \quad (2.30)$$
The variance $\sigma^2$ here is usually unknown a priori. Many times, we can treat it as a hyperparameter and infer it using observational data. For the prior distribution, we assume a uniform distribution in the absence of additional knowledge. Here, we use a Jeffrey's prior which can be formulated as:

$$p(\sigma^2) = \begin{cases} 
\frac{1}{\sigma^2} & \text{for } \sigma^2 > 0 \\
0 & \text{otherwise}
\end{cases}$$  \hspace{1cm} (2.31)

The posterior distribution of the uncertain inputs shown in Equation 2.29 can be estimated by drawing samples from it. Usually, Markov chain Monte Carlo (MCMC) sampling strategy should be applied when the space of the inferring parameters is multidimensional or the underlying structure of the distribution is complex. The specific MCMC algorithm used here is the adaptive Metropolis algorithm (Haario et al., 2001).

The MCMC algorithm is prohibitive expensive for computational demanding models since it usually requires tens of thousands of MCMC steps and each step involves a model simulation with different uncertain input values. PC method provides an efficient way to perform the MCMC algorithm: we first construct a PC surrogate and then sample the surrogate instead of the original model.

### 2.4 Remarks

- All surrogate methods, PC and GP, are ensemble based and are non-intrusive; that is no code modification is necessary and all that is needed is the ability to run the forward model at specified values of the uncertain inputs.

- PC method using projection approach requires model simulations on quadrature points, which is not flexible. The whole set may need to be discarded if a
single sampling point falls outside the operating regime of the forward model. PC method using regression approach and GP method provide more flexible sampling strategy.

- Bayesian optimization provides an efficient framework for global optimization when the objective is a computationally expensive black box function (HYCOM is one such case): we don’t know the convex properties and the derivatives of the underlying function.

- Bayesian inference delivers the complete posterior probability distribution of the uncertain inputs given observational data. However, the MCMC sampling strategy is impractical for computationally expensive models. PC surrogate provides a framework to accelerate the MCMC sampling procedure by first constructing an accurate and efficient surrogate of the original model and then sampling the surrogate instead of the original model.
Chapter 3

Model and ensemble prediction

The functional nature of PC method requires the uncertain inputs to be represented by continuous random variables. This is easy enough to do with parametric uncertainties, such as the UQ study we did for an oil plume model (Wang et al., 2015), but is more complicated to do with field quantities, such as initial conditions in an ocean model, because the number of uncertain inputs in the PC method is limited as sampling cost grows exponentially with the number of dimensions, the so-called “curse of dimensionality”. For a typical ocean general circulation model, it is impractical to perturb the field quantity in each grid point independently since the sample size required should be approximately the same as the degrees of freedom in the system ($\sim 10^5 - 10^8$) (Moore, 1999). One may argue that the actual degree of freedom of the model is far less than the dimension of the state vector as the components in the state vector are actually inter-dependent and are somehow linked to each other by various dynamical processes and constraints. We can then reduce the dimensionality of the problem by relying on reduced state space methods, such as singular vector and breeding vector techniques.

Most ensemble methods differ in their generation of perturbations. Singular vectors, also called optimal perturbations, quantify the growth of small perturbations
over a finite period of time and the leading singular vectors are the fast-growing perturbations which lead to large forecast uncertainty (Buizza and Palmer, 1995; Palmer et al., 1998; Moore, 1999; Yamaguchi and Majumdar, 2010). In order to calculate the singular vectors, a tangent linear version and an adjoint version of the dynamical model are needed. The breeding vector technique perturbs the “errors of the day” (Toth and Kalnay, 1997; Yin and Oey, 2007), in which the difference between the control forecast and the perturbed forecast is scaled down at fixed time intervals. Another way to construct the horizontal and vertical structure of the perturbation is Empirical Orthogonal Function (EOF) decomposition of a selected model time interval (Thacker et al., 2012; Li et al., 2016; Iskandarani et al., 2016a), which captures the structure that explains the maximum variance in a linear sense. The EOF decomposition approach is used in the experiment analyzed here. It should be noted that isolating the uncertainty of a single physical process via EOFs is complicated as multiple physical processes may contribute to the same EOF mode. Our interest is focused on the Loop Current and Loop Current eddies for short time forecast and thus a 14 day data-assimilating HYCOM time series was used for EOF analysis. The resulting initial condition EOF modes successfully captured the variability in the position and strength of the LC frontal eddies (Iskandarani et al., 2016a). Once the horizontal and vertical perturbation structures are selected, we need to determine the amplitudes of the perturbation. The PDFs of the uncertain inputs are selected to be uniformly distributed without additional information. When observational data is available, Bayesian inference techniques can be used to refine the input distribution, see Sraj et al. (2013) for an oceanographic application.
3.1 HYCOM setup

The forecast model is the Hybrid Coordinate Ocean Model (HYCOM, http://hycom.org/) with the configuration similar to the one operationally by US Navy during the period 2003-2014. The distinguishing feature of HYCOM is the use of the hybrid coordinate system (Bleck, 2002) in the vertical direction: isopycnic in stratified regions, terrain-following (sigma) in shallow coastal region, and z-level in the unstratified mixed layer. The model has a horizontal resolution of 1/25 degree and 20 vertical layers. The initial condition for the model is from a data-assimilating HYCOM run (Chassignet et al., 2007). Then, the model is integrated forward in time without data assimilation for 30 days from May 1, 2010 to May 30, 2010. This period coincides with the Eddy Franklin (Donohue et al., 2015) shedding event, during which the LC eddy and the LC frontal eddies may play an important role in the oil transport process. The computational domain is open along portions of its southern, eastern and northern boundaries, where values are provided by a lower resolution 1/12 degree North Atlantic HYCOM simulation (Chassignet et al., 2007). The model is forced by the 27 km resolution Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS) wind products (Hodur, 1997).

3.2 PC ensemble

The PC approach treats the perturbation amplitudes as continuous random variables characterized by their PDFs. In practice, specifying these PDFs can be difficult due to the scarcity of the observational data. Without additional information, the PDFs of the uncertain inputs are usually assumed to be uniformly distributed, see Thacker et al. (2015) for a detailed discussion of input uncertainties. The PC approach is however flexible in allowing the user to explore the impact of different inputs PDF at
little extra cost (Thacker et al., 2015) since the entire response surface is available. When there is additional information about the input PDFs, the user can then re-sample the response surface using different input PDFs without incurring additional model simulations.

The two HYCOM ensembles in Iskandarani et al. (2016a) and Li et al. (2016) relied on reduced state space methods (Kleeman, 2011) to characterize the input uncertainties, so that the “amount” of uncertainties analyzed was maximized while minimizing the number of uncertain inputs. More specifically, EOF decompositions were used to identify modes of variabilities in the initial conditions and wind forcing. The spatial patterns of the perturbation were thus provided by the EOFs while the time series were associated with their principal components. The perturbations could then be constructed by the sum of products of spatial patterns and time series as follows:

\[
\begin{align*}
    u(x, t = 0, \xi) &= u_0(x) + \alpha^{ic} \sum_{k=1}^{K^{ic}} \xi_k^i \lambda_k^i \mathcal{U}_k \\
    f(x, t, \xi) &= f_0(x, t) + \alpha^{w} \sum_{k=1}^{K^{w}} \xi_k^w \lambda_k^w \mathcal{F}_k
\end{align*}
\]

where \(u(x, t = 0, \xi)\) is the perturbed initial condition field, \(u_0\) is the unperturbed initial field, \(\alpha^{ic}\) is a coefficient that controls the size of the perturbation, \(\lambda_k^i\) are the singular values, \(\mathcal{U}_k\) are the initial conditions EOFs, \(-1 \leq \xi_k^i \leq 1\) are standardized uncertain input random variables controlling the amplitude of the initial condition EOFs, and \(K^{ic}\) refers to the number of EOF modes retained. The terms \(f, f_0, \alpha^{w}, \lambda_k^w, \mathcal{F}_k\) and \(K^{w}\) are the analogous quantities for the wind-forcing field.

The vector of uncertain input was set \(\xi^\top = (\xi_1^{ic}, \xi_2^{ic})\) in Iskandarani et al. (2016a) as only two initial conditions modes were retained (\(K^{ic} = 2\) and \(K^{w} = 0\)), whereas \(\xi^\top = (\xi_1^{ic}, \xi_2^{ic}, \xi_3^{ic}, \xi_4^{ic}, \xi_1^{w}, \xi_2^{w}, \xi_3^{w}, \xi_4^{w})\) in Li et al. (2016) where 4 initial conditions EOF
modes and 4 wind forcing modes were retained in the perturbations. The PC approach treats the perturbation amplitudes as independent continuous random variables characterized by their PDFs which were assumed to be uniform.

The initial condition EOF modes targeted uncertainties in the LC system and were obtained from a multivariate EOF analysis of a 14-day data assimilating HYCOM simulation. Iskandarani et al. (2016a) analyzed the first two modes and showed that the \( \xi^\top = (1, 1) \) led to a stronger frontal eddy in the northeast corner of the extended LC and an early separation of a LC eddy, whereas setting \( \xi^\top = (-1, -1) \) had the opposite effect. The wind forcing EOF modes were calculated from a 2 month time-series (May and June, 2010) of a 27 km resolution COAMPS simulation. The EOF analysis was performed on the wind velocity vectors \((u \text{ wind and } v \text{ wind})\) and then projected onto the wind speed and the wind stress vectors, which are the actual components of the HYCOM wind forcing inputs.

Figure 3.1 depicts the Sea Surface Height (SSH) spatial patterns associated with the first four initial condition EOF modes. Specifically, the first initial condition EOF mode can be associated with the presence of a frontal eddy in the LC region. The second initial condition EOF mode can be associated with the variability induced from the LC eddy. Figure 3.2 shows the wind speed spatial patterns and the wind stress vector associated with the first four wind forcing EOF modes. Specifically, the variability explained by the first and fourth EOF modes can be related to a strong anticyclone in the southeast of Gulf of Mexico. The variability explained by the second and third EOF modes can be associated with wind patterns along the Tave shelf and the Bay of Campeche.
Figure 3.1: SSH patterns for initial condition EOF mode1-mode4.

Figure 3.3 shows the cumulative variance explained by the initial condition and the wind forcing EOF modes. The exploratory study in Iskandarani et al. (2016a) included only the first two initial condition EOF modes in order to keep the computational cost tractable; their study thus included only about 50% of the total initial condition variance (with $\alpha^{ic}$ set to 1). Li et al. (2016) increased the number of initial conditions modes to four to explore the impact of additional variability on the forecast uncertainty, and included four wind forcing modes to account for additional sources of uncertainties. Note, however, that Li et al. (2016) had to decrease the size of their perturbation by setting $\alpha^{ic} = 0.8$ and $\alpha^w = 0.04$ in order to avoid repeated crashes of the forward model when the full perturbation was applied. Thus only 53.6% of
the variance was accounted for the initial condition EOF modes and 3% in the wind forcing EOF modes.

Figure 3.2: Wind speed patterns for wind forcing EOF mode1-mode4 and wind stress vectors are superimposed on the wind speed field.

The PC surrogate in Iskandarani et al. (2016a) relied on a Galerkin projection with sampling on the Gauss-Legendre quadrature points to determine the coefficients, and was shown to be valid for a period of 40 days when validated against independent model simulations. The eight-dimensional uncertain inputs space of Li et al. (2016) required a different surrogate construction approach and the latter was built using a Basis Pursuit Denoising algorithm (Van Den Berg and Friedlander, 2008). The validity of the surrogate model was also established using independent model simulations.
Here we focus on the first 30 days of the simulation when both surrogate delivered accurate representation of the model output.

![EOF analysis graphs](image)

**Figure 3.3:** The cumulative variance explained by the first 10 EOF modes, the red labels specify the cumulative variance explained by the first 4 EOF modes.

### 3.3 Model-data comparison

There are two common ways to assess the quality of the model simulations. One way is to make the perfect model assumption and treat the control run as the “truth”. Another way is to compare model simulations directly to the observational data. We follow the latter procedure in order to compare our ensemble to independently obtained data. For a “good” ensemble prediction, the forecast uncertainties should be properly represented (Slingo and Palmer, 2011) such that the true evolution appears to be a plausible realization in the ensemble. In order to assess whether the EOF perturbations satisfy this requirement, we compare the envelop of the model ensembles against observational data both at the surface and at depth.

#### 3.3.1 Comparison against satellite SSH

For the surface model-data comparison, we use the AVISO gridded satellite Absolute Dynamic Topography (ADT) product (https://www.aviso.altimetry.fr/en/my-
aviso.html) on a $1/4^\circ \times 1/4^\circ$ spatial resolution and a daily temporal resolution as the SSH observation. Although the use of SSH gridded data may not be well adapted for the precise characterization of the small scale processes, gridded satellite SSH data is suited for studying the relative large scale LC and LC eddies (Leben, 2005; Le Hénaff et al., 2014) which are the focus of this dissertation. Specifically, we compare the LC edge defined by the SSH 17cm contour, which corresponds to the approximate positions of the frontal positions of the LC eddies (Leben, 2005). In order to discard the SSH signal associated with the total volume change in the Gulf of Mexico, the instantaneous spatial mean over the entire Gulf of Mexico has been removed as suggested by Le Hénaff et al. (2012). Figure 3.4 shows the time snapshot of the LC edge every 10 days starting from May 1, 2010 for both HYCOM ensembles. The background color is the gridded satellite SSH data. The LC edge from HYCOM ensemble members (black contours) are compared against that from the gridded satellite SSH data (white contours).

A basic observation for both ensembles is that the satellite data shown in white contour falls within the envelope of the HYCOM ensemble contours shown in black contours (Figure 3.4) near the LC region, which indicates a “good” ensemble since the observational data appears to be a plausible realization of the model ensemble simulations. The LC eddy simulated by the HYCOM ensemble shows a trough in the north part of the eddy and a crest in the northeast and northwest part of the eddy. This is due to the development of a second LC frontal eddy in the northern LC region. In the western Gulf of Mexico, the ensemble members are slightly different from the observational data, especially in the later days. The difference between the observations and ensemble members in the western Gulf of Mexico may be caused by many reasons: 1) each ensemble member is a “free” run without data assimilation; 2) the initial condition perturbation is mainly focused on the LC region and there is little
perturbation in the western Gulf of Mexico. At day 30, it is clear that the LC eddy shedding process is affected by the ensemble perturbations, in which several ensemble members have already shed an eddy while others have not. The difference between the LC edge spread of these two HYCOM ensembles is small, which indicates that the additional ensemble members in the large ensemble did not add much variability for the LC edge.

3.3.2 Comparison against mooring data

In order to compare the model simulations with observational data at depth, we compare both of the HYCOM ensembles against 9 full-depth mooring observations deployed by the Bureau of Ocean Energy Management (BOEM)/Science Applications International Corporation (SAIC) (Hamilton et al., 2016) during the HYCOM simulation period (May 1, 2010 - May 30, 2010). Instead of comparing the point velocity measurements to the model directly, we compare the mean and standard deviation ellipses of the point velocity for the entire 30 days adopted from Xu et al. (2013).

Figure 3.5 shows the comparison of the mean and standard deviation ellipses between the HYCOM ensemble simulations and the mooring data at different depths. In each subfigure, the mean velocity vectors of the ensemble members are shown in black and the standard deviation ellipses of the ensemble members are presented in red. The mean and standard deviation ellipses of the mooring observations are in blue. At some locations (C1 and C2, moorings located at the top right corner), the small HYCOM ensemble (49 members) does not show enough variability compared to mooring observations. The large ensemble (798 members) captures the variability of the mooring observations at C1 and C2 positions better. In general, the mooring data approximately falls within the envelope of the HYCOM ensemble at different depths, which indicates that the ensemble at depth captures reasonably well the observations.
Figure 3.4: SSH 17 cm contours from HYCOM ensemble (black) and AVISO SSH (white) at indicated forecast time. The background is AVISO SSH data. Top: 49-member HYCOM ensemble; Bottom: 798-member HYCOM ensemble.
Figure 3.5: Mean velocity and standard deviation ellipses at different depths built from 9 Bureau of Ocean Energy Management (BOEM)/Science Applications International Corporation (SAIC) full-depth mooring data (blue) and HYCOM ensemble (black and red). The period is from 05/01/2010 to 05/30/2010. Left: 49-member HYCOM ensemble; Right: 798-member HYCOM ensemble.

Both at the surface and at depth, the ensemble is “realistic” as evidenced by the model-data comparison: the ensemble captured the observed evolution, especially in
the LC region. For the LC edge defined by the SSH 17cm contour, the small ensemble can capture similar amount of variability compared with that of the large ensemble at the surface and at depth.
Chapter 4

Sensitivity analysis and information trade-offs

The 798-member HYCOM ensemble aims at quantifying the combined impact of initial condition and wind forcing uncertainties in the Gulf of Mexico. In this chapter, we leverage the PC surrogate to perform sensitivity analysis and quantify the information trade-offs using information theoretic measures.

The construction of the PC surrogate here (Li et al., 2016) relies on Basis Pursuit Denoising technique. One of the key tunable parameter of this approach, noise level ($\sigma$ in Equation 2.14), needs to be adjusted using a computationally expensive K-fold cross validation approach (shown in Appendix A). The brute-force approach, which constructs PC surrogate independently at each HYCOM grid point using Basis Pursuit Denoising with cross validation, is impractical since there are about 100,000 model wet points in the Gulf of Mexico. The EOF-PC approach proposed by Li et al. (2016) is thus used here to reconstruct the SSH field. The basic idea is to reduce the dimensionality of the problem using EOF decomposition, construct PC surrogate in the EOF space and reconstruct the entire field in the end. The number of EOF retained represented 90% of the total variance; and we refer the interested reader to Li et al. (2016) for more details.
4.1 Sensitivity analysis

Figure 4.1 shows the time evolution of the mean and standard deviation of SSH estimated from the PC surrogate. During the 30-day simulation period, the mean SSH field shows the development of a LC frontal eddy indicated by the negative SSH near the LC neck. The LC eddy shows a potential separation at day 30. At the same time, there is another small weak LC frontal eddy developing in the north of the LC eddy which erodes the LC eddy to form a SSH trough. The ensemble standard deviation shows a clear signal starting at day 20 in the LC frontal eddy position and the signal is all over the LC region at day 30.

The PC approach empowers us to quantify the individual contributions and interactions of each uncertain inputs to the model output via the analysis of variance in a straightforward fashion (see details in Appendix C). There are 8 uncertain inputs in total and we sum all 4 initial condition and 4 wind forcing sensitivity indices to form a simplified measure. We will only focus on this simplified measure hereafter. Figure 4.2 shows the first order sensitivity indices of the SSH field to the initial condition uncertainties (left column) and the wind forcing uncertainties (right column) at day 1, day 10, day 20 and day 30. This figure illustrates the relative importance of the initial condition uncertainties and the wind forcing uncertainties on SSH as time evolves. The initial condition uncertainties mainly contribute to the center of the Gulf of Mexico while the main contribution of the wind forcing uncertainties is in the coastal region. Additionally, the wind forcing uncertainties also play a role in the western Gulf of Mexico at day 30. Figure 4.3 shows the interactions between the initial condition uncertainties and the wind forcing uncertainties estimated by the difference between the total order sensitivity indices and first order sensitivity indices. The signal of interactions appears after day 10 and the region of strong interactions is much smaller compared with that of the sensitivity region shown in Figure 4.2.
4.2 Quantifying information trade-offs using relative entropy

The PC approach suffers from the curse of dimensionality, increasing either the variability of the input uncertainties or the sources of the input uncertainties will lead to a much larger ensemble size. A natural question that arises is the model output (SSH) variability loss by using the small ensemble since it is computationally much cheaper than the large one. To address this question, we first explore the SSH variability loss by omitting the variability of the input uncertainties, in which the high order EOF modes are discarded. Then, we explore the SSH variability loss by omitting the sources of the input uncertainties, in which the initial condition uncertainties and the wind forcing uncertainties are treated separately. Since the PC surrogate for the large ensemble has already been built and validated by Li et al. (2016), we can explore the SSH variability loss by resampling the PC surrogate many times using different input perturbations. Several interesting questions that we wish to address are listed in Table 4.1. The information theoretic measure we used here is the relative entropy which can be defined in a discrete form as follows (Kleeman, 2002):

\[
D(p, q) = \sum_i p_i ln\left(\frac{p_i}{q_i}\right)
\]  

(4.1)

where \( p, q \) denote the PDFs of two distributions and \( i \) is the discrete bin index.
Figure 4.1: Time evolution of SSH mean (left) and standard deviation (right) estimated using PC surrogate.
Figure 4.2: Time evolution of the first order sensitivity indices for the initial condition uncertainties (left) and the wind forcing uncertainties (right).
Figure 4.3: Time evolution of the interaction terms in the sensitivity analysis for the initial condition uncertainties (left) and the wind forcing uncertainties (right). The interaction terms are defined by the difference between the total order sensitivity indices and the first order sensitivity indices.
Scenarios | Perturbations
--- | ---
1) SSH variability loss by omitting the high order EOF modes. | $p = (\xi_1, \xi_2, 0, 0, \xi_5, \xi_6, 0, 0)$

2) SSH variability loss by omitting the high order and the wind forcing EOF modes. | $q = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6, \xi_7, \xi_8)$;

3) SSH variability loss by omitting the high order and the initial condition EOF modes. | $p = (0, 0, 0, 0, \xi_5, \xi_6, 0, 0)$

4) SSH variability loss by omitting all EOF modes except the initial condition EOF mode1. | $q = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6, \xi_7, \xi_8)$;

5) SSH variability loss by omitting all EOF modes except the initial condition EOF mode2. | $p = (0, \xi_2, 0, 0, 0, 0, 0, 0)$

$\text{Table 4.1:}$ Left: Different uncertain input perturbation scenarios. Right: The uncertain input distributions in the relative entropy function form, where the shown variables will be perturbed uniformly from -1 to 1 and 0 means no perturbation. The perturbation amplitude is described by the standardized random variable $\xi = \frac{\theta - \theta_{\text{min}}}{\theta_{\text{max}} - \theta_{\text{min}}}$ where $\theta$ represents the actual uncertain input, the standard random variable, $\xi$, ranges from -1 to 1.

We calculate the probability in the discrete form, where the possible range each PDF ($p$ and $q$) is divided into small bins. The probability of a specific bin is equal to the number of samples in this bin divided by the total number of samples (100,000):

$$p_i = \frac{N_i}{\sum_{i=1}^{K} N_i}$$ (4.2)

where $N_i$ represents the number of samples in bin $i$, and $K$ is the number of bins. We choose $K$ equals 20 for all PDF calculations in our results. Previous investigations using different bin numbers show that the number of bins will not affect the results much when more than 20 bins are used in the calculation.

We first investigate the SSH variability loss by omitting the variability of the input uncertainties (scenario 1 in table 4.1). Figure 4.4 shows the relative entropy of the SSH between the ensemble that perturbs all 8 EOF modes and the ensemble that perturbs 2 EOF modes of each uncertain input sources at different times. The
result shows that the SSH variability loss by omitting the variability in the input uncertainties (EOF modes 3,4 of each uncertain sources) is little.

Next, we investigate the SSH variability loss by omitting the sources of the input uncertainties. Since Figure 4.4 shows that the high order EOF modes contribute little to the SSH variability loss, we focus only on the low order EOF modes here. Figure 4.5 shows the SSH variability loss by omitting the wind forcing input uncertainties (scenario 2 in table 4.1), in which the ensemble that perturbs only the first 2 initial condition EOF modes is compared with the fully perturbed ensemble. The impact of omitting the wind forcing EOF modes is mainly in the coastal region and little influence can be found in the LC region and the western Gulf of Mexico at day 30.

Figure 4.4: SSH variability loss by omitting the variability of the input uncertainties (scenario 1 in table 4.1) measured by the relative entropy between the ensemble that perturbs all 8 EOF modes and ensemble that perturbs 2 EOF modes of each uncertain input sources.
Relative entropy \( (p = 2IC; q = 4IC4WF) \)

**Figure 4.5:** SSH variability loss by omitting the wind forcing input uncertainties (scenario 2 in table 4.1) measured by the relative entropy between the fully perturbed ensemble and the ensemble that perturbs only its first 2 initial condition EOF modes.

Relative entropy \( (p = 2WF; q = 4IC4WF) \)

**Figure 4.6:** SSH variability loss by omitting the initial condition input uncertainties (scenario 3 in table 4.1) measured by the relative entropy between the fully perturbed ensemble and the ensemble that perturbs only its first 2 wind forcing EOF modes.
Relative entropy (\(p = \text{IC01}; q = 4\text{IC4WF})\)

Figure 4.7: SSH variability loss by retaining only the 1st initial condition EOF mode (scenario 4 in table 4.1) measured by the relative entropy between the fully perturbed ensemble and the ensemble that perturbs only its 1st initial condition EOF mode.

Relative entropy (\(p = \text{IC02}; q = 4\text{IC4WF})\)

Figure 4.8: SSH variability loss by retaining only the 2nd initial condition EOF mode (scenario 5 in table 4.1) measured by the relative entropy between the fully perturbed ensemble and the ensemble that perturbs only its 2nd initial condition EOF mode.

The SSH variability loss by omitting the initial condition EOF modes is shown in Figure 4.6 (scenario 3 in table 4.1). The coastal signal disappears in this case, which indicates that the first 2 wind forcing EOF modes dominate the SSH signal in the coastal region. Instead, strong signal is observed in the LC region and its magnitude grows as time evolves, which indicates that the initial condition uncertainties are the
dominant contributors to the SSH variability in the LC region. These results are consistent with the sensitivity analysis shown in Figure 4.2, in which the SSH in the LC region is more sensitive to the initial condition EOF modes while in the coastal region the SSH is more sensitive to the wind forcing EOF modes. Since the first two initial condition EOF modes contribute the most to the SSH variability in the central Gulf of Mexico, we further investigate the influence of each individual initial condition EOF mode. Figure 4.7 shows the SSH variability loss by retaining only the 1st initial condition EOF mode and Figure 4.8 shows the SSH variability loss by retaining only the 2nd initial condition EOF mode. Without the 2nd initial condition EOF mode (shown in Figure 4.7), the SSH variability loss is localized in the north side of the Loop Current, especially at day 30. Without the 1nd initial condition EOF mode (shown in Figure 4.8), the SSH variability loss is substantial in localized region along the Loop Current and its importance grows with time. In the central Gulf of Mexico, two initial condition EOF modes are necessary in order to capture most of the SSH variability. In the western Gulf of Mexico, the SSH variability loss is little for both cases.

In summary, for SSH, the model output variability loss by omitting the variability of the input uncertainties is small and two EOF modes for each uncertain input sources may be enough. Additionally, the SSH variability loss by omitting the wind forcing uncertainties is small in the LC region. These results suggest that the small HYCOM ensemble that perturbs only the first two initial condition EOF modes is suitable for studying the LC region since the number of HYCOM realizations required to build the PC surrogate grows exponentially with the number of uncertain inputs. It should be noted that the relative entropy approach here could be generalized to any ensemble simulations to balance the size of the ensemble and the model output variability contained in the ensemble.
Chapter 5
Statistical analysis and Bayesian inference

In this chapter, we use the PC surrogate to explore model statistics in some detail, calculate ensemble covariance for data assimilation, quantify predictability using information theory, and refine the uncertain inputs using observational data. The focus here is on the short-term dynamics of Loop Current and Loop Current eddies. As shown in the previous chapter, 2 initial condition EOF modes can capture the SSH variability resonably well, thus we will use the small ensemble to carry out the analysis.

5.1 Ensemble statistics

5.1.1 Contour boxplot

Ensemble simulations provide us a way to derive statistics of the model, which leads to an estimate of the confidence of the model prediction. However, mining useful information from the ensemble can be challenging since it usually involves a large number of single model simulations. In descriptive statistics, the boxplot is a standard way to display the distribution of a set of data points based on the five number summary: minimum, first quartile (25% percentile), median, third quantile (75%
percentile) and maximum. This simplest possible boxplot provides the full range of variations (from minimum to maximum) and the most likely range of variations (from first quantile to third quantile) in the data points. The boxplot provides an elegant way to visualize and summarize an ensemble of data points and the key step to generate the boxplot is to sort the data points by order. An important aspect of the boxplot analysis is that only actual model realizations, instead of an algebraic construction, are used, so that dynamical balances are respected for each ensemble member. However, in oceanography and meteorology, we are usually interested in functions (e.g. time series) and contours (e.g. eddy contours) instead of point values. The recently developed contour boxplot (Whitaker et al., 2013), built on the notion of data depth, enables us to find valuable information from ensemble of functions and contours. Here, we are particularly interested in answering: 1) what is the most representative ensemble member? 2) what are the outliers of the ensemble?

Simply speaking, the contour boxplot can be considered as a generalization of the conventional boxplot. Both of these methods require an ordering of the data. The difference between the boxplot and the contour boxplot is that the former can only be applied to point values while the latter can be applied to functions and contours. At the heart of the contour boxplot, a measure of centrality is defined by the concept of data depth proposed from the statistics community (Whitaker et al., 2013). The data depth concept quantifies the centrality or depth of a data sample with respect to an ensemble of data samples. The deeper data sample is considered to be more representative than the shallower data sample. The deepest data sample (the median) can be considered as the most representative ensemble member and the shallowest data samples can be considered as outliers. An important aspect of the box plots analysis is that only the actual model realizations instead of the algebraic constructions are used, which means the dynamical balances are realized.
for this setting of the uncertain inputs. We refer the interested reader to Whitaker et al. (2013); Mirzargar et al. (2014) for more descriptions of the methodology as well as an application in Meteorology. We apply the contour boxplot method to the LC edges obtained from the 49-member HYCOM ensemble. The key effort of the contour boxplot is to sort the 49 ensemble contours by their data depth defined in Whitaker et al. (2013).

Figure 5.1: Contour boxplot of the Loop Current edge derived from the 49-member HYCOM ensemble. All HYCOM ensemble contours are color coded according to the legend and their associated uncertain input perturbations are shown in the bottom left panel. The satellite SSH Loop Current edge is shown in black for reference.

Figure 5.1 shows the contour boxplot for the ensemble of the LC edges defined by the 17 cm SSH contour at day 30. All the contours are color coded according to the legend (e.g. median shown in yellow and mean in magenta) and their associated uncertain input perturbations (normalized) are shown in the bottom left panel of the figure. The satellite SSH observation is shown in black for reference. The satellite
observation shows that a LC eddy has already separated from the LC while the mean of the ensemble shows that the LC is still in its extended position. On the other hand, the most representative ensemble member identified by the median of the ensemble (in yellow) shows a similar LC eddy shedding stage compared with that of the satellite observations. We mark the “shallowest” three ensemble members as outliers shown in red. It is clear that these outliers are still in their early stage of the LC eddy shedding process, which is somewhat “slow” compared with other ensemble members. In the normalized uncertain input perturbations panel (bottom left panel in Figure 5.1), these outliers are the ensemble members with extreme negative perturbations in both of the first two initial condition uncertainties.

Next, we connect the uncertain input perturbations to the associated physical process. Figure 5.2 shows the initial perturbations approximated by the SSH difference between the perturbed and unperturbed simulations one day after the start of the simulation. The initial perturbation patterns are shown according to their normalized uncertain inputs presented in the middle of the figure. The bottom left subfigure represents the most negative perturbations for both of the initial condition EOF modes; the top right subfigure represents the most positive perturbations for both of the initial condition EOF modes. Most negative perturbation on EOF1 contributes to an SSH ridge in the northeast and southwest corner of the LC, which weakens the frontal eddy identified in the unperturbed state. A positive EOF1 perturbation would strength it. The dynamical process associated with the EOF1 perturbation (the $\xi_1$ direction) can be explained by the strengthening or weakening of the LC frontal eddies. Most positive EOF2 perturbation has a ridge at the location of the frontal eddy that does extend as far southwest as EOF1 perturbation. The EOF2 perturbation (the $\xi_2$ direction) may be associated with the variability induced by the LC eddy. These signals strongly affect the intensity of the LC frontal eddies, which play an
important role in the LC eddy shedding process (Le Hénaff et al., 2012, 2014). The outliers identified by the contour boxplot (Figure 5.1) are located at the most negative perturbation of the initial condition EOF model, which is consistent with the related dynamical process: negative perturbation on initial condition EOF model is associated with the weakening of the LC frontal eddies, which delays the LC eddy shedding event.

![SSH initial condition perturbation patterns](image)

**Figure 5.2:** SSH initial condition perturbation patterns. The SSH difference between the perturbed and the unperturbed simulations one day after the start of the simulation. The 17cm SSH contour of the unperturbed simulation is added for reference.

### 5.1.2 An exploration of SSH PDF

The estimate of the full PDF usually requires a large sample size. Traditionally, only the low order statistical moments are calculated from the direct ensemble members,
and these low order statistical moments are most meaningful when the underlying
PDF can be approximated by a normal distribution. The PC surrogate provides an
efficient way to estimate the full PDF of a model output. We thus investigate whether
the underlying distribution of the SSH field in the Gulf of Mexico is normal using
the PC surrogate constructed by the 49-member HYCOM ensemble. We first explore
the pointwise SSH PDF at selected locations in the Gulf of Mexico. In Figure 5.3,
the locations of four selected points (A1-A4) and the LC edges are shown in the
left subfigure. In the right subfigure, the SSH PDFs from selected locations are
presented. Clearly, the SSH PDFs are not always normally distributed. For example,
the SSH PDF at point A3 along the LC edge shows a bimodal distribution. Next, we
investigate where in the Gulf of Mexico the SSH is normally distributed. In order to
quantify the normality of the SSH at each grid point in the Gulf of Mexico, we need
a metric to measure the distance between an arbitrary distribution and a normal
distribution. We quantify the difference between the ensemble SSH distributions
and their Gaussian counterparts using relative entropy. The Gaussian counterpart
of a distribution is obtained from an analytical form using the mean and variance
estimated empirically from the ensemble simulations.

Figure 5.3: The SSH PDFs of selected points at day 30. The yellow contours on the
map shows the Loop Current edges in the ensemble simulations. The red contour is the
Loop Current edge for the unperturbed simulation.
Figure 5.4 shows the relative entropy map between the SSH PDF and its Gaussian counterpart in each grid point. We only plot the region where the associated relative entropy is larger or equal to 0.4 and the LC edges are also shown in the figure for reference. The strongest non-Gaussian signal appears in the LC region (indicated by red color), especially in the place where the variation of the LC edges is high.

![Relative entropy map](image)

**Figure 5.4:** The relative entropy of SSH PDFs to their Gaussian counterparts at day 30 (the color shows only the region where the relative entropy is greater than 0.4). The black contours are HYCOM Loop Current edges for reference.

### 5.1.3 Ensemble covariance

The PC surrogate has the potential to be used for data assimilation. It is important for a data assimilation system to accurately estimate the model forecast covariance
functions. The PC approach provides an efficient way to calculate the covariance between different model output fields. Equation 2.4 shows an example of how to calculate variance using PC surrogate in an efficient manner. Similar to the variance, the covariance can also be estimated using the PC surrogate in the following discrete form:

\[
E[(u - E[u])(v - E[v])] = \sum_{n=1}^{P} \hat{u}_n \hat{v}_n ||\Psi_n||^2 + \text{truncation error} \quad (5.1)
\]

where \(u\) and \(v\) represent two model outputs; \(\hat{u}\) and \(\hat{v}\) denote the associated PC coefficients; \(\Psi\) is the basis function and \(P\) is the number of terms in the truncated series.

Figure 5.5 shows the covariance between SSH at one point \((-86.12^\circ E, 23.841^\circ N)\) and the SSH at each grid point in the Gulf of Mexico at different days. Figure 5.6 shows the covariance between SSH at one point \((-86.12^\circ E, 23.841^\circ N)\) and the SST at each grid point in a vertical section along \(25.98^\circ N\) in the Gulf of Mexico at different days. In both cases, the difference between the PC covariance and the empirical covariance is small.
Figure 5.5: Covariance between SSH at one point (indicated by the black asterisk) and the SSH in the entire Gulf of Mexico. Empirical estimate is shown on the left, PC estimate is shown in the middle and the difference between the empirical estimate and the PC estimate is shown on the right.
Figure 5.6: Covariance between SSH at the same point shown in Figure 5.5 and the temperature along a vertical section (25.98°N). Empirical estimate is shown on the left, PC estimate is shown in the middle and the difference between the empirical estimate and the PC estimate is shown on the right.

5.1.4 Predictability

The 20 years of AVISO SSH data provides us an unique way to quantify the predictability of the ensemble prediction. We use this relative long record of observational data to form an AVISO climatology. The basic idea is to compare the “uncertainty” between the ensemble simulations and the climatological data. If the ensemble simu-
lations at a specific space and time contains more “uncertainty” compared with that of the climatological data \(^1\), we may conclude that predictability is lost since the model provides a prediction which is no better than a random draw from the climatological data. The different definitions of the “uncertainty” can lead to different approaches. One approach is to characterize the “uncertainty” using the standard deviation which will lead to a fairly reasonable approximation if the model and the climatology is unbiased (no shifts in the means) and the underlying distribution is Gaussian. Another approach is to characterize the “uncertainty” using its PDF which encapsulates all its statistical information. Here, we explore both of these definitions. Reducing the information contained in the “uncertainty” by using standard deviation instead of PDF ignores the difference in the first order statistical moments (signal), and implicitly assumes a Gaussian distribution for both model and climatological PDFs. Figure 5.7 shows the predictability limit defined by the ratio between the model standard deviations and the climatological standard deviations. The ensemble prediction will lose its predictability when this ratio is greater than one: the standard deviation in the ensemble prediction is larger than that from the observations.

At day 30, a very strong signal is detected localized in the LC pinch-off region, which is probably due to the abrupt SSH changes caused by the LC eddy shedding process.

\(^{1}\)For climatological data, there is no time dimension, which means all historical data at a specific spatial point will be used to form the “uncertainty”.
Figure 5.7: The SSH predictability measured by the ratio between the standard deviations between the HYCOM ensemble simulations and the climatological data. The red contours identify the region where the ratio is greater than 1.

Information theory provides a complete characterization of the “uncertainty” in the model simulations and the climatological data via their PDFs. In equation 4.1, when $q$ is chosen to be the climatological distribution and $p$ is selected as the model forecast distribution, the relative entropy can be considered as a measure of model predictability. The predictability limit is then defined as the time when the forecast PDF no longer differs from the climatological PDF (DelSole and Tippett, 2007). The relative entropy measure has been recently applied to quantify oceanic predictability (Kim et al., 2009), atmospheric and ENSO predictability (Schneider and Griffies,
In a predictability study, the climatological distribution is expected to contain all possible scenarios and thus has a broader distribution than that of the forecast. The forecast distribution usually broadens as the forecast time increases because of the forecast error growth. Finally, the forecast will lose its predictability when the difference between the two distributions vanishes, which means the forecast is no better than a randomly drawn state from the climatology. The ideal climatological PDF should be obtained from long-term historical observations. In the absence of the observational data, climatological PDF can be constructed from long-term model simulations by invoking the perfect model assumption, in which the observed state can be captured by the model simulations, see Tang et al. (2008) for an example. The 20 years AVISO sea surface height observations provide a way to characterize the climatological PDF. We face several challenges when quantifying the SSH predictability in the Gulf of Mexico using information theory. One is the definition of climatology. Previous investigation shows that our HYCOM forecast PDF was broader than the AVISO climatological PDF in some grid cell, which violates the definition of relative entropy predictability measure. As pointed out by Kleeman (2002), the definition of the predictability formulated in Equation 4.1 is a perfect model measure, which takes no account of the accuracy of the model prediction. Another challenge is the calculation of the high dimensional entropy metrics. When we treat the PDF of the field quantity as a joint distribution, the dimensionality of the problem will make it computationally intractable.

Nevertheless, we show the area averaged SSH PDFs of the model simulations and the climatological data in 2 degree boxes. Figure 5.8, Figure 5.9 and Figure 5.10 present the PDFs of the model simulations and the climatological data at day 10,
day 20 and day 30, respectively. The AVISO climatological PDFs are shown in red and the model PDFs are shown in blue. The black asterisk represents the AVISO observation on that day. A basic observation is that the model PDFs broaden as time evolves and the difference between the model PDFs the climatological PDFs vanishes in grid boxes near the LC region at day 30. The limitation of this result is that the averaged SSH PDFs may be different compared with the original SSH PDFs.

**Figure 5.8:** Comparison of area-averaged (2 degree boxes) SSH PDFs at day 10. The AVISO climatological SSH PDFs is in red and the HYCOM SSH PDFs is in blue.
Figure 5.9: Comparison of area-averaged (2 degree boxes) SSH PDFs at day 20. The AVISO climatological SSH PDFs is in red and the HYCOM SSH PDFs is in blue.

Figure 5.10: Comparison of area-averaged (2 degree boxes) SSH PDFs at day 30. The AVISO climatological SSH PDFs is in red and the HYCOM SSH PDFs is in blue.
5.2 Updating the initial guess with observations

When observational data is available, we can correct the input uncertainties by applying inference or optimization algorithms.

5.2.1 AVISO along track observations

AVISO gridded data provides a great way to investigate the dynamical processes and perform model-data comparison in the entire Gulf of Mexico, however, we may need to account for the interpolation error for optimization and inference purposes. Instead, we use the AVISO along track Absolute Dynamic Topography (ADT) product from satellite mission Envisat, Jason-2 and Jason-1 on its new orbit from May 1, 2010 to May 30, 2010 as SSH observations. We preprocessed the SSH observations by removing the respective instantaneous spatial mean over the Gulf of Mexico in order to discard the signal induced by the volume changes in the Gulf of Mexico (Le Hénaff et al., 2012). The quality of the SSH observations deteriorate in the coastal region. In order to avoid this issue and focus only on the LC region, we further restrict the SSH data points used in this study. Figure 5.11 shows the locations of the SSH observations color coded based on different satellite missions. There is a total of 2249 satellite data points in this region during the one month period.

Figure 5.12 shows the preprocessed SSH observations for three different time intervals. We can clearly observe the development of a LC frontal eddy indicated by the negative SSH located in the west side of the LC eddy between day 10-20 in the west of the LC eddy. A LC eddy separation can be observed between day 20-30.
**Figure 5.11:** AVISO along track data used in the experiment starting from May 1 to May 30, 2010. Three satellite missions are used here: Envisat (red), Jason-2 (blue) and Jason-1 on its new orbit (black).

**Figure 5.12:** Preprocessed along track SSH for different time segments.
5.2.2 Bayesian optimization approach

The satellite observations allow us to update our knowledge of the uncertain inputs. Bayesian Optimization aims to minimize a scalar objective function. Here, we select the objective function to be the root mean square error between the along track satellite observations and the model simulations over the entire 30-day period. The objective function can be expressed as follows:

\[
g(\xi_1, \xi_2) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (H_i - M_i(\xi_1, \xi_2))^2}
\]

(5.2)

where HYCOM simulations \( M_i \) are chosen as the closest spatial point compared with the associated observations \( H_i \). \( N \) is the total number of observations and \( \xi \) represents the uncertain inputs.

Bayesian optimization is well suited here since the objective function shown in Equation 5.2 is computationally expensive since we need to run the forward model to estimate the objective function. It should be noted that we take advantage of the existing PC surrogate and sample the model simulation directly from the surrogate. Although the PC surrogate is used here as the forward model instead of the original HYCOM, Bayesian optimization, however, can be performed using direct model simulations. Bayesian optimization algorithm locates the minimum of the objective function adaptively by identifying the most likely locations of the minimum while simultaneously exploring large portions of the parameter space.

The key step of Bayesian optimization is to define an acquisition function which provides the optimization algorithm the next evaluation point. We adopt the Lower Confidence Bound (LCB) acquisition function here. Bayesian optimization maintains a Gaussian Process (GP) model at each optimization step. We use the constant mean and squared exponential covariance function in the GP model. A total 15
optimization steps are used in this experiment and the optimization cycle is repeated 10 times to avoid bad initialization.

Figure 5.13 shows the objective function modeled by the GP mean at the end of the optimization cycle. The blue dots represent past evaluation points and the black dot is the next evaluation point determined by the acquisition function. The objective function shows a quadratic shape in the $\xi_1$ direction and near linear shape in the $\xi_2$ direction. The optimization result provides the best uncertain input perturbations at $(\xi_1 = -0.08, \xi_2 = -0.99)$ which corresponds to the most negative perturbation in the 2nd initial condition EOF mode and very small negative perturbation in the 1st initial condition EOF mode. Figure 5.14 shows the associated acquisition function at the end of the optimization cycle. Figure 5.15 shows the minimum observed objective function versus the number of optimization steps (we add one more evaluation at each step) for the entire 10 optimization cycles (shown in color) as well as their mean (shown in black). After 10 optimization steps, the algorithm finds a reasonable global minimum in all optimization cycles and the results converge to a minimum root mean square error of about 0.143 m.
Figure 5.13: Modeled objective function with Gaussian Process at the end of the optimization cycle. All previous evaluation points are marked with blue dots. The next proposed evaluation point is shown in black. The surface is the mean function of the Gaussian Process model.
Figure 5.14: The acquisition function at the end of the optimization cycle. The black dot represents the next proposed evaluation point.

Figure 5.15: The minimum observed objective function (RMS error) values versus the number of function evaluations. 10 random restarts are performed in the experiment. Colored curves represent each individual optimization cycle while the thick black curve is the mean of all optimization cycles.
5.2.3 Bayesian inference approach

In order to obtain the complete posterior PDFs of the inferred parameters, the Bayesian inference technique is applied in this experiment. Now, we can substitute all the parameters in the experiment into the posterior distribution of the inferring parameters and rewrite Equation 2.29 as follows:

\[
p(\{\xi_1, \xi_2\}, \sigma^2 | \{H_i\}_{i=1}^N) \propto \left\{ \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(H_i - M_i)^2}{2\sigma^2}\right]\right\} p(\sigma^2) p(\xi_1) p(\xi_2) \tag{5.3}
\]

where \(\xi_1\) and \(\xi_2\) are uncertain inputs and their prior distributions are uniform; \(\sigma^2\) is the hyperparameter and its prior distribution is selected as the Jeffrey prior shown in Equation 2.31. Inferring the posterior distribution of the parameters requires sampling the posterior distribution presented in Equation 5.3. An adaptive Metropolis algorithm is used to accurately and efficiently sample this posterior distribution (50,000 iterations are performed). The PC approach is used here to accelerate the MCMC algorithm.

Figure 5.16 shows the MCMC chains for the three inferred parameters (\(\xi_1\), \(\xi_2\) and \(\sigma^2\)). The \(\xi_1\) chain appears to be concentrated near approximately -0.05, the \(\xi_2\) chain is concentrated in the lower end of the parameter range, which is approximately -1 and the chain for \(\sigma^2\) spans between 0.019 to 0.022 which is consistent with the empirical estimate using the existing 49-member HYCOM simulations.

The MCMC chains are then used to determine the posterior distributions. A kernel density estimation approach (Botev et al., 2010) is used here (the first 10,000 iterations associated with the burn-in period are discarded). Figure 5.17 shows the joint distribution between \(\xi_1\) and \(\xi_2\). There is a well-defined peak in this joint distribution as expected from the MCMC chain shown in Figure 5.16. The maximum a posteriori
(MAP) estimation of this joint distribution is \( \xi_1 = -0.05914 \) and \( \xi_2 = -0.99892 \), which is consistent with the result from the Bayesian optimization approach.

Figure 5.18 shows the root mean square error between the model simulations and the along track satellite observations for each of the HYCOM ensemble member during the 30 days simulation period. The minimum (HYCOM ensemble member 4 in Figure 5.18) can be clearly identified in the same region as the result from the Bayesian optimization (Figure 5.13) and the MAP estimation (Figure 5.17) from the posterior distribution. The fundamental difference between the Bayesian optimization and Bayesian inference is that the former technique provides only the global optimal values for the model uncertain inputs while the latter approach delivers the entire posterior distribution for the model uncertain inputs. The whole posterior distribution clearly provides some benefits since the distribution contains much more information compared with global optimal values (e.g. It makes much more sense to show distribution instead of the optimal values when the underlying distribution is near uniformly distributed). However, the global optimal values estimated using Bayesian optimization provides reasonable results when only limited number of model simulations are allowed.
Figure 5.16: MCMC chain samples for $\xi_1$ (left), $\xi_2$ (middle) and $\sigma^2$ (right).
Joint posterior PDF

\[ \text{MAP} = (\xi_1 = -0.05914, \xi_2 = -0.99892) \]

Figure 5.17: Joint posterior distribution of $\xi_1$ and $\xi_2$.

Figure 5.18: Root mean square error between model simulations and along track satellite observations with different initial condition perturbation $\xi_1$ and $\xi_2$. 
Next, we make a connection between the best perturbation result and the associated dynamical processes. Specifically, we explore the Loop Current evolution by sampling along the $\xi_1$ direction while keep $\xi_2 = -1$ and compare the SSH field between specific model simulations (simulation 1,4,7 shown in Figure 5.18) and observations. Figure 5.19 shows the SSH snapshots at different times for both model simulations (left three columns) and AVISO observations (right column). The ensemble member with $\xi_1 = 0$ (second from left) is the one with the minimum RMS error; the eddy shedding process of this ensemble member is proceeding “slower” than observations. On the other hand, the realization with positive perturbation in the $\xi_1$ direction (strengthening of the frontal eddies) shows a better agreement with observation on the Loop Current separation. At the same time, this positive perturbation also leads to the development of a frontal eddy on the north side of the Loop Current eddy which does not seem to be present in observations.

Figure 5.19: SSH snapshots at day 10 (top), day 20 (middle) and day 30 (bottom). The AVISO observations are located on the right, and the model simulations are located on the left. The left three columns represent simulations with most negative perturbation on $\xi_1$ (left), zero perturbation on $\xi_1$ (middle) and most positive perturbation on $\xi_1$ (right).
Chapter 6

Summary and conclusions

The main contributions of this dissertation are the implementation of an efficient uncertainty quantification (UQ) methodology for a regional ocean forecast and the application of statistical inference which can perform more efficiently using this implementation. The UQ methodology is based on PC expansion, which is used to construct a surrogate of the original model for efficient model evaluations. The implementation is tailored to limit the computational cost of an ensemble forecasting system, which provides the potential to apply it to real-world data assimilation system.

This UQ approach provides us an unique opportunity to represent the uncertainties in the model via their PDFs. The success of this approach hinges on capturing the dominant features of the uncertainty with the relatively modest number of uncertain input parameters so that the ensemble size remains small. In this dissertation, we only rely on EOF analysis to identify the perturbation and limit the dimensionality of the problem. For a mesoscale forecast in the central Gulf of Mexico, affected mainly by the Loop Current and Loop Current eddies, a modest ensemble size of about 50 can capture the variability in the Loop Current dynamics. The UQ implementation here also provides us the opportunity to perform inference without additional assumption on the Gaussianity of the statistics. Traditional, direct Bayesian Inference using
MCMC algorithm requires a large number of forward model samples, which is impractical for computationally expensive models. With the efficient surrogate, we can efficiently sample the posterior distribution of the uncertain inputs given observations using MCMC algorithm and thus provides us a new way to solve inverse problem for computationally expensive ocean models. We summarize the details of the relevant results in this dissertation as follows:

We quantified the uncertainties in the circulation forecast of the Gulf of Mexico caused by uncertain initial conditions and wind forcing. We have analyzed two ensembles of the HYCOM Gulf of Mexico simulations starting from May 1, 2010 to May 30, 2010 during the DWH oil spill. These two ensembles differ in the sources and the variability included in their input uncertainties. Specifically, we compare the model simulations against observations at the surface and at depth to address whether the ensemble perturbations represent the realistic uncertainties. We quantify the information trade-offs between the size of the ensemble and the model output variability contained in the ensemble. The predictability of the ensemble prediction is quantified using information theory. We update the knowledge of the input uncertainties using satellite along track observations.

The results of the surface and subsurface data-model comparison suggest that the uncertainties generated by the EOF modes are “realistic” since the observational data falls within the envelop of the ensemble simulations. A PC surrogate for the 798-member HYCOM ensemble was built to investigate the information trade-offs between the number of model simulations and the amount of variability included in the model output. The results show that the model output variability loss by omitting variability of the input uncertainties (high order EOF modes) is small, which suggests that adding additional input uncertainty sources may be more useful than adding high order EOF modes. The model out variability loss by omitting sources of the input
uncertainties is consistent with the global sensitivity analysis. Including only two initial condition EOF modes provides reasonable results for studying the dynamical processes in the LC region.

The contour boxplot method provides a way to rank functions and contours, in which descriptive statistics can be derived from the ranking result, e.g. the most representative ensemble member and the outliers. The result is consistent with the dynamical processes: weakening of the LC frontal eddies delays the LC eddy shedding event. The detailed investigation of the SSH PDF shows a strong non-Gaussian signal in the LC region, which indicates a bifurcation behavior in this region. The comparisons between PC covariance and empirical covariance at the surface and along a vertical section suggest the difference between these two approaches is small. The SSH predictability study suggest that the HYCOM ensemble simulations lose its predictability after 20 days. However, further investigations are needed in order to quantify the predictability using the forecast PDF.

When observational data is available, we apply two approaches to update our knowledge of the uncertain inputs. When the number of model realizations is limited, we find a reasonable well estimate of the optimal uncertain input parameters using less than 15 model simulations using Bayesian optimization. When the complete posterior distribution of the uncertain inputs is desired, we infer the posterior distribution using the Bayesian inference framework. Specifically, we leverage the PC surrogate to accelerate the MCMC algorithm using 49 model simulations. The posterior distribution is highly concentrated and the result is consistent with that from the Bayesian optimization approach.

The application of uncertainty quantification techniques in ocean modeling is at its beginning stage. The most challenging part is to reduce the dimensionality of the problem since the nature of this new uncertainty quantification approach requires us
to build a functional relationship between the uncertain inputs and model output. The implications of our results will lead to many interesting future works. The efficiency of this uncertainty quantification implementation can be applied for model calibration to guide the selection of model parameters. The availability of the full output PDF will open the door for data assimilation and predictability study in a non-Gaussian paradigm.
Appendix A

Cross-Validation

Algorithm 1 Cross validation error estimation

\[
x \leftarrow \text{ensemble input}; \ F \leftarrow \text{QoI associated with } x;
N_p \leftarrow 10; \ Nr \leftarrow 10; \ Pr \leftarrow 0.8;
\text{dr}_{\text{test}} = [\ \text{linspace}(0.001, 0.01, 10), \ \text{linspace}(0.02, 0.08, 10) ];
\alpha_r, \alpha_v, \alpha (\alpha_r \cup \alpha_v = \alpha; \ \alpha_r \cap \alpha_v = \emptyset; \ \text{size}(\alpha_r) = P_r \times \text{size}(\alpha));
\text{for } i=1 \text{ to } N_p \text{ do}
\quad \text{randomly permute the set index } \alpha_r;
\quad \alpha_v = \alpha \setminus \alpha_r;
\text{end for}
\text{for } i=1 \text{ to } N_p \text{ do}
\quad \text{for } i=1 \text{ to } \text{length}(\text{dr}_{\text{test}}) \text{ do}
\quad \quad \text{Building } F_r \text{ using BPDN with different permuted } x_{\alpha_r};
\quad \quad \text{L2 norm}(F(x_{\alpha_v}) - F_r(x_{\alpha_v}));
\quad \text{end for}
\quad \text{select the } \sigma \text{ corresponding to the minimum L2 norm;}
\text{end for}
\]

Symbols:

\( N_p \) — number of random permutation;
\( Nr \) — number of refinement length;
\( Pr \) — ratio of the reconstruction set and the total set;
\( \text{dr}_{\text{test}} \) — array of \( \sigma \) ratio at the beginning;
\( \alpha_r \) — reconstruction set index;
\( \alpha_v \) — validation set index;
\( \alpha \) — total set index;
\( F_r \) — PC reconstruction of the QoI with BPDN;
Appendix B

Adaptive Metropolis Algorithm

Figure B.1 shows a 2d example of the Metropolis algorithm (Metropolis et al., 1953). The algorithm starts with an initial guess $\xi_{t=0}$. Then, a candidate $\xi'$ is sampled from a proposal distribution $\xi' \sim N(\xi_0, cov)$. $cov$ is the covariance which can be chosen a priori.

Figure B.1: 2D example of Metropolis algorithm.
The Metropolis algorithm accept or reject the proposed candidate based on the following criteria:

$$\xi_1 = \begin{cases} 
\xi' & \text{if } r > \alpha \\
\xi_0 & \text{otherwise}
\end{cases} \quad (B.1)$$

where $r = \frac{\pi(\xi')}{\pi(\xi_0)}$ and $\alpha \sim \mathcal{U}(0,1)$.

The basic idea of the adaptive Metropolis algorithm (Haario et al., 2001; Roberts and Rosenthal, 2009) is to update the proposal distribution using the knowledge we have so far. In other word, the covariance of the proposal distribution depends on the history of the Markov chain. Here, the proposal distribution $N(\xi_t|\xi_0, \ldots, \xi_{t-1})$ is a Gaussian distribution with mean at the current point $\xi_t$ and covariance $C_t(\xi_0, \ldots, \xi_{t-1})$.

**Algorithm 2** Adaptive Metropolis algorithm

1: At time $t - 1$, we have sampled the states $\xi_{t-1} = (\xi_0, \xi_1, \ldots, \xi_{t-1})$
2: Draw a candidate $\xi'$ from a Gaussian distribution centered on the current state: $\xi' \sim N(\xi_{t-1}, cov)$ where $cov$ is the adaptive covariance
3: Calculate the ratio

$$r = \frac{\pi(\xi')}{\pi(\xi_{t-1})}$$

4: Draw a random number $\alpha \sim \mathcal{U}(0,1)$
5: Accept or reject candidate according to:

$$\xi_t = \begin{cases} 
\xi' & \text{if } r > \alpha \\
\xi_{t-1} & \text{otherwise}
\end{cases}$$

6: Repeat
Appendix C

Sensitivity Analysis

Let $L_2(K^d)$ be the space of real-valued squared-integrable functions over the $d$-dimensional hypercube $K^d$. The function $u(\xi) \equiv u(\xi_i, \ldots, \xi_d)$ is defined in the $d$-dimensional hypercube cube $K^d$.

C.1  Sobol-Hoeffding decomposition:

Any $u \in L_2(K^d)$ has an unique hierarchical orthogonal decomposition of the form (Hoeffding, 1948):

$$u(\xi) = u(\xi_1, \ldots, \xi_k) = u_0 + \sum_{i=1}^{d} u_i(\xi_i) + \sum_{i=1}^{d} \sum_{j=i+1}^{d} u_{i,j}(\xi_i, \xi_j) + \sum_{i=1}^{d} \sum_{j=i+1}^{d} \sum_{k=j+1}^{d} u_{i,j,k}(\xi_i, \xi_j, \xi_k) + \cdots + u_{1,\ldots,d}(\xi_1, \ldots, \xi_d)$$  \hspace{1cm} (C.1)

The number of terms in Equation C.1 is equal to:

$$\sum_{j=0}^{d} \binom{d}{j} = 2^d$$
The ensemble notation of Equation C.1 can be expressed as:

\[ u(\boldsymbol{\xi}) = \sum_{i \subseteq \mathcal{D}} u_i(\boldsymbol{\xi}_i), \quad \text{where } \mathcal{D} = \{1, \ldots, d\}. \quad \text{(C.2)} \]

The functionals in the decomposition satisfy the following orthogonality relations:

\[ \int_K u_i(\boldsymbol{\xi}_i) d\xi_j = 0, \quad \forall i \subseteq \mathcal{D}, j \in i, \]
\[ \int_{K^d} u_i(\boldsymbol{\xi}_i) u_j(\boldsymbol{\xi}_j) d\boldsymbol{\xi} = \langle u_i, u_i \rangle = 0, \quad \forall i, j \subseteq \mathcal{D}, i \neq j. \quad \text{(C.3)} \]

Given that the decomposition in Equation C.1 is orthogonal, The variance of \( u \) can be decomposed into:

\[ \mathbb{V}[u] = \sum_{i \subseteq \mathcal{D}} \mathbb{V}[u_i] \quad \text{(C.4)} \]

### C.2 Sensitivity indices

Sensitivity indices can be defined as (\(?\)):

\[ S_i(u) = \frac{\mathbb{V}[u_i]}{\mathbb{V}[u]} \leq 1, \quad \sum_{i \subseteq \mathcal{D}} S_i(u) = 1. \quad \text{(C.5)} \]

The order of the sensitivity indices \( S_i \) is equal to the cardinality of set \( i \): \( |i| = \text{Card}(i) \).

Since there are \( 2^d - 1 \) contributions to the total variance, there will exist the same number of sensitivity indices as well. Thus, sensitivity analysis usually needs more “abstract” characterizations. Next, we define the first order and total order sensitivity indices.

The first order sensitivity indices \( S_{\{i\}} \) measures the fraction of the variance due to the parameter \( \xi_i \) only without any interaction with others. In this case \( \sum_{i=1}^d S_{\{i\}} = 1 \), the effects on the variability of all interactions between parameters can be ignored. In
other word, the impact of the parameters can be studied separately. The total order sensitivity indices $T_{\{i\}}$ characterizes the variability due to parameter $\xi_i$, together with its interactions with other parameters. $\sum_{i \in D} T_{\{i\}}$ can be larger than 1, and the excess part characterizes the impact of the interaction terms.

Fig C.1 shows a schematics of the relationship between sensitivity indices, first-order sensitivity indices and total sensitivity indices with $d = 3$.

The values of different sensitivity indices satisfy the following relations:

$$\sum S_i = 1, \quad sensitivity \ indices$$
$$\sum S_{\{i\}} \leq 1, \quad first \ order \ sensitivity \ indices$$
$$\sum T_{\{i\}} \geq 1, \quad total \ order \ sensitivity \ indices$$

\[ C.6 \]

![Figure C.1: Illustration of different sensitivity indices.](image)
C.3 Polynomial Chaos estimation of sensitivity indices

We can assess the Sobol sensitivity indices by manipulating the PC series representations without any additional model simulations (?).

Consider the truncated PCE:

\[ u(\xi) = \sum_{\alpha \in A} u_\alpha \Psi_\alpha(\xi) \quad (C.7) \]

The Sobol-Hoeffding decomposition in Equation C.2 can be represented with PC approximation by:

\[ u_i(\xi_i) = \sum_{\alpha \in A(i)} u_\alpha \Psi_\alpha(\xi_i) \quad (C.8) \]

where the multi-index set \( A(i) \) can be given by:

\[ A(i) = \{ \alpha \in A; \alpha_i > 0 \text{ for } i \in \mathbf{i}, \ k_i = 0 \text{ for } i \notin \mathbf{i} \} \quad (C.9) \]

The first-order sensitivity indices can be given by:

\[ S_{\{i\}} = \frac{\sum_{\alpha \in \tau_{\{i\}}^S} u_\alpha^2 \langle \Psi_\alpha^2 \rangle}{\sum_{\alpha \in A \setminus A_0} u_\alpha^2 \langle \Psi_\alpha^2 \rangle} \quad (C.10) \]

where \( \tau_{\{i\}}^S = \{ \alpha \in A; \alpha_i > 0, \alpha_j \neq i = 0 \} \)

The total sensitivity indices corresponding to the input \( \xi_i \) can be obtained from:

\[ T_{\{i\}} = \frac{\sum_{\alpha \in \tau_{\{i\}}^T} u_\alpha^2 \langle \Psi_\alpha^2 \rangle}{\sum_{\alpha \in A \setminus A_0} u_\alpha^2 \langle \Psi_\alpha^2 \rangle} \quad (C.11) \]
where $\tau^T_{\{i\}} = \{\alpha \in A; \alpha_i > 0\}$
References


