Reconstructing Uncertain Response Surfaces with Application to Lagrangian Ocean Data

Rafael Carvalho Goncalves

University of Miami, rafaelcgon@gmail.com

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

Recommended Citation
https://scholarlyrepository.miami.edu/oa_dissertations/2225

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
UNIVERSITY OF MIAMI

RECONSTRUCTING UNCERTAIN RESPONSE SURFACES WITH APPLICATION TO LAGRANGIAN OCEAN DATA

By
Rafael Carvalho Gonçalves

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
December 2018
RECONSTRUCTING UNCERTAIN RESPONSE SURFACES WITH APPLICATION TO LAGRANGIAN OCEAN DATA

Rafael Carvalho Gonçalves

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

Rick Lumpkin, Ph.D.
Oceanographer, NOAA/AOML, Miami, Florida

Guillermo Prado, Ph.D.
Dean of the Graduate School
Motivated by the problem of predicting the transport of material in the ocean, the present work went after two general problems. One is to use a probabilistic Lagrangian oil model to forecast the fate of an oil spill. The other is to use a dense array of Lagrangian observations to estimate the near surface submesoscale velocity field. Both problems were tackled with methodologies based on the reconstruction of response surfaces.

In the first part, an uncertainty quantification methodology based on a non-intrusive polynomial chaos approach is developed for the forecasting of oceanic oil spills. This allows the model’s output to be presented in a probabilistic framework so that its predictions reflect the uncertainty in its input data. The new capability is illustrated by simulating the far-field dispersal of oil in a Deepwater Horizon blowout scenario. The uncertain input consisted of ocean currents and oil droplet size data, and the main model output analyzed is the ensuing oil concentration in the Gulf of Mexico. A 1331 member ensemble was used to construct a surrogate for the model which was then mined for statistical information. The mean and standard deviations in the oil concentration were calculated for up to 30 days, and the total contribution of each input parameter to the models uncertainty was quantified at different depths. Also, probability density functions of oil concentration were constructed by sampling the surrogate and used to elaborate probabilistic hazard maps of oil impact. The sur-
rogate performance was constantly monitored in order to demarcate the space-time zones where its estimates are reliable.

In the second part, a dense array of drifters data is used to reconstruct the near-surface velocity field. The reconstruction is carried out by a procedure known as Gaussian process regression, that statistically interpolates the observations in both space and time. Because the spacing of the drifters evolves with the flow causing the resolution that they provide to vary in space and time, it is important to be able to characterize where and when the estimated velocity field is more or less accurate, which we do by providing fields of interpolation errors. The interpolation procedure uses a covariance function to characterize the flow correlations in space and time. One novelty in this approach is that it allows the data to determine the correlation scales along with the appropriate amplitude of observational noise at these scales. This part is divided in two segments, with two distinct analyses being conducted: in the first one, velocity estimates from 320 drifters are used to reconstruct the surface flow field inside a frontal cyclonic eddy during a 12 hours period. Results show the presence of strong convergence zones inside the cyclone with characteristics of submesoscale fronts, displaying strong relative vorticity and strain rate. In the second segment, a framework to estimate the flow field over a submesoscale front is presented. The framework is based on adjusting the coordinate system to the alignment of the drifters. An extensive testing considering 14 different covariance functions show that this strategy improves considerably the velocity reconstructions over a strong front when compared to a Cartesian coordinate system.
Dedicated to the memory of my beloved grandmother, vó Anita.
I would like to thank my advisor, Mohamed Iskandarani, for giving me the opportunity to work with him, and for opening up the doors of the uncertainty quantification world to me. Mo has all the qualities anyone could ask for in an advisor; he knows the right moment to be supportive and patient, and when he needs to push the student in the right direction. And he is always available to help, not only with research matters, but with any problem in general. I am truly grateful for having Mo as my friend.

My sincere thanks to my committee members, Matthieu Le Hénaff, Rick Lumpkin, Tamay Özgökmen and Arthur Mariano for their support, attention, and time taken to help me with my research.

I would like to thank Carlisle Thacker for the research collaborations, discussions and friendship. Carlisle has being my co-mentor during my Ph.D., and his wisdom, knowledge and passion for science have truly inspired me. I also would like to thank Shitao Wang and Ashwanth Srinivasan for the collaboration and research discussions.

I gratefully acknowledge financial support by the Brazilian Ministry of Science, Technology and Innovation (CNPq- Council for Scientific and Technological Development) through a Ph.D. scholarship from the Science Without Borders program, and by grants from the Gulf of Mexico Research Initiative to the Deep-C and CARTHE Consortia.
I would like to thank my wife, Lilian, for the unconditional love and support that she gave me over the last 6 years. She moved to Miami after she got her Ph.D., and she put her carrier on hold to be by my side. That’s way more than I could ever ask, and I’ll always be grateful. I would like to thank my parents, Paulo and Regina, who have always supported every step that I took since before I even knew how to walk. Besides the distance, they are always with me.

At last, I would like to thank all my friends in Miami, including the MPO class of 2012, the Brazilian Mafia, the soccer group and from the wetlab. I really enjoyed living in Miami the past 6 years, and that’s all because of them.
# Contents

List of Figures ix

List of Tables xix

1 Introduction 1

1.1 Motivation ................................................. 1

1.2 Reconstructing uncertain response surfaces .................. 4

2 Methodology 7

2.1 The Polynomial Chaos approach ............................ 7

2.2 Gaussian Process Regression ............................... 9

2.2.1 Covariance functions ................................... 12

2.2.2 Optimization procedure ................................. 14

3 A framework to quantify uncertainty in simulations of oil transport in the ocean 16

3.1 Overview .................................................. 16

3.2 Model description .......................................... 19

3.2.1 Lagrangian oil fate model .............................. 19

3.2.2 Hydrodynamic model .................................... 21

3.2.3 Model set up ............................................. 23

3.3 Input parameters uncertainty ............................... 24
List of Figures

3.1 Rosin-Rammler distributions considering $d_{50} = 200 \mu m$ (blue) and $d_{50} = 900 \mu m$ (red). For both distributions, the spreading parameter is $\alpha = 1.8$ .......................................................... 26

3.2 Percentage of the total variance explained by the 1st and 2nd EOFs combined for the zonal (U, left panels) and meridional (V, right panels) velocity components, in three selected depths. The black dots depict the position of the Deepwater Horizon wellhead. ............................................. 29

3.3 Surface velocity vectors of the unperturbed velocity field (left) and of the perturbations constructed from the first (center) and second (right) EOFs. The magenta dots depict the position of the Deepwater Horizon wellhead. ................................................................. 30

3.4 Surface velocity vectors at 19-May-2010 for the unperturbed velocity field $u_a = u_0$ (upper left) and for three perturbed fields $u_a = u_0 + u_1 + u_2$ (upper right), $u_a = u_0 + u_1$ (lower left), and $u_a = u_0 + u_2$ (lower right). The magenta dots depict the position of the Deepwater Horizon wellhead. ................................................................. 31

3.5 Principal components 1 and 2 of the velocity field from HYCOM, starting at 20-April-2010. ................................................................. 31
3.6 Vertical component of relative vorticity normalized by planetary vorticity from surface velocity fields on 19-May-2010. Upper left: $\mathbf{u}_a = \mathbf{u}_0$ (unperturbed field, i.e. $\xi_1 = \xi_2 = 0$). Upper right: $\mathbf{u}_a = \mathbf{u}_0 + \mathbf{u}_1 + \mathbf{u}_2$ ($\xi_1 = \xi_2 = 1$). Lower left: $\mathbf{u}_a = \mathbf{u}_0 + \mathbf{u}_1$ ($\xi_1 = 1$ and $\xi_2 = 0$). Lower right: $\mathbf{u}_a = \mathbf{u}_0 + \mathbf{u}_2$ ($\xi_1 = 0$ and $\xi_2 = 1$).

3.7 Lateral strain rate normalized by planetary vorticity from surface velocity fields on 19-May-2010. Upper left: $\mathbf{u}_a = \mathbf{u}_0$ (unperturbed field, i.e. $\xi_1 = \xi_2 = 0$). Upper right: $\mathbf{u}_a = \mathbf{u}_0 + \mathbf{u}_1 + \mathbf{u}_2$ ($\xi_1 = \xi_2 = 1$). Lower left: $\mathbf{u}_a = \mathbf{u}_0 + \mathbf{u}_1$ ($\xi_1 = 1$ and $\xi_2 = 0$). Lower right: $\mathbf{u}_a = \mathbf{u}_0 + \mathbf{u}_2$ ($\xi_1 = 0$ and $\xi_2 = 1$).

3.8 Left: RMSE of the oil concentration in the upper layer after 30 days. The RMSE was computed through a set of 1024 model realizations using randomly chosen sets of the uncertain input parameters and their respective surrogate. The values of oil concentration (in $10^4 kg/m^3$) from the model realizations and their surrogates in points A and B are presented in the scatter plots (center and right, respectively). The red lines are the linear regression of the scatter plots, and Corr. and Sl. are the correlation coefficients and slopes, respectively. The magenta dot indicates the DWH wellhead location.
3.9 Mean oil concentration between 0 – 5 meters depth (in $kg/m^3$, left), and mean over standard deviation ratio (right), on days 10 (29 April 2010), 20 (9 May 2010), and 30 (19 May 2010) of the oil spill. Both mean and standard deviation were extracted from the PC coefficients. The contours on the plots of 19 May 2010 indicate the position of the oil spill on 17 May 2010 (cyan) and 20 May 2010 (blue) as observed from a composite of satellite images. The black contours enclose cells for which the surrogate reproduces the model outputs with a $RMSE \leq 0.1$, and the black dot indicates the DWH wellhead location.

3.10 Mean oil concentration between 300 – 600 meters depth (in $kg/m^3$, left), and mean over standard deviation ratio (right), on days 10 (29 April 2010), 20 (9 May 2010), and 30 (19 May 2010) of the oil spill. The black contours enclose cells for which the surrogate reproduces the model outputs with a $RMSE \leq 0.1$, and the black dot indicates the DWH wellhead location.

3.11 Mean oil concentration between 900 – 1200 meters depth (in $kg/m^3$, left), and mean over standard deviation ratio (right), on days 10 (29 April 2010), 20 (9 May 2010), and 30 (19 May 2010) of the oil spill. The black contours enclose cells for which the surrogate reproduces the model outputs with a $RMSE \leq 0.1$, and the black dot indicates the DWH wellhead location.
3.12 Global sensitivity index at the upper layer for the 1st and 2nd EOF modes, and the 50th percentile of the droplet size distribution \((d_{50})\), on days 10 (29 April 2010), 20 (09 May 2010) and 30 (19 May 2010) of the oil spill. The blue contour encloses cells for which the surrogate reproduces the model outputs with a \(RMSE \leq 0.1\). The green line on the 19 May 2010 plots indicates the grid cells used to compute the total amount of oil that reached the northwestern coast of Florida during the first 30 days of the oil spill (Figure 3.15).

3.13 Same as figure 3.12 but for the mid-depth layer.

3.14 Global sensitivity index for the deep layer for the three input parameters on day 30.

3.15 Amount of oil that reached the coast of Florida (green line on figure 3.12) after 30 days considering different input parameters. \(\xi_1\), \(\xi_2\) and \(\xi_3\) are the standardized random variables used to propagate uncertainty through the 1st and 2nd EOFs, and the 50th percentile of the droplet size distribution, respectively. The red dots are the outputs of simulations considering \(\xi_1 \in [-1,1], \xi_2 = 0, \xi_3 = 0\), the black dots are from simulations with \(\xi_2 \in [-1,1], \xi_1 = 0, \xi_3 = 0\), and the blue dots are from simulations with \(\xi_3 \in [-1,1], \xi_1 = 0, \xi_2 = 0\). The big dots are outputs from actual realizations of the model, while the smaller ones were obtained from the PC surrogate.
3.16 10th (left panels) and 90th (right panels) percentiles of the PDF of oil concentration \((kg/m^3)\) between 0 and 5 meters depth for days 15 (04 May 2010), 20 (09 May 2010) and 25 (14 May 2010) of the oil spill, considering \(10^5\) samples of the PC proxy. The blue contour encloses cells for which the surrogate reproduces the model outputs with a \(RMSE \leq 0.1\), and the black dot indicates the DWH wellhead location.

3.17 Probability of oil concentration (in %) to reach at least \(10^5kg/m^3\) between 0 and 5 meters depth on days 15 (04 May 2010), 20 (09 May 2010) and 25 (14 May 2010) of the oil spill. The blue contour encloses cells for which the surrogate reproduces the model outputs with a \(RMSE \leq 0.1\), and the black dot indicates the DWH wellhead location.

4.1 (A) and (B): Sea surface temperature and chlorophyll a concentration from MODIS/AQUA on 07 February 2016 showing a hammerhead feature where the LASER drifters were released. (C) Aircraft survey of sea surface temperature with the initial drifter positions (white dots). The highlighted area in (A) and (B) indicates the boundaries of (C).

4.2 Two dimensional histogram of drifter pair separation and time throughout the first 48 hours after the last drifter deployment. The size of the bins are 50 meters in the pair separations axis and 30 minutes in the time axis. Only bins with more than 50 samples are shown. The solid black line indicates the mode of the pair separations distribution in each 30 minutes bin. The vertical dashed line indicates the end of the 12 hours period used for the velocity reconstruction.

4.3 Position and displacement velocity of the drifters 4 (a) and 12 (b) hours after the last drifter deployment.
4.4 Comparison of correlations estimated from the observations and the optimized SEs used in the covariance function. The SEs were normalized by their respective total variance ($\sigma_1 + \sigma_2 + \sigma_N$, see table 4.1). Right (left) panels contain comparisons for the u (v) component. Upper panels: simultaneous correlation as a function of drifters separations computed using spatial bins of 200 meters (blue), and SE for the zonal (solid red) and meridional (dotted red) directions. Lower panels: Lagrangian and Eulerian lagged auto-correlation (solid and dotted blue, respectively), and SE for the time dimension (red).

4.5 Horizontal divergence normalized by planetary vorticity and velocity vectors of reconstructed velocity fields in a 50 meters resolution grid (velocity vectors sampled every 15 grid points). (a) and (b): Reconstruction using covariance function with only one scale per dimension for each component, with $\sigma = \sigma_1$, $r_t = r_{t,1}$, $r_y = r_{y,1}$ and $r_x = r_{x,1}$ in (a), $\sigma = \sigma_2$, $r_t = r_{t,2}$, $r_y = r_{y,2}$ and $r_x = r_{x,2}$ in (b) (see table 4.1 for the values of each hyper-parameter). In (c), the covariance function was defined as the sum of the two covariance functions used in (a) and (b).

4.6 (a): Cumulative histograms of absolute error of $v$ and $u$ reconstructions at verification points. (b): Cumulative histogram of the absolute error normalized by the absolute observed value. (c) and (d): Difference between error estimates from the posterior covariance ($\text{Err}_Q$) and the actual errors computed from the observations for $v$ and $u$, respectively. A total of 8171 verification points were used in these histograms.
4.7 Snapshots of the reconstructed velocity field (black vectors) sampled every 15 grid points superimposed over speed ($ms^{-1}$). The time of each snapshot is referenced on 7 February 2016 at 14:15 UTC. The magenta vectors are the drifters velocities at the time of each snapshot.

4.8 Horizontal divergence of the reconstructed velocity field normalized by planetary vorticity. The time of each snapshot is referenced on 7 February 2016 at 14:15 UTC. The green vectors are the drifters velocities at the time of each snapshot.

4.9 Vertical relative vorticity of the reconstructed velocity field normalized by planetary vorticity. The time of each snapshot is referenced on 7 February 2016 at 14:15 UTC. The green vectors are the drifters velocities at the time of each snapshot.

4.10 Lateral strain rate of the reconstructed velocity field normalized by planetary vorticity. The time of each snapshot is referenced on 7 February 2016 at 14:15 UTC. The magenta vectors are the drifters velocities at the time of each snapshot.

4.11 Normalized histograms of relative vorticity (a), horizontal divergence (b) and lateral strain rate (c), all quantities normalized by f. The histograms were built from velocity reconstructions taken in 15 minutes intervals from 7 February 2016, 14:15 UTC to 8 February 2016, 2:15 UTC. Only grid points with $Err_Q(u)$ and $Err_Q(v)$ smaller than 0.04 m/s were considered.
5.1 Positions of the drifters used in the analysis of chapter 5. (A) shows the original positions, (B) and (C) show the positions with the center of mass removed at each time step and (D) shows the positions of the red dots in (C) in the new coordinate system. In (A) and (B), the blue/red dots are the initial/final positions. In (C), the red dots are the positions of the observations used in the analysis. The green line is a 2nd order polynomial fit of the red dots, which was defined as the along front direction (y') in the new coordinate system. 

5.2 Three different settings of training/test data in the original Cartesian coordinate system.

5.3 Medians of the absolute errors of velocity for the reconstructions using the training/test data setting presented in figure 5.2. The covariance function labels are: SE = squared exponential, M32 = Matérn 3/2, M52 = Matérn 5/2 and RQ = rational quadratic.

5.4 Box plots of the absolute errors of velocity for the reconstructions using the training/test data setting presented in figure 5.2A. The horizontal black line inside each box is the median, and the whiskers extend to 1.5 of the inter-quartile range. The outliers were omitted for better visualization.

5.5 Same as figure 5.4 for the training/test data setting presented in figure 5.2B.

5.6 Same as figure 5.4 for the training/test data setting presented in figure 5.2C.
5.7 Observations used as training (red vectors) and test (black vectors) data, and velocity reconstructions at the test data locations using $C_1$ (blue) and $C_2$ (green) with the rational quadratic covariance function. The training/test setting is the same presented in figure 5.2C. The magenta line indicates the location of the curvilinear along front axis. The velocity snapshots were taken at the first (A) and last (B) time steps of the three hours period consider in the analysis.

5.8 Selected optimized covariance functions with $C_{1b}$ normalized by their total variance. A, B and C show the cross-front, along front, and time dimension functions, respectively.

5.9 Horizontal divergence normalized by planetary vorticity at 2 h and 45 min of the 3 hours period, from regressions using the rational quadratic covariance function with $C_1$ (A), $C_{1b}$ (B), $C_2$ (C) and $C_3$ (D). Black vectors are the reconstructed velocities sampled every 40 grid points. Green vectors are the velocity observations used in the regression. The black line delineates the area inside which the estimated error from the posterior covariance is smaller than 0.05 m/s.

5.10 Horizontal divergence normalized by planetary vorticity at 2 h and 45 min of the 3 hours period, from reconstructions with $C_{1b}$ using as covariance function: SE (A), RQ (B), SE+SE (C) and M32 + M32 (D). Black vectors are the reconstructed velocities sampled every 40 grid points. Green vectors are the velocity observations used in the regression. The black line delineates the area inside which the estimated error from the posterior covariance is smaller than 0.05 m/s.
5.11 Snapshots of horizontal divergence normalized by planetary vorticity from the velocity reconstruction using $C_{1b}$ with the rational quadratic covariance function. Black vectors are the reconstructed velocities sampled every 40 grid points and yellow vectors are the velocity observations used in the regression. The black line delineates the area inside which the estimated error from the posterior covariance is smaller than 0.05 m/s.

5.12 Same as fig. 5.11 for relative vorticity normalized by planetary vorticity.

5.13 Same as fig. 5.11 for lateral strain rate normalized by planetary vorticity.
List of Tables

4.1 Optimized hyper-parameters of the covariance function. . . . . . . . . . 64

5.1 Optimized hyper-parameters of selected covariance functions with $C_1b$.

All covariance functions presented an observation noise of 0.007 m/s. 89
Chapter 1

Introduction

1.1 Motivation

Predicting the pathways of material transport in the ocean is important for several applications, like the study of the distribution of planctonic species, search and rescue operations, and mitigation actions for pollutant spills. Simulations of material transport are usually dealt with in a Lagrangian framework, i.e. a quantity of interest is treated as a collection of particles whose displacements are computed through an Eulerian representation of the ocean velocity field. The main difficulty of such predictions arises exactly from their Lagrangian nature, as even regular flows with simple time dependence can induce chaotic behavior in particles trajectories (Aref, 1984). Uncertainty in the initial position of the particles impose severe limitations in the Lagrangian predictability, specially in areas with large velocities and high velocity gradients (Huntley et al., 2011). Furthermore, simulated trajectories accumulate even the smallest errors in the Eulerian velocity, leading them to quickly diverge from reality (Griffa et al., 2004; Huntley et al., 2011). In most applications, the velocity field is obtained from ocean general circulation models (OGCMs), and uncertainties
arise from truncation errors associated with the model numerics and grid choices, subgrid-scale parameterizations and the specification of initial and boundary conditions.

Particle dispersion at scales not resolved by the OGCMs might impose further limitations to Lagrangian predictability when controlled by processes within the same scale range (local dynamics). The behavior of particle dispersion is dictated by the kinetic energy wavenumber spectrum, and is therefore related to the nature of the energy cascade in the ocean (Bennett, 1984; Bracco et al., 2004; Özgökmen et al., 2012). The ocean circulation is forced at large scales, and kinetic energy is transferred to the mesoscale range by instabilities of the large scale currents. The resultant mesoscale flow is dominated by geostrophically balanced eddies, which according to the geostrophic turbulence theory, should transfer energy towards larger scales (inverse energy cascade), and enstrophy towards smaller scales (forward enstrophy cascade) (Charney, 1971). The relative dispersion under geostrophic turbulence presents a cubic growth in time at scales within the wavenumber range of the inverse energy cascade, and is controlled by local dynamics (Bracco et al., 2004). This is the part of the spectrum represented in current operational OGCMs, which can resolve scales down to the borderland between mesoscale and submesoscale $O(10)$ km. At separation scales on the order of the enstrophy cascade range (smaller than 10 km), relative dispersion under geostrophic turbulence grows exponentially with time and is controlled by non-local dynamics (Bennett, 1984), more specifically by the mesoscale eddy field. In this case, the relative dispersion is not dependent on the subgrid scales from OGCMs, and the spatial resolution wouldn’t be an issue. However, recent studies have suggested a different picture for the kinetic energy wavenumber spectrum, showing that the strain field from mesoscale eddies induce the formation of sharp submesoscale fronts with significant ageostrophic circulation (Capet et al.,
2008; McWilliams et al., 2009; Molemaker et al., 2010). Frontal instabilities along these fronts can generate submesoscale eddies and smaller features, leading to a forward energy cascade and therefore breaking the geostrophic turbulence paradigm. Such forward energy cascade at the submesoscale range would result in a strong local influence of the relative dispersion at these scales, inducing a much faster dispersion than it would be expected under geostrophic turbulence dynamics. In accordance with that, Poje et al. (2014) presented evidence of locally controlled dispersion at the submesoscale range by analyzing the trajectories of hundreds of surface drifters released in the Northern Gulf of Mexico. In the light of these studies, submesoscale transport properties should be accounted for in material transport forecasts by either increasing the spatial resolution of operational OGCMs or by implementing adequate sub-grid parameterizations (Özgökmen et al., 2012; Haza et al., 2012).

The first part of this thesis addresses the quantification of the uncertainties in simulating the transport of oil spilled in the ocean, which is relevant for response efforts and for the elaboration of risk assessments. Forecasts of the transport of oil in the ocean by deterministic simulations are highly uncertain, not only because of the inherent uncertainty from the Eulerian velocity, but also because they rely on several parametrizations to represent oil buoyancy, oil degradation and unresolved scales of motion. The objective of this part is to establish an uncertainty quantification (UQ) framework for oil transport simulations which provides most likely scenarios along with confidence levels and error estimates. The second part of the thesis is devoted to apply observations to improve estimates of the Eulerian velocity field of the ocean surface. More specifically, dense array of velocity observations obtained from surface drifters trajectories, with high resolution in space and time, are applied to resolve the velocity field in the submesoscale range. The first goal of this part is to provide a reliable characterization of the Eulerian velocity field, which could be directly applied
to estimate material transport, to locally improve coarser Eulerian velocity estimates from altimetry observations or numerical models, or even be applied to improve the initial conditions of numerical forecasts. The second goal is to characterize the sub-mesoscale activity sampled by the drifters in terms of derived kinematic quantities and dominant length and time scales.

1.2 Reconstructing uncertain response surfaces

The two parts of this thesis are linked not only by the motivation of accurately estimating material transport in the ocean, but also by the methodology to approach each problem. Both require estimating a response surface from sampling points.

In the UQ part, we want to estimate probability density functions (PDFs) of model outputs from PDFs of model input parameters. A Monte Carlo estimation requires a large number of model realizations in order to sample the input parameter space, which might not be feasible for computationally costly numerical models. Our approach is based on constructing a surrogate for the model by establishing the dependence of an output of interest in the uncertain input parameters, using fewer model realizations than it would be required in a pure Monte Carlo approach. The surrogate is then sampled to build the output PDFs. In other words, with a few samples of input/output pairs we reconstruct the full response surface of a model output.

A variety of regression or interpolation methods can be used to build the surrogate (Iskandarani et al., 2016b). The technique used here is a spectral regression with orthogonal polynomial basis functions known as Polynomial Chaos (PC). Our approach to PC is non-intrusive, which requires running the model a few times to compute the PC coefficients by quadrature, with the input parameters set according to the quadrature rule. The PC approach has been used in oceanography for forward propagation of uncertainty in OGCM boundary and initial conditions (Thacker et al., 2012; Iskan-
It has also been applied in Bayesian inverse problems aimed to estimate uncertain input parameters from observations, with the PC surrogate used to build the posterior distribution. In Sraj et al. (2014), airborne expandable bathythermograph temperature data was used to infer the wind drag coefficient in a general ocean circulation model.

In the second part of the thesis, the goal is to reconstruct the Eulerian surface velocity field from sparse observations. We want the response surface of the horizontal velocity components as a function of space and time. For this problem we can not use the PC approach, as it requires observations at predetermined input values to solve the quadrature computations. The Gaussian Process regression (GPR) was chosen, as it does not restrict the input values. The GPR has a wide range of applications, from geostatistics to supervised machine learning (Rasmussen and Williams, 2006), and include, in particular, the problems of forward propagation of uncertainty in numerical models in a similar fashion as the above-mentioned PC approach (Thacker et al., 2015; Iskandarani et al., 2016a). The technique is mathematically equivalent to objective analysis, which was first applied in oceanography by Bretherton F. P. et al. (1976). The authors presented its capabilities for optimum interpolation and for the design of observational arrays. This latter application takes advantage of the analysis error provided by a posterior covariance function to optimize the observational effort (Barth and Wunsch, 1990), and it has been applied on the design of networks for expendable bathythermograph measurements (White and Bernstein, 1979; Sprintall and Meyers, 1991; Festa and Molinari, 1992; White, 1995), for the deployment of autonomous profiling floats from the Argo array (Roemmich et al., 2004), and on the optimization of trajectories from gliders (Leonard et al., 2007; Davis et al., 2009; Leonard et al., 2010). Applications of objective analysis for optimum interpolation include estimates of the geostrophic streamfunction from different sorts of observations,
as from velocity estimates of neutrally buoyant floats (McWilliams, 1976), from the combination of surface drifters and satellite altimetry data (Le Traon and Hernandez, 1992), from hydrographic observations and shipboard acoustic Doppler current profiles (Chereskin and Trunnell, 1996), and from satellite radiometer estimates of surface velocity (Wilkin et al., 2002). Other applications include estimates of the velocity field from high frequency radar observations (Kim et al., 2008), and of sea surface temperature and salinity from satellite data (Walker and Wilkin, 1998; Melnichenko et al., 2016). The difference between applications of GPR in the machine learning community and objective analysis is that in the first, the hyper-parameters of the covariance function (correlation scales and signal to noise ratio) are determined directly by an optimization procedure, in which their probability conditioned on the observations is maximized, while in the latter, they are estimated empirically from observations, or obtained from climatology.
Chapter 2

Methodology

2.1 The Polynomial Chaos approach

The application of the PC formulation to uncertainty quantification is based on constructing a surrogate for a model which represents the dependence of an output of interest on the uncertain data. This surrogate is subsequently used to explore the uncertain parameter space using a large ensemble to tally robust estimates of the output’s statistical distribution. A PC surrogate uses a spectral series of the form

\[ C(x, t, \xi) = \sum_{k=0}^{P} \hat{C}_k(x, t) \psi_k(\xi) \] (2.1)

where \( C(x, t, \xi) \) represents the output of interest\(^1\), \( \hat{C}_k(x, t) \) represents the PC coefficients, \( \psi_k(\xi) \) are the multi-dimensional orthogonal-polynomial basis functions representing the variations of \( C \) in the uncertain space, and \( P \) represents the truncation of the series to \((P + 1)\) terms. The notation \( C(x, t, \xi) \) emphasizes the dependence of the model output not only on space \( x \) and time \( t \), but also on the vector of uncertain data \( \xi \). The basis functions are chosen to be orthogonal with respect to an inner

\(^{1\text{In our application of PC presented on chapter 3, } C(x, t, \xi) \text{ is oil concentration.}}\)
product where the PDF of the uncertain variables $\xi$ appears as the weight function. The multi-dimensional basis functions are product of 1D basis functions. In chapter 3 we will consider 3 uncertain input parameters, so, in our 3D uncertain space, the basis-functions take the form:

$$\psi_k(\xi) = P_{m_1}(\xi_1)P_{m_2}(\xi_2)P_{m_3}(\xi_3)$$

(2.2)

where $P_{m_d}(\xi_d)$ represents the 1D basis function of degree $m_d$ in the variable $\xi_d$, $d = 1, 2, 3$. As the standardized random variables $\xi_d$ are uniformly distributed, $P_{m_d}(\xi_d)$ are Legendre polynomials of degree $m_d$. The polynomial degree in each dimension varies between $0 \leq m_d \leq M$, and we have used a total degree truncation for the multi-dimensional basis i.e. $m_1 + m_2 + m_3 \leq M$. The total number of basis functions is then $P + 1 = \frac{(M+3)!}{M!3!}$.

In chapter 3, we use a non-intrusive Galerkin projection to compute the series coefficients and the associated integrals are approximated with tensorized Gauss quadrature:

$$\hat{C}_k(x, t) = \frac{\sum_{i=1}^{Q} \psi_k(\xi_{i}^Q)C(x, t, \xi_{i}^Q)\Omega_{i}^Q}{\sum_{i=1}^{Q} \psi_k(\xi_{i}^Q)\psi_k(\xi_{i}^Q)\Omega_{i}^Q}$$

(2.3)

where $\Omega_{i}^Q$ and $\xi_{i}^Q$ are the multi-dimensional quadrature weights and nodes, respectively. The number of uni-dimensional quadrature points needed to maintain an exact evaluation of the denominator is $q = (M + 1)$ and hence the number of 3D quadrature points needed is $Q = (M + 1)^3$. The determination of the coefficients thus require an ensemble with $Q$ members and this constitutes the bulk of the computational cost.

In our application of PC, the maximum degree of the polynomial was set to $M = 10$, which required $Q = 1331$ realizations of model to compute the coefficients. Once
the coefficients are available, the series can be used instead of the model to compute any required statistical information. Some statistical information can be obtained directly from the coefficients, e.g. the expected value of the output of interest over the uncertain input space is simply given by \( \hat{C}_0(x, t) \) while the standard deviation is \( \sum_{k=1}^{P} \hat{C}_k^2 \| \psi_k \|^2 \).

### 2.2 Gaussian Process Regression

The GPR will be applied to estimate the near-surface flow in the vicinity of drifters by interpolating their velocities, assuming that the drifters faithfully follow the water parcels. In doing so, we want to ignore what cannot be resolved and focus on scales allowed by the separations of the drifters. By interpolating to a sufficient number of points, we intend to get a useful picture of the evolving submesoscale flow, in particular time-varying fields of horizontal components of velocity, and from these fields we intend to infer horizontal divergence, relative vorticity, and rate of strain. Our approach is to interpolate the horizontal components \( u \) and \( v \) independently, with no cross-correlation between \( u \) and \( v \) being assumed, and filter the unresolved small scales by attributing them to errors in the drifters velocities. In GPR, the scales present in the interpolated fields are controlled by a covariance function. Because drifters tend to separate as they are carried by the flow, we use a covariance function with up to two scale parameters \(^1\) for each of the zonal, meridional and temporal coordinates so that the interpolation might be tuned to benefit from both smaller and larger separations of neighboring drifters.

The zonal \( u \) and meridional \( v \) components of surface velocity are function of space-time \( p = (x, y, t) \) (where \( x, y \) and \( t \) are the zonal, meridional and time coordinates, respectively). Drifters observations are available at \( n \) space-time point and are de-

\(^1\)In chapter 5, covariance functions with 1 and 2 scales are tested.
noted by \((u_i^d, v_i^d) = (u(p_i^d), v(p_i^d))\) with \(i = 1, \ldots, n\). The superscript \(d\) refers to the observational data and the subscript \(i\) to the \(i\)-th such observation. As the GPR formulation is the same for \(u\) and \(v\) it will be presented only for \(u\), and the estimator of \(u\) at a target point \(q\) will be denoted as \(u^t(q)\).  

The main assumption of GPR is that the quantity of interest, \(u(p)\), follows a Gaussian process \(^2\), which is completely specified by its mean function \(\bar{u}(p)\) and covariance function \(\mathcal{K}(p, q)\) (Rasmussen and Williams, 2006):

\[
\begin{align*}
    u(p) &\sim \mathcal{GP}(\bar{u}(p), \mathcal{K}(p, q)). 
\end{align*}
\]  

The Gaussian process (2.4) characterizes prior beliefs in terms of a prior mean field \(\bar{u}\), which accounts for what we might guess for the velocity field in the absence of data, and \(\mathcal{K}\), which specifies covariances that account for how we expect the velocities at any two points \(p\) and \(q\) might jointly deviate from our guesses at those points. While (2.4) presumes \(p\) to represent any and all points in a continuum, in practice it is necessary to work with a finite number of points, so equation (2.4) becomes a multivariate normal distribution for the velocity components at these points with the prior mean field replaced by a vector of velocities and the covariance function replaced by a matrix of covariances.

Assuming that \(u^d\) are noisy observations of \(u\), and that the noise is independent, normally distributed and with zero mean and variance \(\sigma_N^2\), we can write the

\(^1\)Because we are interpolating from points where we have drifter data to points where we do not, it is convenient to distinguish which is which notationally with superscripts \(d\) for data and \(t\) for target.  
\(^2\)A Gaussian process is defined as a collection of normally distributed random variables, from which any finite samples have a joint Gaussian distribution (Rasmussen and Williams, 2006).
multivariate Gaussian prior distribution of \( \mathbf{u}^t \) and \( \mathbf{u}^d \) following from 2.4 as:

\[
\begin{bmatrix}
\mathbf{u}^d \\
\mathbf{u}^t
\end{bmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
\mathbf{ar{u}}^d \\
\mathbf{ar{u}}^t
\end{bmatrix},
\begin{bmatrix}
\mathbf{K}^{dd} + \sigma_N^2 \mathbf{I} & \mathbf{K}^{dt} \\
\mathbf{(K}^{dt})^\top & \mathbf{K}^{tt}
\end{bmatrix}
\right)
\]

(2.5)

where \( \mathbf{ar{u}}^d \) and \( \mathbf{ar{u}}^t \) refer to the prior estimate of \( \mathbf{u} \) at the observation and target points, respectively, \( \mathbf{K}^{dd} \) refer to the covariance matrix between pairs of observation points, \( \mathbf{K}^{tt} \) to the covariance matrix between pairs of target points, \( \mathbf{K}^{dt} \) to the covariance matrix between observation and target points, \( (\mathbf{K}^{dt})^\top \) is the matrix transpose of \( \mathbf{K}^{dt} \), and \( \mathbf{I} \) is the \( n \times n \) identity matrix. More specifically, the entries of the different vectors and matrices are given by

\[
\begin{align*}
\bar{u}^d_i &= \bar{u}(\mathbf{p}^d_i), & K^{dd}_{i,j} &= \mathcal{K}(\mathbf{p}^d_i, \mathbf{p}^d_j), & i, j = 1, 2, \ldots, n \\
\bar{u}^t_i &= \bar{u}(\mathbf{q}_i), & K^{tt}_{i,j} &= \mathcal{K}(\mathbf{q}_i, \mathbf{q}_j), & i, j = 1, 2, \ldots, m \\
K^{dt}_{i,j} &= \mathcal{K}(\mathbf{p}^d_i, \mathbf{q}_j), & i = 1, 2, \ldots, n, & j = 1, 2, \ldots, m
\end{align*}
\]

(2.6)

By conditioning the prior distribution on the observations we obtain the joint Gaussian posterior distribution:

\[
\begin{align*}
\mathbf{u}^t|\mathbf{u}^d & \sim \mathcal{N}(\mathbf{	ilde{u}}^t, \mathbf{Q}) \\
\mathbf{	ilde{u}}^t &= \mathbf{\bar{u}}^t + (\mathbf{K}^{dt})^\top \left(\mathbf{K}^{dd} + \sigma_N^2 \mathbf{I}\right)^{-1} (\mathbf{u}^d - \mathbf{\bar{u}}^d) \\
\mathbf{Q} &= \mathbf{K}^{tt} - (\mathbf{K}^{dt})^\top \left(\mathbf{K}^{dd} + \sigma_N^2 \mathbf{I}\right)^{-1} \mathbf{K}^{dt}
\end{align*}
\]

(2.7)

The expressions in equation 2.7 are the same as the ones resulting from Optimal Interpolation (Gandin, 1966) with \( \mathbf{\bar{u}} \) representing the background estimate, \( \mathbf{K}^{dd} \) the background error covariance matrix at the observation points, \( \sigma_N^2 \mathbf{I} \) the observational error covariance matrix, \( \mathbf{K}^{tt} \) the background covariance matrix at the target points, and \( \mathbf{K}^{dt} \) as the background covariance matrix between observation and target points. The posterior mean \( \mathbf{\tilde{u}}^t \) and the posterior covariance \( \mathbf{Q} \) are usually referred to as the
analysis in data assimilation.

2.2.1 Covariance functions

A key aspect of GPR, as of optimum interpolation, is the choice of the covariance function from which the covariance matrices used in the mapping are drawn from. The main requirement is that the covariance matrices must be positive semidefinite (Bretherton F. P. et al., 1976; Rasmussen and Williams, 2006), and ideally they should reflect the real space/time correlations of the field to be estimated. In applications of optimum interpolation focused on the mapping of meso and large scales circulation, covariance functions were chosen to resemble empirical estimates of the covariance matrix computed from observations and from climatology (McWilliams, 1976; Carter and Robinson, 1987; Mariano and Brown, 1992; Wilkin et al., 2002). For these scales, geostrophy and non-divergence constrains are imposed in the covariance function by the addition of cross-covariance terms between the horizontal velocity components (Bretherton F. P. et al., 1976; McWilliams, 1976; Hua et al., 1986). As our goal is to characterize the submesoscale activity, and at this scale range the flow is significantly ageostrophic, with significant horizontal divergence, the above mentioned constrains are not desirable. Given that the drifter observations to be used sampled the flow field at several separation scales (see Figure 4.2 in section 4.3), we would like to explore the possibility of capturing more than one correlation scale per space-time coordinate dimension. Previous applications of multi-scale analysis involve low-pass filtering the observations, which are used to build the larger scale. The latter is then used as prior mean for the smaller scale, which is constructed from the high-pass filtered observations. This process can be implemented sequentially, as in Li et al. (2015), or simultaneously, as in Le Traon (1990) and Le Traon and Hernandez (1992). Our approach differs from the above, as the multi-scale is added in the covariance
function, and no filtering of the observations is required. Another difference is that, in GPR, the correlation scales and signal to noise ratio are determined directly by an optimization procedure, in which their probability conditioned on the observations is maximized.

In the analysis of chapter 4, the covariance is defined by the squared exponential function (SE):

$$K_{SE}(p, p') = \sigma^2 \exp \left( -\frac{\mathbf{l}_t^2}{2} - \frac{\mathbf{l}_y^2}{2} - \frac{\mathbf{l}_x^2}{2} \right),$$  \hspace{1cm} (2.8)

where

$$\mathbf{l}_t = \frac{(t - t')}{r_t},$$
$$\mathbf{l}_y = \frac{(y - y')}{r_y},$$
$$\mathbf{l}_x = \frac{(x - x')}{r_x},$$ \hspace{1cm} (2.9)

$\sigma^2$ is the signal variance, $r_t$ is the correlation time scale, and $r_x$ and $r_y$ are the horizontal correlation scales in the zonal and meridional directions, respectively. The parameters that define the covariance function $r_t$, $r_y$, $r_x$ and $\sigma^2$, along with the noise variance $\sigma^2_N$, are referred as hyper-parameters.

In the analysis of chapter 5, combinations between four of the most used functions in machine learning (Rasmussen and Williams, 2006) are tested. Those are the above mentioned SE, the Matérn 3/2 (M32)

$$K_{M32}(p, p') = \sigma^2 \left( 1 + \sqrt{3}l_t \right) \left( 1 + \sqrt{3}l_y \right) \left( 1 + \sqrt{3}l_x \right) \exp \left[ -\sqrt{3} (l_t + l_y + l_x) \right],$$ \hspace{1cm} (2.10)

the Matérn 5/2 (M52)
\[ K_{M52}(p, p') = \sigma^2 \left( 1 + \sqrt{5}l_t + \frac{5}{3}l_t^2 \right) \left( 1 + \sqrt{5}l_y + \frac{5}{3}l_y^2 \right) \left( 1 + \sqrt{5}l_x + \frac{5}{3}l_x^2 \right) \exp \left[ -\sqrt{5} (l_t + l_y + l_x) \right], \]  

(2.11)

and the Rational Quadratic (RQ)

\[ K_{RQ}(p, p') = \sigma^2 \left( 1 + \frac{l_t^2}{2\alpha} \right)^{-\alpha} \left( 1 + \frac{l_y^2}{2\alpha} \right)^{-\alpha} \left( 1 + \frac{l_x^2}{2\alpha} \right)^{-\alpha}, \]  

(2.12)

where \( \alpha \) is a "scale mixture" parameter. The RQ can be regarded as an infinite sum of SE functions with different length scales, and the limit of RQ for \( \alpha \to \infty \) is the SE with the same length scale (Rasmussen and Williams, 2006).

### 2.2.2 Optimization procedure

To optimize the hyper-parameters of the covariance function, we applied the marginal likelihood method, described in Section 5.4.1 of Rasmussen and Williams (2006). Considering the set of hyper-parameters \( \theta \) that specify the covariance function \( K(\theta) \), the optimization procedure seeks the maximization of the probability of the hyper-parameters conditioned on the observations \( p(\theta|u_d) \). Using Bayes theorem, this is equivalent to maximizing the log marginal likelihood:

\[ \log p(u^d|\theta) = -\frac{1}{2} (u^d)^\top B^{-1} u^d - \frac{1}{2} \log |B| - \frac{n}{2} \log 2\pi, \]  

(2.13)
where \( \mathbf{B} = (\mathbf{K}^{dd} + \sigma^2_n \mathbf{I}) \) and \( n \) is the number of observations. The partial derivative of the log marginal likelihood function with respect to a hyper-parameter \( \theta_j \in \theta \) is:

\[
\frac{\partial}{\partial \theta_j} \log p(u^d | \theta) = \frac{1}{2} (u^d)^\top \mathbf{B}^{-1} \frac{\partial \mathbf{B}}{\partial \theta_j} \mathbf{B}^{-1} u^d - \frac{1}{2} \text{Tr}(\mathbf{B}^{-1} \frac{\partial \mathbf{B}}{\partial \theta_j} ) - \frac{1}{2} \text{Tr}(\beta \beta^\top - \mathbf{B}^{-1} \frac{\partial \mathbf{B}}{\partial \theta_j} ),
\]

(2.14)

where \( \beta = \mathbf{B}^{-1} u^d \). Equation 2.14 can be applied with any gradient based optimization algorithm to find the values of \( \theta \) that maximize the log marginal likelihood. Here, we used the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Byrd et al., 1995).
Chapter 3

A framework to quantify uncertainty in simulations of oil transport in the ocean

3.1 Overview

Oil spilled in the ocean poses a threat to the environment as its consequences might be catastrophic to the biota and to the communities of the impacted areas. Oil fate models are important instruments that can provide the scientific background needed for oil spill risk assessment, and for decision-making during an oil spill incident. In an emergency response, forecasting the pathways and the oil fate is essential for awareness, preparation and mitigation actions of the probable impacts.

Several oil fate models of increasing skill have been developed over the decades to represent the complex behavior of the oil in the ocean. They evolved from simple transport models, with limited fate algorithms, to current state-of-art models, which simulate the transport of the oil in three dimensions, including processes of advection, diffusion, stranding, emulsification and oil weathering (Cekirge et al., 1995; Chao et al., 2003; Berry et al., 2012). In general, oil fate models treat the spill as a
collection of Lagrangian particles, that are advected by an Eulerian velocity field. This velocity field, which is set as an input for the oil model, includes the effect of ocean currents, waves and winds, and is usually provided by the outputs of numerical models.

Regardless of how sophisticated the oil fate models are, and the amount of processes they can represent, errors in the input data will inevitably lead the results to diverge from observations (Sebastião and Guedes Soares, 2007). Oil fate predictions have to rely on data from oceanic and atmospheric forecasts, which have limited predictability and inherent uncertainty. In many cases, little is known about the incident itself, and important information, that is essential for setting up the initial conditions and other input parameters, is missing. Information like the exact location where the spill occurred, the amount of oil that was spilled, the flow rate of the oil, in case of a blowout, will likely influence the outcome of an oil fate simulation. With that in mind, the application of UQ techniques in oil fate forecasts is essential to provide meaningful information for the spill response efforts (Galt, 1997a,b).

The goal of UQ is to estimate uncertainty in the model’s output given uncertainties in its inputs. It provides means of estimating the most likely output, along with confidence levels and error estimates. Although it should be the next step for the advancement of operational oil spill forecast (Hodges et al., 2015), applications of UQ in oil fate models are still scarce in the literature. This chapter presents a framework for quantifying uncertainty in the transport of oil during oil fate simulations based on PC expansions. The method is based on constructing a computationally cheap surrogate for the model where an output is defined as a polynomial expansion of the uncertain input parameters. The surrogate is built through an ensemble of model realizations, and once constructed, it should be able to approximate the model output for any given value of the input parameters inside a pre-established range. As
obtaining outputs from the surrogate is computationally cheap, the task of sampling the model hundreds of thousands of times is feasible, and PDFs of the output can be obtained relatively fast. The ability of estimating PDFs at low computational cost is particularly important for emergency response efforts, when decisions have to be made fast. The PC approach also provides an easy and fast way to quantify the contribution of each uncertain input parameter in the output uncertainty (sensitivity analysis). This property has applications in oil spill risk assessments (Neves et al., 2015).

The capabilities of the PC approach are explored in a simulation of the Deepwater Horizon (DWH) blowout. In this accident, about 4.9 million barrels of crude oil along with a large amount of gas were released in the Gulf of Mexico (McNutt et al., 2011) following the explosion of a drilling platform on 20 April 2010. The spill lasted for 87 days, with oil being continuously released in the ocean at a depth of 1500 meters. The fact that the accident occurred in deep water increases drastically the complexity of the problem of simulating the oil spill. High pressure along with other ambient variables, such as temperature, salinity and ocean currents, affect the upward movement of the plume formed by oil and gas (Johansen, 2003; Yapa et al., 2012).

The dissolution of gas and oil components along with the entrainment of sea water causes the plume to lose buoyancy and momentum; eventually the plume reaches a neutrally buoyant level. At this level, oil droplets and gas bubbles separate from the plume, that now is composed of dissolved gas and oil mixed with sea water, and start moving upwards (Socolofsky and Adams, 2002). The initial plume stage of a deep water blowout is referred to as the near-field, and is usually simulated through integral models (Socolofsky et al., 2011). The stage when oil droplets separate from the plume and move upward, while being advected by the ocean currents, is the far-field, which is simulated through Lagrangian transport models mentioned above (Mariano et al.,
Here, the DeepC Oil Model (DCOM) is applied to simulate the far-field of the DWH spill. Two sources of uncertainty are considered. The advecting ocean currents, which are provided by the outputs of a general ocean circulation model (OGCM), and the oil droplet size distribution, which affects the buoyancy of the spilled oil. The forward propagation of uncertainty is carried out in a hindcast simulation of the first 30 days of the oil spill. We present a framework to estimate the uncertainty in the simulated oil distribution, to quantify the contribution of each uncertain input parameter considered and to construct probabilistic hazard maps of oil impact. Also, we test the skill of the PC surrogate in representing the model’s output by comparing it to a set of actual realizations of the model.

This chapter is organized as follows. Section 3.2 describes the models used, and the set up of the simulations. Section 3.3 discuss the sources of uncertainties and the choice of the input parameters. Section 3.4 displays the results, which include the mean and standard deviation of the oil distribution, sensitivity analysis, and probabilistic hazard maps. Summary and discussion are presented in section 3.5.

3.2 Model description

3.2.1 Lagrangian oil fate model

The DeepC Oil model (DCOM) is a 3D advection-diffusion and weathering code that simulates the transport and evolution of oil accidentally released from a subsea well, pipeline or a surface source such as a leaking ship. The algorithms for transport and transformations are largely derived from the works of Proctor et al. (1994); Al-Rabeh et al. (2000); Spaulding et al. (2000); Zheng and Yapa (2000); Lehr et al. (2002); Zheng et al. (2003); Tkalich et al. (2003); Korotenko et al. (2010); Socolofsky et al.
(2011). The intended functionality of the model is to predict where, when and in what state the oil is most likely to be found based on information about the ocean currents, winds, waves, oil properties and other environmental conditions. The model is based on a Lagrangian formulation where the oil spill is represented by a large number of particles with time-varying position, mass, volume and composition. The model takes as input the release conditions, initial position and properties of the oil along with time-varying environmental conditions including winds, waves, water currents, and water temperature and salinity. The position and composition of the particles are then updated based on environmental conditions. The present article focuses solely on the evolution of the oil caused by transport, the effect of other processes have been turned off. We thus focus on describing in some details the transport processes including the buoyancy calculation.

The positions of the particles \( \mathbf{x} \) are evolved in time \( t \) according to the following equation

\[
\frac{d\mathbf{x}}{dt} = \mathbf{u}_a + \mathbf{u}_s + k \mathbf{w}_b, \tag{3.1}
\]

Here \( \mathbf{u}_a \) represents the sum of advective processes, namely ocean currents, Stokes drift and wind drag. These velocities must be obtained from operational forecasting systems and account for the transport due to dynamical processes that can be resolved on the forecasting systems computational grids. The unresolved processes are represented by \( \mathbf{u}_s \) and account for subgrid scale turbulence; DCOM uses a random walk model (Griffa, 1996) which adds a random displacement to the particles at each timestep. The last term on the right hand side of equation 3.1 is the buoyant velocity of the oil droplets, \( \mathbf{w}_b \). The buoyancy of the droplets is a function of their size, and \( \mathbf{w}_b \) is computed following a two equations approach. For droplet sizes below a critical diameter \( d_c \), so that the Reynolds number is low, \( \mathbf{w}_b \) is given by Stokes law. For
diameters greater than $d_c$ (higher Reynolds number), $w_b$ follows a formulation proposed by Zheng and Yapa (2000). In DCOM, Equation 3.1 is solved by a 4th order Runge-Kutta scheme.

DCOM uses a probabilistic approach to represent the range of droplet sizes found in a deep water oil spill: each model particle is assigned an oil droplet diameter sampled from a distribution. DCOM uses the Rosin-Rammler distribution, as that was found to be a good fit based on different laboratory and field experiments (Yapa and Chen, 2004; Brandvik et al., 2013; Johansen et al., 2013; Aman et al., 2015). Its cumulative probability function, $P(d)$, is

$$P(d) = 1 - \exp \left[ \ln \left( \frac{1}{2} \right) \left( \frac{d}{d_{50}} \right)^\alpha \right]$$  (3.2)

where $d$ is the droplet size, $d_{50}$ is the 50th percentile of the distribution, and $\alpha$ is the spreading parameter.

### 3.2.2 Hydrodynamic model

To force DCOM, we used ocean currents, temperature and salinity data from a Gulf of Mexico simulation performed with the HYbrid Coordinate Ocean Model (HYCOM, http://hycom.org). HYCOM is a generalized vertical coordinate model, which is widely used for research and operational tasks (Bleck, 2002; Chassignet et al., 2003, 2006, 2009). Typically, the vertical coordinate is isopycnic in the stratified interior, z-level or pressure coordinates in the mixed layer, and terrain-following in coastal regions. Here, outputs of a data assimilative 2003-2012 hindcast were used (Chassignet and Srinivasan, 2015). The hindcast is based on the 1/25-degree GoM-HYCOM system, which is used in near real time by the Naval Research Lab (http://www7320.nrlssc.navy.mil/hycomGOM/). The model is configured with 20 vertical layers, and its domain extends from approximately 98W to 76.5W, and from
18.9N to 31.5N. It includes a realistic bottom topography, derived from a recent high resolution bathymetry available from Florida State University. The hindcast is forced at the surface with the North American Regional Reanalysis (NARR) atmospheric product (Mesinger et al., 2006). Surface fluxes of heat and momentum are calculated through bulk formula using fields of wind stress, 10 m wind speed, 2 m air temperature, 2 m water vapor mixing ratio, precipitation, and incoming solar radiation. A correction to the NARR derived radiation fluxes were derived by comparing model surface temperature with Advanced Very High Resolution Radiometer (AVHRR) estimates of surface temperature fields. The NARR wind speed was corrected by regression to QuikSCAT wind speeds. Monthly climatological Gulf of Mexico river data from USGS database (http://waterdata.usgs.gov/nwis/sw) and the RIVDIS database (http://www.RivDis.sr.unh.edu) were specified for the freshwater discharge (treated as a virtual salt flux). The hindcast simulation uses bi-weekly climatological lateral open boundary conditions derived from a North Atlantic climatological run.

The data assimilation is implemented with the Tendral-Statistical Interpolation System (Halliwell et al., 2014). The primary data set for constraining the model are daily along track sea level anomalies. The data were obtained from multiple operational satellite altimeters, available from Collecte Localisation Satellites in a delayed time quality controlled mode. Daily satellite sea surface temperature, obtained from the National Oceanic and Atmospheric Administration (NOAA), in-situ profiles of temperature and salinity from the World Ocean Atlas 2009, and data from cruises during the DWH incident were also used. For an evaluation of the hindcast, the reader is referred to Chassignet and Srinivasan (2015).
3.2.3 Model set up

The numerical experiments were set to represent the first 30 days of the DWH oil spill. Despite the fact that the blowout occurred at 1500 meters depth, studies have indicated that the transition between the near-field and far-field in the DWH spill occurred around 1200 meters (Socolofsky et al., 2011; North et al., 2011). So, a thousand particles were released every 30 minutes from a source located at 88.3°W and 28.7°N at a depth of 1200 meters. The velocities used to transport the particles ($u_a$ in Equation 3.1) were given by the outputs of the HYCOM hindcast, from 20 April 2010, to 19 May 2010 and displacements of the particles were computed with a time step of 30 minutes. The oil density was set to 858 kg/m³, while the sea water density was computed using temperature and salinity from the HYCOM hindcast. In our experiments only ocean currents were considered.

We define our quantity of interest as the oil concentration at some specific location after a period of time. Despite the difference in droplet sizes between the particles, it was considered that each particle carried the same amount of oil. This way, the simulated oil spill had the same flow rate for the different realizations of the model, as the number of particles released in each realization was the same. The oil concentration was computed by binning the particles in a grid, where the number of particles at each grid cell was multiplied by a constant oil mass, and divided by the volume of the cell. Oil concentrations were obtained for three different depths: between 0 – 5, 300 – 600 and 900 – 1200 meters depth, which will be referred to as upper layer, mid-depth layer and deep layer, respectively. In all of these layers, the particles were binned in a grid with 0.1 degrees resolution in both zonal and meridional directions.
3.3 Input parameters uncertainty

The choice of the input parameters to be perturbed has to fall on those whose uncertainty are more likely to have a higher influence on the quantity of interest. On the other hand, the number of parameters to be perturbed is limited by the “curse of dimensionality”, i.e. the number of simulations required to build the PC surrogate grows exponentially with the number of uncertain parameters considered (or dimensions in the uncertain space). Here, we are interested in the sources of uncertainty that controls the transport of oil in a deepwater blowout. For that matter, we focused solely on the the oil droplet size distribution, which controls the droplets rise velocity, and the advecting currents, which is the main factor responsible for the horizontal spread of the oil throughout the water column. Other important sources of uncertainty, which won’t be considered in this study, include the subgrid scale parametrization, wind drag at the surface, Stokes drift, and the oil degradation.

A key aspect of UQ is the choice of the range that an input parameter is to be perturbed. We would like this range to be wide enough to account for all possible values of the parameter. On the other hand, the response of the quantity of interest to the perturbations inside the chosen range should be fairly smooth, so that the PC expansion can approximate the behavior of the model. If the range of the input values is too broad, the response of the model might be more complicated and would require a longer series to represent accurately. Also, the parameter space might present different regimes, with each regime being characterized by a local uncertainty range. If such local ranges are known, they should be accounted for separately.

In the PC approach, it is convenient to describe the input uncertainty by standardized random variables $|\xi| \leq 1$. Considering the vector of uncertain input parameters $\theta = (\theta_1, \ldots, \theta_d, \ldots, \theta_D)^T$, where $D$ is the number of uncertain inputs, each parameter is related to a standardized random variable by:
\[ \theta_d = \frac{\theta_{d}^{\text{max}} + \theta_{d}^{\text{min}}}{2} + \xi_d \frac{\theta_{d}^{\text{max}} - \theta_{d}^{\text{min}}}{2} \] (3.3)

where \( \theta_{d}^{\text{min}} \leq \theta_d \leq \theta_{d}^{\text{max}} \) is the range of the \( d \)-th uncertain input. An important and difficult step in UQ is the definition of the PDFs of the uncertain inputs. Due to the lack of information about the input parameters considered, we defined the PDFs of all input parameters with an uniform distribution. A description of the input parameters and their perturbation range is presented in the following subsections.

### 3.3.1 Droplet size distribution uncertainty

In situ measurements of the droplet size distribution during accidents like the DWH blowout are usually difficult and highly uncertain. Several conditions influence the size of the droplets, as they break up/coalesce into smaller/bigger droplets; these include the ambient state (pressure, temperature, ocean currents, etc), the intensity of the flux at the source, the oil viscosity, and oil-water surface tension (Yapa et al., 2012; Zhao et al., 2014). The use of dispersants as a measure to prevent the oil from reaching the surface further increases the uncertainty of the droplet size distribution, as they reduce the size of the droplets by decreasing the oil-water interfacial tension.

The Rosin-Rammler distribution, used here to represent the droplet size distribution, is defined by \( \alpha \) and \( d_{50} \) (equation 3.2), which control its shape and range. Due to the high uncertainty of the droplet size distribution and its importance in determining the vertical velocity of the oil, both parameters would be good candidates to be perturbed. In a preliminary experiment, which won’t be presented here, both parameters were considered uncertain; \( d_{50} \), however, turned out to have a much bigger impact on the oil fate than \( \alpha \). Henceforth, only \( d_{50} \) will be perturbed here in order to reduce the dimensionality of the uncertain space, while \( \alpha \) is set to 1.8.

To define a range of values for \( d_{50} \), we considered that the oil plume was under the
influence of dispersants, as about 771,000 gal of dispersants were released next to the wellhead during the DWH spill (Kujawinski et al., 2011). Our choice for the range of $d_{50}$ was based on the results from Zhao et al. (2014), where a droplet formation model was applied to estimate the droplet size distribution in the DWH blowout considering different degrees of reduction on the oil-water interfacial tension. For a 1000 fold reduction, their model produced a distribution with $d_{50} = 260 \mu m$, while for a 10 fold reduction, the droplet median diameter was $d_{50} = 890 \mu m$. We defined the range of $d_{50}$ to be between 200 $\mu m$ and 900 $\mu m$. Figure 3.1 shows the droplet size distributions in the extreme cases, when $d_{50} = 200 \mu m$ and $d_{50} = 900 \mu m$.

Figure 3.1: Rosin-Rammler distributions considering $d_{50} = 200 \mu m$ (blue) and $d_{50} = 900 \mu m$ (red). For both distributions, the spreading parameter is $\alpha = 1.8$.
3.3.2 Eulerian velocity field

There are several sources of uncertainty that are present in the modeled HYCOM Eulerian velocity fields. In addition to truncation errors associated with the model numerics and grid choices, the solution is quite sensitive to choices in subgrid scale parameterizations and the specification of initial and boundary conditions. Quantifying the effects of all sources would require thousands of HYCOM realizations with different combinations of values for each parameter and for each collection of initial and boundary conditions. Unfortunately, that task is impractical due to the overwhelming computational cost. Instead, we would like to be able to propagate uncertainty in the velocity field using a reduced number of parameters. Moreover, we would like to perturb the velocity field in every grid cell without running extra realizations of HYCOM, using only the outputs of a single run, and with the perturbations evolving in time. Our approach to this problem was to characterize the perturbations as a sum of products of spatial patterns and time series using empirical orthogonal function (EOF) decomposition. The spatial patterns are given by the multivariate EOFs of the zonal and meridional components of the velocity field, and the time series are given by the respective principal components. A similar approach was applied by Thacker et al. (2012) to propagate uncertainty through boundary conditions in a HYCOM simulation of the Gulf of Mexico circulation.

The perturbations on the Eulerian velocity field were carried out in the following manner:

\[ u_a = u_0 + \sum_{k=1}^{K} \xi_k u_k, \]  

(3.4)

where \( u_0 \) is the unperturbed horizontal velocity field, \( u_k \) is the k-th perturbation of the horizontal velocity field, \( \xi_k \) is the standardized random variable that controls the k-th perturbation, and \( K \) is the number of modes used. The perturbations are
given by:

\[ u_k = N_T c_k r_k, \]  \hspace{1cm} (3.5)

where \( c_k \) is the k-th EOF, \( r_k \) is its respective the principal component, and \( N_T \) is a normalization factor, which here is the number of time steps considered in the velocity field.

The goal of this approach is to generate a distribution of velocity fields that capture the uncertainty in the HYCOM velocity field \( u_0 \) used in the DeepC Oil model. This task is similar to that of characterizing an error-covariance matrix for data assimilation and suffers from the same limitation of insufficient information to quantify the error variances of velocity at each grid point and their mutual covariances. For our approach to be practical, these uncertainties must be parameterized by just a few numbers, the coefficients \( \xi_n \) used in the expansion, each adjusting the strength of a pattern characterizing co-varying uncertainties.

The issue becomes how to design these patterns. Clearly, their design should reflect the uncertainties likely to have the greatest impact on the particular outputs of the DeepC Oil model that are of greatest interest. And even then, there are more ways to design them than can be explored here. We have chosen to base them on an EOF decomposition of the HYCOM velocity field during the interval of the simulation, and this way, assign greater uncertainty in regions of higher variability. To insure that the perturbations are appropriate to the region of the spill, the EOF decomposition was restricted to the area limited by latitudes 23°N and 31°N and longitudes 82°W and 92°W. The first two EOFs, which represent 56% of the total variance (41% the first mode, and 15% the second), were used to propagate uncertainty in the velocity field. Figure 3.2 presents maps of the percentage of the variance that is explained by both modes together at different depths. This plot illustrates where the imposed perturbations will mostly impact the velocity field.
Figure 3.2: Percentage of the total variance explained by the 1st and 2nd EOFs combined for the zonal (U, left panels) and meridional (V, right panels) velocity components, in three selected depths. The black dots depict the position of the Deepwater Horizon wellhead.

The unperturbed surface velocity ($u_0$) and the perturbations ($u_1$ and $u_2$) are presented on figure 3.3 for days 10, 20 and 30 of the DWH oil spill (29-April-2010, 09-May-2010 and 19-May-2010, respectively). Both modes are dominated by variability
from meanders of the Loop Current and Loop Current frontal eddies, although variability related to coastal currents are also present. These perturbations may intensify, weaken or deform circulation features. Figure 3.4 presents 3 perturbed velocity fields for day 30, when both $u_1$ and $u_2$ are strong (see Figures 3.3 and 3.5). These velocity fields ($u_a$) were obtained considering $\xi_1 = 1$ and $\xi_2 = 1$ ($u_a = u_0 + u_1 + u_2$), $\xi_1 = 1$ and $\xi_2 = 0$ ($u_a = u_0 + u_1$), and $\xi_1 = 0$ and $\xi_2 = 1$ ($u_a = u_0 + u_2$).

**Figure 3.3:** Surface velocity vectors of the unperturbed velocity field (left) and of the perturbations constructed from the first (center) and second (right) EOFs. The magenta dots depict the position of the Deepwater Horizon wellhead.
Figure 3.4: Surface velocity vectors at 19-May-2010 for the unperturbed velocity field $u_a = u_0$ (upper left) and for three perturbed fields $u_a = u_0 + u_1 + u_2$ (upper right), $u_a = u_0 + u_1$ (lower left), and $u_a = u_0 + u_2$ (lower right). The magenta dots depict the position of the Deepwater Horizon wellhead.

Figure 3.5: Principal components 1 and 2 of the velocity field from HYCOM, starting at 20-April-2010.
The effect of the perturbations on the dynamics are illustrated in the maps of vertical vorticity and lateral strain, which are presented on figures 3.6 and 3.7. The vorticity maps show the intensification of frontal cyclones and on the Loop Current in the perturbed velocity fields, most notably for \( u_a = u_0 + u_1 + u_2 \). The strain rate maps show intensification in the Loop current system and over the continental shelf break of the northern Gulf of Mexico coast, including the Mississippi delta region. The circulation in the latter is dominated by wind driven currents, and by a cyclonic eddy east of the DWH wellhead. The increased strain rate in the perturbed velocities indicates an enhanced dispersion around the DWH site. An opposite effect occurs for negative values of \( \xi_1 \) and \( \xi_2 \), where for the same time, the perturbations induce a weakening in the vorticity and strain rates of the same features (not shown).

**Figure 3.6**: Vertical component of relative vorticity normalized by planetary vorticity from surface velocity fields on 19-May-2010. Upper left: \( u_a = u_0 \) (unperturbed field, i.e. \( \xi_1 = \xi_2 = 0 \)). Upper right: \( u_a = u_0 + u_1 + u_2 \) (\( \xi_1 = \xi_2 = 1 \)). Lower left: \( u_a = u_0 + u_1 \) (\( \xi_1 = 1 \) and \( \xi_2 = 0 \)). Lower right: \( u_a = u_0 + u_2 \) (\( \xi_1 = 0 \) and \( \xi_2 = 1 \)).
Figure 3.7: Lateral strain rate normalized by planetary vorticity from surface velocity fields on 19-May-2010. Upper left: $u_a = u_0$ (unperturbed field, i.e. $\xi_1 = \xi_2 = 0$). Upper right: $u_a = u_0 + u_1 + u_2$ ($\xi_1 = \xi_2 = 1$). Lower left: $u_a = u_0 + u_1$ ($\xi_1 = 1$ and $\xi_2 = 0$). Lower right: $u_a = u_0 + u_2$ ($\xi_1 = 0$ and $\xi_2 = 1$).

3.4 Results

3.4.1 PC representation of oil concentration

In order to use the PC expansion as a surrogate for the model, it is important to check its consistency against actual model outputs. A metric that approximates the relative root-mean-square error (RMSE) of the PC surrogate was computed by using differences between the surrogate and the model at a large number of different combinations of the random inputs (Alexanderian et al., 2012a):
RMSE = \left( \frac{\sum_{\xi \in S} |C(x, t, \xi) - \tilde{C}(x, t, \xi)|^2}{\left( \sum_{\xi \in S} |C(x, t, \xi)|^2 \right)^{1/2}} \right)^{1/2} \quad (3.6)

Here, $S$ is an ensemble of 1024 independent realizations of the model, where the values of the uncertain input parameters $\xi(\xi_1, \xi_2, \xi_3)$ are randomly chosen from the assumed three-dimensional uniform distribution, $C$ represents the oil concentration computed from the model’s outputs, and $\tilde{C}$ is the respective PC surrogate. Figure 3.8 shows the scatter plot of all 1024 samples of $\tilde{C}$ and $C$ at two grid cells on 19 May 2010. Point A illustrates a grid cell where the PC expansion represents satisfactorily the model outputs, with $RMSE = 0.13$. The scatter plot for this grid cell shows little spread around the linear regression line, which has a slope $s = 0.96$, and the concentrations present a correlation coefficient of 0.98. On the other hand, point B illustrates a grid cell where the PC expansion poorly approximate the model outputs, with $RMSE = 0.59$. In this case, many samples of the surrogate produced negative concentrations. Here, PC expansions with $RMSE \leq 0.1$ will be considered as suitable to be used as a surrogate for the model.

As presented in section 2.1, the mean and standard deviation of the quantity of interest can be computed directly from the coefficients of the PC expansion. They are a proxy for the mean and standard deviation of the distribution of all possible model runs inside the range of the uncertain parameters. Maps of the mean oil concentration in the upper layer (between 0 and 5 meters depth) are presented in the left panels of figure 3.9 at days 10, 20 and 30 of the spill (29 April 2010, 09 May 2010 and 19 May 2010). As the values range within several orders of magnitude, the maps are presented in a logarithmic scale. The panels on the right show the ratio of the mean over the standard deviations. Smaller values (in red) indicate relatively higher
standard deviation, and therefore, higher uncertainty. In most cases, places with higher mean oil concentration presented relatively smaller uncertainty. On the edges of the oil plume, where concentrations are smaller, the standard deviation reached values which were equal or higher than the mean. Uncertainty is also high in the regions dominated by the mesoscale variability of the Loop Current system.

Although the scope of this chapter is mainly to present the PC approach to simulations of oil transport in the ocean, a comparison of our results to the observed oil slick from the DWH spill may bring some insight of what processes were well represented, and what was missed from our simulations. We compared the results of oil concentration in the surface layer to daily composite products from NOAA’s Satellite Analysis Branch (Streett, 2011), which is composed by synthetic aperture radar and visible/near infra-red multispectral satellite imagery. Only qualitative comparisons can be made, as the daily composites just provide information about the presence of the oil slick, not giving any information that can be related to oil concentration. The detection of the oil slick through SAR imagery presents some limiting factors

---

**Figure 3.8:** Left: RMSE of the oil concentration in the upper layer after 30 days. The RMSE was computed through a set of 1024 model realizations using randomly chosen sets of the uncertain input parameters and their respective surrogate. The values of oil concentration (in $10^4 \text{kg/m}^3$) from the model realizations and their surrogates in points A and B are presented in the scatter plots (center and right, respectively). The red lines are the linear regression of the scatter plots, and Corr. and Sl. are the correlation coefficients and slopes, respectively. The magenta dot indicates the DWH wellhead location.
that may affect the quality of the composites. The most important one is that the slick cannot be detected under strong (greater than 25 knots) or low (less than 4 knots) wind conditions, as these conditions induce false positive and false negative, respectively. Other important factors that affect the quality and availability of the composites are cloud cover and limited satellite image coverage, which may not cover the whole spill area every day (Streett, 2011; Dietrich et al., 2012). From the period that we are considering (initial 30 days of the oil spill), only 2 daily composites are available; from 17 May 2010 and from 20 May 2010, although the quality of the last one is poor due to the lack of imagery available. Thereby, we compared the these two composites with the simulated oil concentration in the surface layer on day 19 May 2010 (fig. 3.9). During this period, part of the oil slick was advected southwards into a frontal cyclone north of the Loop Current, while another portion was advected westward along the Louisiana coast by wind driven currents (Walker et al., 2011). The first was well captured by the simulations, while the westward advection of the oil was underestimated. In the overall, the observed spread of the oil was mostly inside the uncertainty range of our simulations. On the other hand, the simulated oil transport to the east, which resulted in oil arriving in the northwestern coast of Florida by 9 May 2010, and in the Straits of Florida by 14 May 2010, was not observed during the oil spill. In the simulations, the oil was advected eastward/northeastward during the first 10 days of the spill (see day 29 April 2010 on figure 3.9). Part of the oil advected eastward was then transported to the south, where it reached the Loop Current by 9 May 2010. From there, it went through the Straits of Florida by 14 May 2010. During this period, when simulations presented a strong eastward transport, observations do not shown any evidence that oil was transported east of 87°W (Liu et al., 2011a; Walker et al., 2011). The discrepancy between simulations and observations could be simply explained by ocean circulation patterns that were
not present in the HYCOM velocity field. On the other hand, it could also be caused by the neglected effects of wind and waves. Numerical experiments of Le Hénaff et al. (2012) showed that the direct effect of the wind on the DWH oil slick might have constrained the spread of the oil closer to the Louisiana coast in the initial stages of the oil spill, which kept the oil from reaching the Loop Current, and thereafter the Straits of Florida. Huntley et al. (2011) tested the direct effect of the wind in different stages of the DWH spill, and obtained better results with a parameterized wind effect for the oil transport in coastal areas, although this effect was negligible away from the shelf. The effect of waves through stokes drift in the transport of the surface oil slick is very similar to the direct effect of wind, and as Carratelli et al. (2011) described, it is hard to distinguish both conceptually and practically. Dietrich et al. (2012) used a high resolution finite element model coupled with a wave model to simulate the transport of oil from the DWH over the continental shelf, and found better results without adding the direct effect of the wind.

The mean and standard deviation for the mid-depth and deep layers are presented in Figures 3.10 and 3.11. The deep layer reached higher concentrations than the mid-depth layer, as ocean currents are stronger in the latter, and consequently, dispersion is higher. Another aspect which influenced this behavior was that particles with smaller droplet sizes got trapped in the deep layer, while all particles that reached the mid-depth range kept moving towards the surface. As both of these layers span a depth range of 300 m (300-600 m for the mid-depth layer, and 900-1200 m for the deep layer), oil concentrations can’t be directly compared to the concentrations in the upper layer, which span a 5 m depth range, as the Lagrangian particles are binned in a larger volume. Uncertainty was high in both layers, with values of standard deviation reaching twice the mean concentration in the edges of the oil plume, where concentrations are low.
Figure 3.9: Mean oil concentration between 0 – 5 meters depth (in kg/m$^3$, left), and mean over standard deviation ratio (right), on days 10 (29 April 2010), 20 (9 May 2010), and 30 (19 May 2010) of the oil spill. Both mean and standard deviation were extracted from the PC coefficients. The contours on the plots of 19 May 2010 indicate the position of the oil spill on 17 May 2010 (cyan) and 20 May 2010 (blue) as observed from a composite of satellite images. The black contours enclose cells for which the surrogate reproduces the model outputs with a $RMSE \leq 0.1$, and the black dot indicates the DWH wellhead location.
Figure 3.10: Mean oil concentration between 300 – 600 meters depth (in kg/m³, left), and mean over standard deviation ratio (right), on days 10 (29 April 2010), 20 (9 May 2010), and 30 (19 May 2010) of the oil spill. The black contours enclose cells for which the surrogate reproduces the model outputs with a $RMSE \leq 0.1$, and the black dot indicates the DWH wellhead location.
Figure 3.11: Mean oil concentration between 900 – 1200 meters depth (in kg/m$^3$, left), and mean over standard deviation ratio (right), on days 10 (29 April 2010), 20 (9 May 2010), and 30 (19 May 2010) of the oil spill. The black contours enclose cells for which the surrogate reproduces the model outputs with a RMSE $\leq 0.1$, and the black dot indicates the DWH wellhead location.

In the upper layer, most of the oil is inside the RMSE $\leq 0.1$ margin in the first 10-15 days (Figure 3.9 shows day 10), indicating that the surrogate is a good representation of the model during this period. After 15 days, the oil spreads towards
and around the Loop Current system where the errors of the surrogate increase above 0.1. At all depth ranges analyzed, the areas with a $RMSE > 0.1$ presented high uncertainty, and for the two deeper layers, most of these areas occurred where the standard deviation was at least twice as high as the mean. Also, in the deeper layers, most of the areas with $RMSE > 0.1$ presented an oil concentration $C \leq 10^{-6} \, kg/m^3$.

The results in the upper layer show that the quality of the PC surrogate deteriorates in areas dominated by mesoscale variability. This is explained by the fact that the perturbations in the velocity field are larger there. Another problem occurs in areas with a relatively small oil concentration. In these areas, the grid cells used to compute the oil concentration are reached by a small number of Lagrangian particles. With that, a slight variation in the number of particles in these grid cells, among different realizations of the model, results in a relatively abrupt change of oil concentration. In this case, the response of the model to the different values of input parameters are not accurately represented by the 10 degree PC surrogate. These problems could be attenuated by increasing the order of the PC expansions, although this measure would require a considerably higher number of model realizations to compute the coefficients. For example, increasing the order from 10 to 11 would require another 397 model realizations. Another measure to reduce the errors in the surrogate would be to increase the area of the grid cells in which the Lagrangian particles are binned. In this case, the drawback is the loss of spatial resolution.

### 3.4.2 Sensitivity analysis

In order to quantify the contribution of each input parameter on the uncertainty of our quantity of interest, we took advantage of the PC coefficients, as they provide an efficient and computationally cheap way to perform variance-based sensitivity analysis. A global sensitivity index was computed as the ratio between the fraction
of the variance due to a specific input parameter and the total variance considering all parameters (Iskandarani et al., 2016b; Alexanderian et al., 2012b). As the index quantifies the total sensitivity of the model to a specific parameter, it also takes into account the interaction between the different parameters.

The sensitivity indices for the three input parameters in the upper layer for days 10, 20 and 30 of the oil spill, are presented in figure 3.12. The blue contour indicates the $\text{RMSE} = 0.10$, presented previously. Overall, the perturbations on the velocity field dominated the variance, with the 1st EOF mode being dominant on days 10 and 30, and the 2nd mode being more significant on day 20. This behavior follows the patterns of the principal components, as on day 20, the absolute value of the 2nd mode is higher than the 1st mode (fig. 3.5). Comparing figures 3.3 and 3.12 elucidates how each mode impacts the oil concentration uncertainty. On day 10, the oil is being transported towards the shelf by wind driven currents, which at this moment are affected by the 1st EOF. On day 20, the oil is reaching the Loop Current, which is perturbed by the 2nd EOF. On day 30, both modes impact the circulation on the Loop current system and closer to the coast. The contribution of the droplet size distribution was more prominent on day 10, but it still reached as much as 60% of the total variance in some areas on days 20 and 30. In the mid-depth layer, the velocity perturbations are still dominant, although the contribution of the droplet size distribution uncertainty is much higher than in the surface layer (figure 3.13). In the deep layer, the uncertainty in the droplet size distribution dominated the variance in the core of the deep plume, while the 1st EOF mode contributed to the uncertainty on its edges. Figure 3.14 shows the sensitivity indices only for day 30, as on the other days the results followed the same pattern. The effect of the 2nd EOF at this depth range is negligible, which reflects that the 2nd EOF signal around this area is relatively weak.
Figure 3.12: Global sensitivity index at the upper layer for the 1st and 2nd EOF modes, and the 50th percentile of the droplet size distribution ($d_{50}$), on days 10 (29 April 2010), 20 (09 May 2010) and 30 (19 May 2010) of the oil spill. The blue contour encloses cells for which the surrogate reproduces the model outputs with a $RMSE \leq 0.1$. The green line on the 19 May 2010 plots indicates the grid cells used to compute the total amount of oil that reached the northwestern coast of Florida during the first 30 days of the oil spill (Figure 3.15).
Figure 3.13: Same as figure 3.12 but for the mid-depth layer.

Figure 3.14: Global sensitivity index for the deep layer for the three input parameters on day 30.

In the upper layer, in the areas outside the $RMSE = 0.1$ contour, where the PC
surrogate did not approximate the output of the model accurately, the variance was largely dominated by the EOF modes. This indicates that the perturbations of the velocity field induced a complex response in the oil concentration in those areas that could not be captured by the 10th order polynomial expansion. This problem could be attenuated if we increased the size of the grid cells in which we binned the Lagrangian particles to compute the oil concentration. As the Eulerian velocity is what controls the horizontal displacement of the particles, a coarser grid would reduce the impact of the velocity perturbations in the uncertainty of our quantity of interest. Similarly, a bigger depth range of the layer used to compute the oil concentrations would reduce the impact of the droplet size distribution uncertainty. If instead of defining our quantity of interest as oil concentration in a grid, we define it as the amount of oil that reached some sensitive area, the response of the model to perturbations in each parameter will be different. Let’s consider that our quantity of interest now is the amount of oil that reaches the shoreline of northwestern Florida (green line on day 19 May 2010 plot of figure 3.12) during the first 30 days of oil spill. Figure 3.15 shows the curves of the total amount of oil that reaches this area as a function of each uncertain parameter. For each curve, the PC surrogate was sampled 10,000 times with the values of one specific parameter being randomly chosen from an uniform distribution, while the other two parameters were kept zero. These results indicate that the uncertainty in this quantity of interest was largely dominated by the droplet size distribution, with the amount of oil varying by an order of magnitude between the simulation considering a distribution with $d_{50} = 0.2\,mm$ (related to $\xi_3 = -1$), and a simulation considering $d_{50} = 0.9\,mm$ ($\xi_3 = 1$). The increased impact of $d_{50}$ in this case is not surprising, as the area considered to compute the amount of oil that reached the shoreline is 10 times bigger than the horizontal area of the grid cells used to compute the oil concentrations. On the other hand, these curves don’t show the
combined effect of the parameters, as the global sensitivity index does.

Figure 3.15: Amount of oil that reached the coast of Florida (green line on figure 3.12) after 30 days considering different input parameters. $\xi_1$, $\xi_2$ and $\xi_3$ are the standardized random variables used to propagate uncertainty through the 1st and 2nd EOFs, and the 50th percentile of the droplet size distribution, respectively. The red dots are the outputs of simulations considering $\xi_1 \in [-1, 1], \xi_2 = 0, \xi_3 = 0$, the black dots are from simulations with $\xi_2 \in [-1, 1], \xi_1 = 0, \xi_3 = 0$, and the blue dots are from simulations with $\xi_3 \in [-1, 1], \xi_1 = 0, \xi_2 = 0$. The big dots are outputs from actual realizations of the model, while the smaller ones were obtained from the PC surrogate.

3.4.3 Probabilistic maps

Probability density functions of oil concentration were constructed at each grid point by sampling the PC expansion 100,000 times with randomly chosen values of the variables $\xi_1$, $\xi_2$ and $\xi_3$. These PDFs provide valuable information for impact assessment and for contingency planning during oil spill events. By extracting the 10th percentile of each distribution, we can get a measure of “best case scenario” of the oil spread,
where there is a 90% chance of the concentration to reach that value for the uncertain parameters space considered. On the other hand, the 90th percentile provides a measure of “worst case scenario”, where in only 10% of the uncertain parameters space the oil concentration suppress its value. Figure 3.16 presents the maps of the 10th and 90th percentiles of the PDFs of oil concentration in the upper layer at days 15, 20 and 25. These plots highlight the importance of taking into account the uncertainty of the model in the oil fate forecast. For instance, in the worst case scenario, the oil reaches the coast of Florida, near Panama City, on day 15, while on the best case scenario, the oil reaches this area after day 20. Also, in the worst case scenario, the oil slick would reach the Mississippi Delta by day 25, which didn’t happen in the best case scenario.

Another way to use these PDFs is to build probabilistic hazard maps given a threshold value (Madankan et al., 2014). Considering that in our simulations we didn’t take into account any type of oil degradation, it is hard to define what concentration would represent some risk to the environment, which could be used as threshold. Our choice here was arbitrary, as we picked a concentration that was two orders of magnitude smaller than the maximum concentration obtained in the upper layer. Figure 3.17 shows the probability of the oil concentration to reach at least $10^{-5}$ $kg/m^3$ in the upper layer on days 15, 20 and 25. According to these maps, it is more likely that the oil would reach the coast of Florida after 20 days. Also, the maps show a high probability for the oil slick to reach the Mississippi Delta by day 25.
Figure 3.16: 10th (left panels) and 90th (right panels) percentiles of the PDF of oil the concentration (kg/m$^3$) between 0 and 5 meters depth for days 15 (04 May 2010), 20 (09 May 2010) and 25 (14 May 2010) of the oil spill, considering $10^5$ samples of the PC proxy. The blue contour encloses cells for which the surrogate reproduces the model outputs with a $RMSE \leq 0.1$, and the black dot indicates the DWH wellhead location.
Figure 3.17: Probability of oil concentration (in %) to reach at least $10^{-5} kg/m^3$ between 0 and 5 meters depth on days 15 (04 May 2010), 20 (09 May 2010) and 25 (14 May 2010) of the oil spill. The blue contour encloses cells for which the surrogate reproduces the model outputs with a $RMSE \leq 0.1$, and the black dot indicates the DWH wellhead location.

3.5 Summary and discussion

The purpose of this chapter was to present a methodology to account for model uncertainty in oil fate simulation during a deepwater spill. The ensemble-based PC approach was used to propagate uncertainty in the Eulerian velocity field and droplet size distribution of a Lagrangian oil fate model. The method aims to represent a model output as a polynomial series expansion of model input parameters, and this representation can then be used as a surrogate for the model. In order to test the capabilities of the PC expansion, we applied the technique on the first 30 days of the DWH oil spill, and checked the quality of the PC representation of the model during the simulation period. A variance-based sensitivity analysis was carried out through the PC coefficients, where the total contribution of each source of uncertainty considered was quantified at different depths. Finally, we used the PC surrogate to construct probability density functions of oil concentrations, and applied them to elaborate probabilistic hazard maps.

The biggest challenge we faced was to find a reasonable way to perturb the Eule-
rian velocity field without having to run extra realizations of HYCOM, and prescribe
the perturbations with a small number of parameters. By using the EOF modes to
accomplish that, we assumed that the uncertainty in the velocity field from HYCOM
had patterns of variability that resemble the variability it exhibited during the pe-
riod analyzed. This way, higher uncertainty was assigned where the flow was more
energetic, like in regions dominated by mesoscale variability, which is a reasonable
assumption. On the other hand, if HYCOM missed completely some physical process
or feature, like a mesoscale eddy, due to misspecification of some parametrization (or
boundary/initial conditions), this source of uncertainty was not accounted for by the
EOF modes. Our approach targets the effect of relatively small deviations from the
Eulerian velocity field, but, as in any other approach to estimate oil transport in the
ocean, a reliable representation of the ocean state is paramount.

It is important to point out that, as our focus was to demonstrate the capabilities
of the methodology, we were not concerned about accurately reproducing observations
of the oil slick from the DWH spill. Actually, several important aspects prevented our
results from correctly reproducing the observations of the surface oil slick. Firstly, as
discussed on section 3.4.1, we did not account for the direct impact of wind or Stokes
drift in the transport of oil, which are likely to impact the spread of the surface oil
slick closer to the shore. Another important factor is that the predictability limit of
Lagrangian trajectories in the area affected by the oil spill ranges from about 1.5-6
days (Mariano et al., 2011), which makes the task of forecasting the oil transport for
long periods impractical without updating the location of the oil with observations.
Applications of Lagrangian models to simulate the transport of the oil from the DWH
by Liu et al. (2011b), Macfadyen et al. (2011), Huntley et al. (2011) and Mariano
et al. (2011) highlight the importance of applying observations to correct the location
of the oil in order to accomplish a faithful forecast. Also, we neglected oil weathering
processes, like biodegradation and evaporation, which might decrease the horizontal distribution of oil by thousands of kilometers (Adcroft et al., 2010; Mariano et al., 2011; North et al., 2015). Finally, submesoscale processes which are not resolved by the HYCOM velocity field have been identified as important factors in the dispersion of the DWH oil spill (Poje et al., 2014).

Nevertheless, the capabilities of the PC approach presented can improve significantly the oil fate forecast, and support contingency planning during oil spill events. In order to apply the methodology operationally in a 7 days forecast, the order of the polynomial expansion can be reduced to decrease the computational cost of calculating the expansion coefficients, and provide a faster response. In previous experiments within the same scenario, we were able to satisfactorily reproduce our quantity of interest up to 10 days using a 6th order PC expansion. In this case, the computation of the PC coefficients required only 343 realizations of the oil fate model, against the 1331 used to evaluate the 10th order PC expansions presented. Further, other sources of uncertainty should be considered, such as oil weathering, wind drag, and stokes drift. The last two can be added in the Eulerian velocity field, and be perturbed along with the ocean currents through EOF decomposition. This way, the number of uncertain input parameters wouldn’t change, and no extra realizations of the model would be necessary. At last, alternative ways to perturb the velocity field are possible, and should be the focus of further investigation. Considering a 6th order PC expansion, and limiting the number of variables used to perturb the velocity field, we can propagate uncertainty on an ensemble of HYCOM realizations, in a similar fashion as Iskandarani et al. (2016a). If we are able to reduce the number of uncertain parameters to two, as we did in the present study, a maximum of 49 realizations of HYCOM would be necessary. This number can be further reduced by the application of adaptive quadrature techniques to compute the PC coefficients (Constantine et al.,
2012; Winokur et al., 2013; Iskandarani et al., 2016b).
Chapter 4

Reconstruction of submesoscale velocity field from surface drifters

4.1 Overview

The Lagrangian Submesoscale Experiment (LASER) was conducted with the arduous and challenging task to observe submesoscale activity in the Northern Gulf of Mexico. Submesoscale processes comprise horizontal spatial scales of $\mathcal{O}(0.1 - 10)$ km and time scales of $\mathcal{O}(1 - 100)$ hours, which are too short to be resolved by satellite altimeters, and extremely hard to observe by ship surveys. In LASER, the task was approached by an extensive drifter deployment, which provided measurements of submesoscales at unprecedented high resolution (D’Asaro et al., 2018; Haza et al., 2018; Novelli et al., 2017). The objective of this chapter is to reconstruct the surface flow field sampled by a subset of those drifters, and to present estimates of its differential kinematic properties, namely horizontal divergence, the vertical component of relative vorticity and lateral strain rate, and dominant space-time scales. One specific goal here is to identify submesoscale processes in an Eulerian frame, providing not only statistics about the flow field, but also a visualization of its evolution in time.
Submesoscale features have been known to oceanographers for some time. Spiral eddies with length scales of 0(1 – 10) km were first observed in the sun glitter of photographs taken from space by the Apollo Mission in the late 1960s. These spiral eddies were later one of the targets of the space shuttle Challenger Mission 41-G in 1984, whose sun glitter images revealed a large occurrence of these features embedded in complex fields of shear currents (Scully-Power, 1986; Munk et al., 2000). Although those images suggested a potential relevance of the spiral eddies to the ocean circulation, they only provided brief snapshots of these processes and did not yield information on the underlying dynamics (Munk et al., 2000). As mentioned above, the relatively small length and time scales impose an observational constraint that are challenging for traditional ocean sampling methods, which contributed to the slow progress in understanding the submesoscale dynamics in the 1980s and 1990s (McWilliams, 2016).

The increased computer power over the last decade enabled theoretical studies about submesoscale turbulence based on numerical models, and their results have drawn attention to the potential role of submesoscales on the ocean energy cascade, mixing and material transport. The flow regime at these scales is only partially constrained by geostrophy (0(1) Rossby number) and hydrostatic balance, and represents a transition zone from the quasi-two dimensional mesoscale flows to smaller scale and more three dimensional flows (Thomas and Ferrari, 2008; McWilliams, 2016). Studies have shown that submesoscale frontal instabilities extract energy from the mesoscale strain field and transfer it to smaller scales (Capet et al., 2008). By providing a forward energy cascade route to dissipation, the submesoscales might play a key role in solving the dilemma of how energy is dissipated from geostrophically balanced flows (Molemaker et al., 2010), as geostrophic turbulence theory predicts an inverse kinetic energy cascade from meso to large scales, and an enstrophy cascade from
meso to smaller scales (Charney, 1971). Also, an active submesoscale field has direct consequences for oceanic transport and mixing as it controls dispersion on local scales (Özgökmen et al., 2012), in contrast to non-local dispersion under an enstrophy cascade regime (Bennett, 1984).

Observations of submesoscales are still scarce, but their growth in recent years have evidenced their importance to ocean dynamics and material transport. In Flament et al. (1985), a succession of satellite infrared images of sea surface temperature (SST) off the California shelf presented the evolution of a cold filament under frontal instability on its cyclonic side, and hydrographic measurements showed the presence of frontal subduction. Analyzing a similar feature, Flament and Armi (2000) were able to estimate velocity gradients from clusters of drifters and ship drift over a cyclonic front sharper than 1km. The authors found horizontal velocity shears larger than \(7.5f\) (\(f\) is the planetary vorticity), and horizontal convergence larger than \(0.8f\).

D’Asaro et al. (2011) used a neutrally buoyant Lagrangian float and a towed Triaxus profiling vehicle to take measurements at a frontal zone in the Kuroshio extension. The authors identified an enhanced turbulent activity at a sharp front, marked with vigorous vertical displacements of the float, and verified that the enhanced turbulence was extracting energy from the front itself by symmetric instability. Poje et al. (2014) used two-particle statistics from about 300 surface drifters to study the structure of velocity fluctuations from 200-m to 50-km scales. Their results show that particle dispersion was controlled by local scales in the submesoscale range, indicating the presence of an energetic submesoscale flow field, and that the underlying Eulerian kinetic energy spectrum at these scales was shallower than the one expected in an enstrophy cascade regime. Some recent studies have identified submesoscale activity by estimates of differential kinematic properties from the velocity field, as velocity
gradients with $O(f)$ are expected at these scales\(^1\). In Shcherbina et al. (2013), ADCP velocity measurements from two vessels running on parallel tracks were used to estimate the full gradient tensor of the horizontal velocity field in the North Atlantic Mode Water region. Their results show the presence of strong cyclonic vorticity, up to $3f$, over a mainly weak anticyclonic background. In Berta et al. (2016) and Ohlmann et al. (2017), estimates of flow kinematics at multiple scales were computed from clusters of drifters. In both studies, convergence around $O(1 - 10)f$ were found in submesoscale fronts. The latter study also estimated relative vorticity larger than $5f$ over submesoscale eddies. Rascle et al. (2017) used Sun glitter images taken from airplane to observe fine-scale sea surface roughness, and X-band radar to estimate the velocity gradients over submesoscale fronts, with both data sets collected during LASER. Velocity estimates from the X-band radar in a 500 m resolution grid showed the presence of an intense front, characterized by across-front convergence and along-front shear, and with cyclonic vorticity reaching up to $5f$. In the sea surface roughness anomaly estimated from the sun glitter images of the same area, they measured a 50 m wide front, and the sharpness of the roughness anomaly indicated velocity gradients on the order of $80f$.

The reconstruction of Eulerian velocity field from Lagrangian observations have been accomplished through a wide range of techniques for various purposes. For studies of large scale circulation, Lagrangian observations are usually grouped inside spatial bins, where the mean and eddy components of the flow are estimated. Applications of this technique include studies of the circulation in the California current system (Brink et al., 2000), in the Adriatic Sea (Poulain, 2001), in the Nordic Sea (Jakobsen et al., 2003) and in the North Atlantic basin (Fratantoni, 2001). An improvement of this approach was presented by Lumpkin (2003) in order to avoid

\(^1\)For large and mesoscales, velocity gradients much smaller than $f$ are expected.
contamination of the estimated mean circulation by seasonal variability. The author treated the observations inside each spatial bin as a time series composed of a mean, harmonic components (semiannual and annual) and a residual term, which were estimated via Gauss-Markov theorem. Variations of this technique were applied by Lumpkin and Garzoli (2005) to retrieve mean circulation along with annual and semiannual variability in the tropical Atlantic, and by Lumpkin and Johnson (2013) to infer a climatology of the global circulation, with estimates of semiannual to inter-annual variability. Recently, Laurindo et al. (2017) proposed a method where one-dimensional polynomials were used to describe horizontal variations of the mean flow within spatial bins, enabling an improved resolution of the large scale circulation. This method was applied by Mariano et al. (2016) to retrieve the mean flow over a month from surface drifters in the Gulf of Mexico, and they were able to reconstruct mesoscale features in the Loop Current system.

A more "instantaneous picture" of the ocean circulation is the subject of data blending and data assimilation methods, which are applicable to operational forecasts and emergency responses. Both consist of improving a prior estimate of the Eulerian velocity field with Lagrangian observations (Griffa et al., 2013), with the former being used with off-line OGCM outputs (Taillandier et al., 2006a; Chang et al., 2011), HF-radar data (Berta et al., 2014) and altimetry data (Berta et al., 2015), and the latter being applied to correct model circulation (Taillandier et al., 2006b; Carrier et al., 2014; Muscarella et al., 2015). Several techniques to assimilate Lagrangian data have been put forward, including optimal interpolation (Molcard, 2003; Ö zgökmen, 2003), extended kalman filter (Kuznetsov et al., 2003) and variational methods (Taillandier et al., 2006a; Carrier et al., 2014). These methods can be separated into pseudo-Lagrangian, when the velocities along drifters trajectories are used as observations, and strictly Lagrangian, when the positions are assimilated through an observational
operator, and the prior velocity field is corrected by minimizing the distance between simulated and observed trajectories (Molcard, 2003; Chang et al., 2011). An idealized study by Özgökmen (2003) showed that the latter class outperforms the former when the sampling interval ($\Delta t$) lays between one fifth and one half of the Lagrangian correlation time scale ($T_L$), but that both present equivalent results when $\Delta t \ll T_L$.

As the goal here is to characterize the fast evolving submesoscales sampled by a large array of drifters, we opted to apply a methodology which enables the reconstruction of instant snapshots of the Eulerian velocity field, while taking advantage of the velocity correlations in space and time. Moreover, as the sampling frequency of our observations ($\Delta t = 15$ min.) is much smaller than the Lagrangian correlation time scale at the sampling site ($T_L \simeq 2 - 3$ days in the Gulf of Mexico (LaCasce and Ohlmann, 2003)), we opted to take a pseudo-Lagrangian regression approach. Its philosophy is based on using recent supervised machine-learning techniques to infer information about the Eulerian velocity field from Lagrangian data. This information includes estimates of the velocity field at unsampled points, querying the observations for the dominant length and time scales experienced by the drifters and inferring the derived quantities. All information is gleamed from the drifter data, without recourse to outside sources.

In this chapter, velocity estimates from 320 drifters are used to reconstruct the surface flow field inside a frontal cyclonic eddy during a 12 hour period. The results show the presence of strong convergence zones inside the cyclone with characteristics of submesoscale fronts, displaying strong relative vorticity and strain rate. In the following chapter, a methodology is presented to infer the velocity field in the case when the drifters are mostly aligned over a front, while the whole front is being advected by a background flow. The chapter is organized as follows: section 4.2 presents the LASER data, section 4.3 describes submesoscale observations sampled
by the drifters, section 4.4 presents results and discussion, and section 4.5 summarize the main findings.

### 4.2 Data

The data used in this study are estimates of positions and velocities of drifters released during the LAgrangian Submesoscale ExpeRiment (LASER), which was conducted in the northern Gulf of Mexico around the DeSoto Canyon. Two ships and two airplanes collected data under El Nino winter conditions from 18 January 2016 to 15 February 2016. The main objective of this experiment was to sample transient submesoscale features in the outflow of the Mississippi river using hundreds of drifters. The biodegradable surface drifters were developed by CARTHE for LASER (Novelli et al., 2017). They were equipped with a small drogue to sample the upper 0.6 m circulation, and they reported their GPS position every 5 minutes. (For more information on the drifter design, the reader is referred to Novelli et al. (2017).) Before we received the data, the GPS positions had been processed to identify drogue loss, trajectories had been filtered with an acceleration-based filter to eliminate positioning errors, positions had been interpolated into regular 15 minutes intervals, and horizontal components of velocity were estimated with centered finite differences of the drifters displacements (Haza et al., 2018). We used only data from drifters with drogues attached.

### 4.3 Submesoscale observations

The following analysis involved 326 drifters that were deployed on the western side of the DeSoto Canyon, about 100 km from the Mississippi River Delta. The deployment started on 7 February 2016, at 2:00 UTC, and it lasted about 13 hours. At the time
of deployment, this area presented strong density gradients induced by the lighter,
fresher, and colder water from the Mississippi river plume and the heavier, saltier,
and warmer waters from the Gulf of Mexico (D’Asaro et al., 2018). The drifters were
released over a strong cyclonic vortex, which can be seen in the satellite images of sea
surface temperature (SST) and chlorophyll a (fig. 4.1 a and b). The observed SST
from an aircraft survey, taken a few hours before the beginning of the deployments,
presents a more detailed picture (fig. 4.1c), with the presence of colder and warmer
filaments spiraling inside the cyclone. All drifters remained inside the eddy for about
40 hours after the last drifter deployment, and during this period they sampled the
flow field at several scales. The time evolution of the drifter pairs separations his-
togram, displayed in fig. 4.2, shows that a large number of surface velocity samples
were obtained at separation distances of $O(100)$ m to $O(10)$ km from 7 February
2016, 14:15 UTC to 9 February 2016, 14:00 UTC. Most of the drifters remained
within 0.1–15 km from each other for about 40 hours, while all drifters were inside
the cyclonic eddy. After that, a storm disrupted the drifters trajectories, and several
clusters moved away from each other (D’Asaro et al., 2018). This event induced the
noisy behavior of the drifters separations mode after $t = 40$ h (solid line in fig. 4.2).

We focused our attention to the initial 12 hours, when there is a shift in the mode of
pair separations distribution from 3.6 km to 7.8 km. Before this shift, a large number
of drifters appear to be moving in and around a filament in the eastern side of the
cyclonic eddy (fig. 4.3a). The interior of the filament can be identified by the presence
of drifters with very small velocity vectors (or no velocity vectors at all), located
between $x = 7$–10 km and $y = 5$–8 km, surrounded by faster flowing drifters circling
around them in a counter-clockwise direction. To the east of the almost motionless
drifters, other drifters are moving westward, indicating the presence of a convergence
zone. The sharp temperature gradient in the area where these drifters were deployed
Figure 4.1: (A) and (B): Sea surface temperature and chlorophyll a concentration from MODIS/AQUA on 07 February 2016 showing a hammerhead feature where the LASER drifters were released. (C) Aircraft survey of sea surface temperature with the initial drifter positions (white dots). The highlighted area in (A) and (B) indicates the boundaries of (C).

(fig. 4.1c) indicates the presence of a strong frontal feature that could be driving such convergence. After the shift in the mode of pair separations distribution, the filament can not be observed in the drifter velocities any more. Most of the drifters which were floating within the filament are now circling around what appears to be a small eddy, with diameter between 2 and 3 km (fig. 4.3b). The presence of the small eddy can be observed in the pair separation histograms by a secondary mode around 2.5 km (fig. 4.2). The presence of another convergence zone is evidenced in fig. 4.3b by the meridional alignment of drifters between x = 0 - 6 km.

Following we present results of the Eulerian velocity field reconstruction during this 12 hours period. The prior mean \( \bar{u}(p) \) was set to zero, so that all the information about the velocity field is inferred from the regression. The covariance function was
Figure 4.2: Two dimensional histogram of drifter pair separation and time throughout the first 48 hours after the last drifter deployment. The size of the bins are 50 meters in the pair separations axis and 30 minutes in the time axis. Only bins with more than 50 samples are shown. The solid black line indicates the mode of the pair separations distribution in each 30 minutes bin. The vertical dashed line indicates the end of the 12 hours period used for the velocity reconstruction.

defined as a sum of two SE (equation 2.8), with each representing one scale per dimension, so that 9 hyper-parameters have to be determined by the optimization procedure ($\theta = \{(r_{ti}, r_{yi}, r_{xi}, \sigma_i, \sigma_N) | i = 1, 2\}$). As it will be shown, the optimization successfully captured 2 distinct scales per dimension: a dominant, larger one, and a smaller scale, with lower variance. In tests considering a sum of three SE (not shown), the optimization did not result in 3 distinct scales. The larger scale was basically the same as with the two SE, while the signal of the smaller scale was split in two, having similar values of $r_{ti}$, $r_{yi}$, $r_{xi}$ and $\sigma_i^2$. 
4.4 Results and discussion

4.4.1 Optimization of the hyper-parameters

We start our analysis by optimizing the covariance function to find the values of $\sigma_1^2$, $r_{t1}$, $r_{y1}$, $r_{x1}$, $\sigma_2^2$, $r_{t2}$, $r_{y2}$, $r_{x2}$ and $\sigma_N^2$ for each velocity component. A total of 16,343 data points were selected in the 12 hours time window from 7 February 2016, 14:15 UTC to 8 February 2016, 2:15 UTC, and the drifter positions $(x, y)$ are defined in Cartesian coordinate system with origin at $28^\circ N$ and $88^\circ W$. Table 4.1 shows the optimized hyper-parameters of the covariance function for the two velocity components. The optimized noise level for $u$ and $v$ was $\sigma_N \approx 0.009$ m/s, and both components presented a dominant larger scale, with signal standard deviation $\sigma_1 \approx 0.1$ m/s, and a smaller scale with signal standard deviation $\sigma_2 \approx 0.03$ m/s. In both velocity components, the larger scale presented a correlation time $r_{t1} \approx 3$ hours, with $u$ having a larger correlation scale in the zonal direction ($r_{x1} = 2.5$ km and $r_{y1} = 1.8$ km), and $v$ presenting larger correlation in the meridional direction ($r_{x1} = 1.4$ km and $r_{y1} = 3.5$ km).
The smaller scale presented a correlation time \( r_{t2} \approx 40 \) minutes, and length scales around 400 meters.

**Table 4.1:** Optimized hyper-parameters of the covariance function.

<table>
<thead>
<tr>
<th>H.P.</th>
<th>u</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_N(m/s) )</td>
<td>0.009</td>
<td>0.01</td>
</tr>
<tr>
<td>( \sigma_1(m/s) )</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>( r_{t1}(hours) )</td>
<td>2.99</td>
<td>3.24</td>
</tr>
<tr>
<td>( r_{y1}(km) )</td>
<td>1.84</td>
<td>3.54</td>
</tr>
<tr>
<td>( r_{x1}(km) )</td>
<td>2.58</td>
<td>1.42</td>
</tr>
<tr>
<td>( \sigma_2(m/s) )</td>
<td>0.028</td>
<td>0.031</td>
</tr>
<tr>
<td>( r_{t2}(hours) )</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td>( r_{y2}(km) )</td>
<td>0.4</td>
<td>0.44</td>
</tr>
<tr>
<td>( r_{x2}(km) )</td>
<td>0.39</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Each dimension of the optimized SE covariance function is plotted in figure 4.4 along with empirically estimated space and time correlations of the drifters velocity components. The correlation of the velocity components as a function of radial distances between observations (blue line in fig. 4.4a and b) was estimated in 200 m spatial bins, and averaged over the 12 hours period used in the analysis. These correlations assume a statistically isotropic and homogeneous flow, but even if that is not the case, they still provide a crude estimate of the dominant scales of motion (Davis, 1985). In general, the spatial part of the SE (solid and dotted red in fig. 4.4a and b) is in good agreement with the empirically estimated spatial correlations, presenting correlation decay of the same order of magnitude. The most striking difference is that the former do not have negative values, which are induced by the cyclonic circulation of the eddy where the drifters were released. The SE can only capture the correlations
in one half of the eddy. Another difference is that the spatial correlation of $v$ appears to be a combination between the broader meridional SE and the narrow zonal SE. This might actually be the case if the flow is anisotropic. The decay in spatial correlations at shorter separations is more abrupt than provided by the SE, particularly for $u$. Figure 4.4c and d present the Lagrangian auto-correlation, which is averaged over all drifters (solid blue), and an estimate of the Eulerian correlation (dotted blue), which was obtained from the zero spatial lag of the space/time lagged correlations. For the latter, spatial bins of 400 meters were considered. The optimized time SE for each velocity component is presented in red. As with the spatial correlations, the negative lagged auto-correlations are caused by the cyclonic flow of the eddy, which is not captured by the SE. In spite of that, the three curves have similar correlation decay, with the SE curve staying between the Lagrangian and Eulerian correlations curves. The fact that all drifters are trapped inside the eddy explains the proximity between the Lagrangian and Eulerian correlation estimates.

To illustrate the effect of the different scales in the regression, figure 4.5 presents a snapshot at $t = 4$ h of the velocity vectors and the divergence field from reconstructions using covariance functions considering only the larger scale (fig. 4.5a), only the smaller scale (fig. 4.5b), and both scales (fig. 4.5c). The two-scale reconstruction shows a cyclonic circulation around a sharp convergence zone. While the larger scale captures the cyclonic circulation, the sharp velocity gradients in the reconstructed velocity field are resolved by the smaller scale.
Figure 4.4: Comparison of correlations estimated from the observations and the optimized SEs used in the covariance function. The SEs were normalized by their respective total variance ($\sigma_1 + \sigma_2 + \sigma_N$, see table 4.1). Right (left) panels contain comparisons for the u (v) component. Upper panels: simultaneous correlation as a function of drifters separations computed using spatial bins of 200 meters (blue), and SE for the zonal (solid red) and meridional (dotted red) directions. Lower panels: Lagrangian and Eulerian lagged auto-correlation (solid and dotted blue, respectively), and SE for the time dimension (red).
Figure 4.5: Horizontal divergence normalized by planetary vorticity and velocity vectors of reconstructed velocity fields in a 50 meters resolution grid (velocity vectors sampled every 15 grid points). (a) and (b): Reconstruction using covariance function with only one scale per dimension for each component, with $\sigma = \sigma_1$, $r_t = r_{t,1}$, $r_y = r_{y,1}$ and $r_x = r_{x,1}$ in (a), $\sigma = \sigma_2$, $r_t = r_{t,2}$, $r_y = r_{y,2}$ and $r_x = r_{x,2}$ in (b) (see table 4.1 for the values of each hyper-parameter). In (c), the covariance function was defined as the sum of the two covariance functions used in (a) and (b).

4.4.2 GPR Validation

To validate the results of the GPR velocity reconstruction, we randomly selected half of the observations to build an estimator for $u$ and $v$ (equation 2.7), while the other half was set as verification data. Using the optimized hyper-parameters from table 4.1, we estimated the velocity components at the space/time locations of the verification data, and used their respective observations to compute the absolute error:

$$Err(v) = |\tilde{v}_i - v_i^d|$$

$$Err(u) = |\tilde{u}_i - u_i^d|,$$

where $Err(u)$ and $Err(v)$ are the absolute errors of the estimates of $u$ and $v$, respectively, $(u_i^d, v_i^d)$ are the withheld velocity observations, and $(\tilde{u}_i, \tilde{v}_i)$ are their corresponding GPR estimates. The normalized cumulative histograms of $Err(v)$ and $Err(u)$ are
presented in figure 4.6a, and the absolute errors normalized by the observed value are shown in figure 4.6b. The reconstruction of both components presented satisfactory results, with more than 90% of the errors staying below 0.02 m/s, and almost 99% below 0.04 m/s. The distribution of $\text{Err}(v)$ had mean of 0.008 m/s, and standard deviation of 0.013 m/s, while the distribution of $\text{Err}(u)$ had mean of 0.007 m/s, and standard deviation of 0.011 m/s. More than 85% of $\text{Err}(v)$ and about 80% of $\text{Err}(u)$ were below 10% of the observed values.

![Figure 4.6](image)

**Figure 4.6:** (a): Cumulative histograms of absolute error of $v$ and $u$ reconstructions at verification points. (b): Cumulative histogram of the absolute error normalized by the absolute observed value. (c) and (d): Difference between error estimates from the posterior covariance ($\text{Err}_Q$) and the actual errors computed from the observations for $v$ and $u$, respectively. A total of 8171 verification points were used in these histograms.

The square-root of the diagonal of the posterior covariance matrix $Q$ (equation 2.7) provides error estimates for the GPR, which we will use to bound our Eulerian velocity reconstructions. We compared these errors estimates to the absolute errors presented before. Figure 4.6c and d present histograms of the difference between the
posterior covariance error estimates ($\text{Err}_Q$) and the absolute errors ($\text{Err}$). Negative values indicate that $\text{Err}_Q$ underestimated the actual error, while for positive values, $\text{Err}_Q$ overestimated $\text{Err}$. On average, $\text{Err}_Q$ overestimated the actual error by 0.01 m/s for both velocity components, indicating that the former is a good metric to bound the velocity reconstructions.

### 4.4.3 Velocity maps and kinematics

Next, we evaluate the evolution of the velocity field sampled by the drifters over the 12 hours period between 7 February 2016 and 14:15 UTC to 8 February 2016 2:15 UTC. We set a grid with 50 meters resolution, and the reconstructions were carried out every 15 minutes. For this step all observations available were used. Figures 4.7, 4.8, 4.9, and 4.10 present the evolution of speed, horizontal divergence ($\delta = \partial u/\partial x + \partial v/\partial y$), relative vorticity ($\zeta = \partial v/\partial x - \partial u/\partial y$) and lateral strain rate ($\alpha = \left[(\partial u/\partial x - \partial v/\partial y)^2 + (\partial v/\partial x + \partial u/\partial y)^2\right]^{1/2}$), respectively. Only grid points with $\text{Err}_Q(u)$ and $\text{Err}_Q(v)$ smaller than 0.04 m/s are shown, and the differential kinematic properties were normalized by the planetary vorticity ($f$). These figures also show vectors of the reconstructed velocity, along with vectors of the drifters velocities.

The velocity vectors present a large cyclonic pattern, with radius on the order of 10 km, which is consistent with the feature observed in the satellite and aircraft images (fig. 4.1). The filament illustrated in figure 4.3a, appears in the maps of speed as a low speed zone elongated in the $SSW - NNE$ direction between $x = 4$–12 km, and $y = 2$–12 km, which gets narrower from $t = 2$ h to $t = 8$ h. The differential kinematic properties on the eastern side of this feature present characteristics of a submesoscale front (Munk et al., 2000), with intense horizontal convergence, reaching up to $8f$, and strong cyclonic vorticity, with values up to $13f$. As the low speed
area gets narrower, the strain rate and relative vorticity under the convergence zone intensifies, indicating the occurrence of strain driven frontogenesis (Mahadevan and Tandon, 2006). At $t = 10$ h, the convergence zone is broken as the drifters in the northern edge distance themselves from the ones in the south, and the northern edge has turned into the small eddy described in section 4.3 (fig. 4.3b). In the western side

**Figure 4.7:** Snapshots of the reconstructed velocity field (black vectors) sampled every 15 grid points superimposed over speed ($\text{m s}^{-1}$). The time of each snapshot is referenced on 7 February 2016 at 14:15 UTC. The magenta vectors are the drifters velocities at the time of each snapshot.
of the larger cyclone, between $x = -4-4$ km, other frontal features can be observed throughout the 12 hours period, with enhanced convergence, cyclonic vorticity and strain rate.

**Figure 4.8:** Horizontal divergence of the reconstructed velocity field normalized by planetary vorticity. The time of each snapshot is referenced on 7 February 2016 at 14:15 UTC. The green vectors are the drifters velocities at the time of each snapshot.
Figure 4.9: Vertical relative vorticity of the reconstructed velocity field normalized by planetary vorticity. The time of each snapshot is referenced on 7 February 2016 at 14:15 UTC. The green vectors are the drifters velocities at the time of each snapshot.
Figure 4.10: Lateral strain rate of the reconstructed velocity field normalized by planetary vorticity. The time of each snapshot is referenced on 7 February 2016 at 14:15 UTC. The magenta vectors are the drifters velocities at the time of each snapshot.

The reconstructed differential kinematic quantities document the presence of fronts which exhibit strong positive relative vorticity and strongly negative horizontal divergence on short space-time scales. These features are typical of mixed-layer submesoscale turbulence (McWilliams, 2016), and were captured by the Lagrangian observations as they tend to cluster in and align over convergence zones (D’Asaro et al., 2018). The histograms of relative vorticity and horizontal divergence computed here
over the whole period (fig. 4.11) are similar to the ones reported by Ohlmann et al. (2017) over submesoscale fronts off the California coast, although the methodology used in Ohlmann et al. (2017) was quite different. Our relative vorticity histogram shows a distribution with mean of \(0.83 f\), standard deviation of \(1.57 f\), and skewness of \(1.73\), while the histogram of horizontal divergence has zero mean, standard deviation of \(0.92 f\) and skewness of \(-0.91\). The distribution of lateral strain rate has a mean of \(1.42 f\), standard deviation of \(1.12 f\) and skewness of \(2.2\). Ohlmann et al. (2017) used clusters of four drifters to estimate the velocity gradients, and obtained a positively skewed relative vorticity distribution, with mean of \(0.68 f\) and standard deviation of \(1.13 f\) (the values of skewness were not presented), and a divergence distribution with negative skewness, mean of \(-0.13 f\), and standard deviation of \(0.95 f\). Shcherbina et al. (2013) estimated the differential kinematics from ADCP velocity measurements obtained by two vessels sailing in parallel tracks; the two vessels covered a 500 km area of intense submesoscale turbulence in the North Atlantic Mode Water region. They found a positively skewed relative vorticity distribution throughout the mixed layer (skewness = 2.5 between 0-50 meters), with a slightly negative mean of \(-0.07 f\), and median of \(-0.32 f\). Values of high strain rate matching high cyclonic vorticity estimates indicated the presence of submesoscale fronts, which were embedded in an anticyclonic background, as suggested by the weak negative mean and median. Their distribution of divergence was nearly Gaussian, with zero mean, and skewness slightly negative \((-0.2)\).

The maximum values of velocity gradients reported here are about 3 times larger than those found by Shcherbina et al. (2013), which could be explained by the relatively coarser space resolution of their observations, as the two vessels were 1 km apart. On the other hand, Rascle et al. (2017) estimated velocity gradients up to \(80 f\) over a frontal feature from fine-scale sea surface roughness anomalies. The sea surface
roughness anomalies were obtained from high resolution (0.5–6 m) sun-glitter images from airplane surveys during LASER, and the observations covered the same cyclonic eddy as the drifters trajectories used here, although 4 days later. The space-times scales captured in the present analysis are only a subset of the scales of the true flow; this is due to the finite resolution of the drifter sampling and of the limitation of the current GPR approach. The observed scales are set by the spatial distribution of the drifters, and hence dictated by the flow being sampled, and by the frequency of observations. Furthermore, the current GPR approach uses only a limited number of stationary space-time scales to represent a flow field which might be energetic at several scales of motion and which might be non-stationary. Increasing the grid resolution of our Eulerian velocity field would not change the intensity of the velocity gradients, as these are dictated by the smoothness of a covariance function whose shortest optimal correlation scales were found to be around 400 meters. The front in Rascle et al. (2017) was 50 meters wide as seen from airplane observations.

![Figure 4.11: Normalized histograms of relative vorticity (a), horizontal divergence (b) and lateral strain rate (c), all quantities normalized by \( f \). The histograms were built from velocity reconstructions taken in 15 minutes intervals from 7 February 2016, 14:15 UTC to 8 February 2016, 2:15 UTC. Only grid points with \( \text{Err}_Q(u) \) and \( \text{Err}_Q(v) \) smaller than 0.04 m/s were considered.](image)
4.5 Summary

Gaussian process regression has been applied to obtain high resolution velocity maps from a dense array of drifter observations. Data from 326 drifters with time resolution of 15 minutes were used in a 12 hour window, when all drifters remained inside a cyclonic frontal eddy. The space and time correlation scales from the covariance function, as well as the data noise level, were obtained from the drifter observations themselves via a Bayesian optimization procedure. The covariance function was defined with two scales per dimension, and the optimization procedure resulted in a dominant larger scale, which captured the cyclonic circulation inside the eddy, and a smaller scale that represented the fast evolving sharp velocity gradients inside the eddy. The reconstructed Eulerian velocity field presents the time evolution of strong submesoscale features displaying significant ageostrophic circulation. Frontal features with horizontal convergence reaching $8f$, lateral strain rate up to $10f$ and cyclonic relative vorticity up to $13f$ were observed.

It is important to point out that as all information about the velocity field were provided by the drifters, the sharp velocity gradients only manifested where (and when) there were enough samples. As drifters move away from each other, beyond the correlation scales from the covariance function, the velocity gradients are smoothed out. The empirically estimated covariances in figure 4.4 display features that were not present in the optimized SE covariance functions: the steep correlation decay at short separation scales, and negative correlations at large separations. The SE however yielded acceptable results, as evidenced by the error verification in section 4.4.2, even though no attempt was made to optimize the type of covariance function nor to account for the varying space-time scales of the flow. Model selection algorithms, including combination of different types of covariance function, can be implemented in a similar fashion as the optimization procedure presented for the hyper-parameters
(Rasmussen and Williams, 2006). Likewise, more complex and non-stationary covariance functions (Paciorek and Schervish, 2006) can be used to capture the changes in the dominant dynamical scales of heterogeneous and non-stationary flows. These improvements could provide a better representation of the flow field, and should be the subject of future research.

Nevertheless, our approach enabled us to not only identify submesoscale activity, but also follow its time evolution for 12 hours. This aspect represents an advantage to other approaches used to identify submesoscales, like point estimates of differential kinematic quantities, which do not provide a visualization of the processes, or static images from vessel or aircraft surveys, as these do not provide the time evolution of an observed process.
Chapter 5

Reconstruction of a submesoscale front

5.1 Overview

In the previous chapter, the multi-scale covariance function provided a velocity reconstruction with two distinct scales: the larger one was attributed to a cyclonic eddy with radius between 2-4 km, while the smaller scale captured the intense velocity gradients over frontal features inside the eddy. The smaller scales of $u$ and $v$ were nearly isotropic and in contrast to what would be expected over frontal features, i.e. larger correlations along the front, and smaller correlations across-front. The problem occurs because the flow length scales are non-stationary and non-isotropic and are not represented correctly with a rectilinear coordinate system. Unless all observed fronts are oriented in a single direction, stationary covariance functions will not properly represent the velocity correlations in a Cartesian coordinate system. The elongated features shown in the previous section relied on the large array of observations at short separations. In this chapter, we propose a framework to reconstruct the velocity field over submesoscale fronts applying GPR with drifter observations. Our strategy to
overcome the limitations of the stationary covariance functions is to select clusters of drifters over a given frontal structure, and adjust the coordinate system according to the alignment of drifters. The hope is that GPR can be useful in identifying the dominant along and across frontal scales and recover a more accurate velocity field along the feature.

Figure 5.1: Positions of the drifters used in the analysis of chapter 5. (A) shows the original positions, (B) and (C) show the positions with the center of mass removed at each time step and (D) shows the positions of the red dots in (C) in the new coordinate system. In (A) and (B), the blue/red dots are the initial/final positions. In (C), the red dots are the positions of the observations used in the analysis. The green line is a 2nd order polynomial fit of the red dots, which was defined as the along front direction ($y'$) in the new coordinate system.
For the following analysis we selected a set of drifters used in the previous chapter which were caught up in a spiral eddy 4 days after their deployment. Figure 5.1A presents the drifters positions during a 3 hours period, with the blue and red dots indicating their initial and final positions, respectively. The group of blue dots between $y = 5 - 10$ km and $x = 10 - 12$ km display two lines of drifters merging in a strong convergence zone, forming a ”Y” shape, or as D’Asaro et al. (2018) described it, a fabric zipper-like structure. Three hours later, the two lines have collapsed into one. The focus of the analysis will be on this convergence structure during the three hour period.

### 5.2 Coordinate system adjustment

Figure 5.1 illustrates how the coordinate adjustment is performed. First, the center of mass of the drifters positions is removed at each time step:

$$
x_2(t) = x(t) - x_c(t),
$$

$$
y_2(t) = y(t) - y_c(t),
$$

(5.1)

Where $x_c(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t)$, and $y_c(t) = \frac{1}{N} \sum_{i=1}^{N} y_i(t)$ are the $(x, y)$ coordinates of the center of mass of all drifters around the spiral eddy ($N = 103$ drifters). This step is necessary to avoid interference in the optimization of the covariance function by the background flow. In figure 5.1A, an average eastward drift can be seen in most drifters, which could result in increased correlations in the zonal direction. With the center of mass removed, the drifters trajectories display more clearly the shape of the spiral eddy (fig. 5.1B). Next, we select the drifters in the western side of the eddy, that are lining up in a strong convergence zone (red dots in fig. 5.1C). A second order polynomial fit is used to draw a coordinate axis following the orientation of the
drifters (green line in fig. 5.1C). The new coordinates \((x', y')\) are computed as follows. For a given observation, the shortest distance from its position to the polynomial fit curve gives the "cross-front" direction \((x')\). The length of the polynomial fit from a predetermined origin to its nearest point to the observation is defined as the "along front" direction \((y')\). Figure 5.1D shows the drifters positions in the new curvilinear coordinate system. The drifter velocities are projected into the new coordinates following the angle between the tangent of the polynomial fit at its closest point to each drifter and the \(x\) axis of the Cartesian coordinate system. The velocities are rotated accordingly, so that \(v'\) and \(u'\) are the along-front and cross-front components, respectively.

5.3 Results and discussion

5.3.1 Evaluating the coordinate system adjustment

To evaluate the performance of the velocity reconstructions using the new curvilinear coordinate system, we compared it with reconstructions using two different Cartesian coordinate systems: one with the original coordinates, and the other with the drifters center of mass removed at each time step. A setting where both velocity components share the same covariance function using the curvilinear coordinate system is also tested. Hereafter, the analysis using the curvilinear coordinate system will be labeled \(C_1\) (\(C_{1b}\) for the case where both velocity components have the same covariance function), while the analysis using the Cartesian coordinates with the drifters center of mass removed and with the original coordinates will be referred as \(C_2\), and \(C_3\), respectively.

For \(C_1\), \(C_{1b}\) and \(C_2\), the prior mean is defined as the velocity of the drifters center of mass, so that it is constant in space but varies in time. For \(C_3\), the prior mean is
defined as the average velocity of all drifters over the 3 hours period, being constant in space and time. For each coordinate system, 14 combinations of the covariance functions presented in section 2.2.1 are tested. This was done to check whether the results are more sensitive to the choice of covariance function or the coordinate system. For each case, the covariance function was optimized using all observations available. The drifters were then divided into sets of training and test data. The training data were used to build a new Gaussian process regressor with the previously optimized hyper-parameters. With the new regressor, velocity reconstructions at the test data locations were used to compute absolute errors ($Err_{vel}$) as:

$$Err_{vel} = \sqrt{Err(u)^2 + Err(v)^2}$$  \hspace{1cm} (5.2)

where $Err(u)$ and $Err(v)$ are computed from eq. 4.1. Three different settings of training/test data were used. In the first one, about half of the observations was randomly chosen as training data, while the other half was used for error verification (fig. 5.2A). In the second setting, 1/12 of the observations was randomly selected as training data, while the rest was used as test data (fig. 5.2B). The last setting was implemented to check the ability of each GPR configuration in extrapolating the results in space (fig. 5.2C). A cluster of 492 observations situated upstream of the frontal feature was used for training, while the remaining 216 observations downstream were used as test data.

As the distributions of $Err_{vel}$ for each configuration were highly skewed towards large errors, the median is presented as a measure of their central tendency (figure 5.3), and box plots are used to illustrate their spread (figures 5.4, 5.5 and 5.6). For the first setting (fig. 5.3A and fig. 5.4), in which about half of the observations was used as training data set, the errors within all configurations were small, with $Err_{vel}$ ranging between 0 – 0.05 m/s, and the distribution medians around 0.01 m/s. The differences
between covariance functions are subtle. The curvilinear coordinate systems $C_1$ and $C_{16}$ slightly outperformed the Cartesian coordinates with every covariance function tested, presenting lower medians and smaller spread.

**Figure 5.2:** Three different settings of training/test data in the original Cartesian coordinate system.

In the second setting (fig. 5.3B and fig. 5.5), where only $1/12$ of the observations was used as training data set, the average distance between test and training points is larger than in the previous setting. With this, the velocity reconstructions depend more heavily on the configuration of the covariance function and on the choice of coordinate system. The most noticeable difference compared to the previous setting is that the covariance function defined by a single SE clearly under performed all other covariance functions with all coordinate systems. Again, $C_1$ and $C_{16}$ presented smaller errors than $C_2$ and $C_3$, and the differences between them have increased compared to the previous setting. The error distributions of $C_3$ presented the largest values with almost every covariance function tested, and its results were more dependent
on the choice of covariance function than the other coordinate systems. The error distributions of $C_1$ and $C_{1b}$ were very close to each other, with the exception of the configurations with SE and SE+M32, in which $C_1$ presented considerably larger errors (O(0.001-0.01) m/s). For both $C_1$ and $C_{1b}$, SE, M32 and M52 provided better results when the covariance function was composed as a sum of two functions (with exception of $C_1$ with SE+M32).

![Figure 5.3: Medians of the absolute errors of velocity for the reconstructions using the training/test data setting presented in figure 5.2. The covariance function labels are: SE = squared exponential, M32 = Matérn 3/2, M52 = Matérn 5/2 and RQ = rational quadratic.](image)
Figure 5.4: Box plots of the absolute errors of velocity for the reconstructions using the training/test data setting presented in figure 5.2A. The horizontal black line inside each box is the median, and the whiskers extend to 1.5 of the inter-quartile range. The outliers were omitted for better visualization.

Figure 5.5: Same as figure 5.4 for the training/test data setting presented in figure 5.2B.
The last training/test data setting is the most challenging one, as the regressors have to extrapolate the observations from the training points to the test points. As expected, the values of $Err_{vel}$ are considerably larger than the ones computed before, reaching values one to two orders of magnitude higher than the largest errors in the previous two settings. The median (fig. 5.3C) and box plots of $Err_{vel}$ (fig. 5.6) show that the curvilinear coordinate system from $C_1$ and $C_{1b}$ provided significantly smaller errors, with narrower distributions, than the Cartesian coordinates from $C_2$ and $C_3$ in almost every covariance function configuration. The exceptions were SE, which provided smaller errors with $C_3$, although in a broader distribution, and M52, which presented smaller errors with $C_2$. The best overall results were achieved by $C_1$ and $C_{1b}$ with the RQ covariance function, given that it is the only configuration where the whole box plot is below 0.1 m/s. Other configurations that also performed well compared to the others were $C_1$ with the functions M32, SE+SE, SE+M32, and SE+M52, and $C_{1b}$ with SE+M52, SE+RQ and M32+RQ.

As we intended, the along front axis of the curvilinear coordinate system provided a more efficient way to carry the information from the training points to the test points. This fact is illustrated in Figure 5.7, which shows velocity vectors estimated with $C_1$ and $C_2$ using RQ, along with the observed velocity. In the first time step of the three hours period considered (fig. 5.7A), most of the $C_1$ velocities (blue vectors) are superimposed on the test points (black vectors), while the $C_2$ velocities (green) diverge from the latter as the distance from the training points (red) increase. The same occurs in the last time step (fig. 5.7B), but this time, the velocity vectors from $C_1$ also diverge from the observations in the southernmost test points. This fact probably occurred because, at this time step, the along front axis (magenta line) is not parallel to the orientation of the drifters. The difference between the alignment of the drifters and the curvilinear axis indicates that the frontal feature is rotating
following the cyclonic circulation of the eddy where it is embedded. This rotation is not treated in our coordinate system adjustment, which induces the larger errors of $C_1$ in the southernmost test points.

![Figure 5.6](image)

**Figure 5.6:** Same as figure 5.4 for the training/test data setting presented in figure 5.2C.

![Figure 5.7](image)

**Figure 5.7:** Observations used as training (red vectors) and test (black vectors) data, and velocity reconstructions at the test data locations using $C_1$ (blue) and $C_2$ (green) with the rational quadratic covariance function. The training/test setting is the same presented in figure 5.2C. The magenta line indicates the location of the curvilinear along front axis. The velocity snapshots were taken at the first (A) and last (B) time steps of the three hours period consider in the analysis.
5.3.1.1 Considerations about the covariance functions

In the previous error analysis, the curvilinear coordinate system presented better results with covariance functions that considered multiple scales per dimension. The best overall performance was obtained by the single RQ, which can be regarded as an infinite sum of SE functions with different correlation scales, despite being defined with only one scale per dimension.

Table 5.1 presents the optimized hyper-parameters of 8 covariance functions tested with $C_{lb}$: one of each type ($SE$, $M32$, $M52$ and $RQ$, with a single length scale), and a sum of identical covariance function of different length scales (i.e. $SE + SE$). Figure 5.8 shows the correlation decay of these covariance functions. As we anticipated from the design of the coordinate system, the along front correlation scales are larger than the cross-front ones. For all of the composed covariance functions ($SE + SE$, $M32 + M32$, $M52 + M52$ and $RQ + RQ$), the optimization procedure resulted in a separation between larger and smaller scales. The smaller scales presented values between 260 m and 1.5 km in the cross-front direction, and values between 660 m and 3.9 km in the along front direction, and can be associated to the sharp velocity gradients over the frontal feature. The larger scales, with the exception of $SE + SE$, presented time and spatial scales that are larger than the sampled domain. For example, $M32 + M32$ had a time scale of 20 hours and an along front scale of 23 km. These are not necessarily related to the scales of the flow field. What they actually mean is that the covariance dependence on these dimensions is small (section 5.1 of Rasmussen and Williams (2006)). On the other hand, they provide a slow correlation decay at large separation scales, and, as it will be shown in section 5.3.2, they allow a smooth transition from the submesoscale front regime to the the background circulation away from the observations. The time and length scales of the single covariance functions were in general $3/2$ to 2 times larger than the smaller scales.
from the composed functions. With the absence of a larger scale, it appears that the optimization procedure tries to balance between resolving the sharp correlation decay of the frontal feature with the smoother transitions at larger separations inside the cyclonic eddy. The exception is the $RQ$, which has the smallest length scales among all of the covariance functions. The small length scales allow for a sharp drop in correlation at short distances, which can be associated with the velocity gradients at the front, while the small $\alpha$ (0.05) provides a smooth transition to larger separations.

**Table 5.1:** Optimized hyper-parameters of selected covariance functions with $C_{1b}$. All covariance functions presented an observation noise of 0.007 m/s.

<table>
<thead>
<tr>
<th>Function</th>
<th>$\sigma$ (m/s)</th>
<th>$r_t$ (hours)</th>
<th>$r_y$ (km)</th>
<th>$r_x$ (km)</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td>0.14</td>
<td>1.15</td>
<td>0.84</td>
<td>0.35</td>
<td>-</td>
</tr>
<tr>
<td>SE</td>
<td>0.2</td>
<td>2.2</td>
<td>3.13</td>
<td>0.6</td>
<td>-</td>
</tr>
<tr>
<td>+ SE</td>
<td>0.03</td>
<td>0.59</td>
<td>0.38</td>
<td>0.22</td>
<td>-</td>
</tr>
<tr>
<td>M32</td>
<td>0.25</td>
<td>5</td>
<td>3.9</td>
<td>1.47</td>
<td>-</td>
</tr>
<tr>
<td>M32</td>
<td>0.3</td>
<td>20</td>
<td>23</td>
<td>2.5</td>
<td>-</td>
</tr>
<tr>
<td>+ M32</td>
<td>0.1</td>
<td>2.6</td>
<td>2.0</td>
<td>0.9</td>
<td>-</td>
</tr>
<tr>
<td>M52</td>
<td>0.16</td>
<td>2.2</td>
<td>1.6</td>
<td>0.67</td>
<td>-</td>
</tr>
<tr>
<td>M52</td>
<td>0.24</td>
<td>5.4</td>
<td>11.3</td>
<td>1.2</td>
<td>-</td>
</tr>
<tr>
<td>+ M52</td>
<td>0.06</td>
<td>1.29</td>
<td>0.9</td>
<td>0.45</td>
<td>-</td>
</tr>
<tr>
<td>RQ</td>
<td>0.27</td>
<td>0.87</td>
<td>0.66</td>
<td>0.26</td>
<td>0.05</td>
</tr>
<tr>
<td>RQ</td>
<td>0.02</td>
<td>24</td>
<td>15.4</td>
<td>10.3</td>
<td>1943</td>
</tr>
<tr>
<td>+ RQ</td>
<td>0.26</td>
<td>1.6</td>
<td>1.96</td>
<td>0.41</td>
<td>0.17</td>
</tr>
</tbody>
</table>
Figure 5.8: Selected optimized covariance functions with $C_{1b}$ normalized by their total variance. A, B and C show the cross-front, along front, and time dimension functions, respectively.

5.3.2 Differential kinematic properties

The impacts of the different combinations of coordinate system and covariance function on the differential kinematic quantities are illustrated after reconstructing the velocity field on a 20 meter resolution grid using all available observations. Figure 5.9 presents maps of divergence from reconstructions with the different coordinate systems using RQ, while figure 5.10 displays the divergence obtained using different covariance functions with $C_{1b}$. In all maps, black contours delineate the areas where the estimated error from the posterior covariance (eq. 2.7) is smaller than 0.05 m/s. The maps of relative vorticity and lateral strain rate present similar characteristics as the maps of horizontal divergence, so they will not be presented.

In figure 5.9, the drifters are aligned over a strong convergence zone in all velocity reconstructions, which is expected over a submesoscale front. In all of them, convergences reached values between 12 f and 15 f over the most intense parts, which were also the areas with larger numbers of drifters. On the other hand, the shape
of the convergence zone varied drastically with the type of coordinate system. The maps of $C_1$ and $C_{1b}$ are almost indistinguishable, with a narrow and long convergence zone following the alignment of the drifters. The maps of $C_2$ and $C_3$ present coarser convergence structures, with parts extending away from the line of drifters, as in the area around $x = 10$ km between $y = 4 - 6$ km in the plot for $C_2$, and the area around $x = 11$ km between $y = 8 - 12$ km in the plot for $C_3$. These extensions are caused by large correlations in the covariance matrix which are not aligned with the frontal structure. Another consequence of the latter is the irregular shape of the contour lines of the posterior covariance error from $C_2$ and $C_3$, while the contours from $C_1$ and $C_{1b}$ follow practically parallel to the line of drifters.

Figure 5.10 present snapshots of divergence reconstructed with $C_{1b}$ using four different covariance functions: the one with the worst performance in the error analysis presented in section 5.3.1 (SE), two of the ones with the best results (RQ and M32 + M32), and one with an average performance (SE+SE). In all of this maps, the strong convergence zone follows the curvature of the along front axis in the curvilinear coordinate system. The maps from SE and SE+SE present strong divergence bands parallel to the convergence zone, and are probably due to the sharper correlation decay at large separations in the cross-front dimension (see blue solid and dotted lines in fig. 5.8A). These bands are absent in the maps from RQ and from M32 + M32, which have a much smoother decay (solid green and dotted black in fig. 5.8A). The divergence bands also appear, although much weaker, in the map from M52+M52 (not shown), which correlation decay is not as smooth as for RQ and M32 (dotted red line in fig. 5.8A). The effect of the different scales represented in the covariance functions is illustrated by the reconstructed velocity vectors. The velocity in SE transitions to the prior mean at relatively short distances from the observations, while in RQ and M32 + M32 the velocities far from the observations follow the cyclonic
curvature of the front. The velocities of $SE + SE$ also transition to the prior mean, although at larger distances than those of $SE$. It is important to point out that these covariance functions were not designed specifically for the physical problem at hand, and the resulting optimized correlations scales are not necessarily indicative of the scales of variability in the flow field. The smaller scales are clearly associated to the front variability, e.g. in the RQ case, the cross-front correlation scale was 260 meters, and the width of the convergence zone in the reconstructed velocity was between 600 and 800 meters. On the other hand, the larger scales can not be directly associated to the spiral eddy scales where the front is embedded. As we limited the observations to only a fraction of the eddy, the optimization procedure was not able to properly extract its information from the observations. With that, we should only trust the velocities around the front, which we delineated by the area inside the 0.05 m/s error margin from the posterior covariance.

In the previous section, the curvilinear coordinate system along with the RQ covariance function presented the smallest errors among all training/test data settings. Hence, we chose this combination, with $C_{1b}$ to show the time evolution of the differential kinematic quantities during the 3 hours analyzed. As before, velocity reconstructions were carried out in a grid with 20 meters resolution, and all available observations were used. Figures 5.11, 5.12 and 5.13 present maps of horizontal divergence, relative vorticity and lateral strain rate, respectively. These maps highlight the presence of the sharp front, with horizontal convergence reaching 15 f at some points, while the relative vorticity reaches 17 f, and the lateral strain rate 20 f. A few changes on the front over time can be noted. In the first two snapshots, two lines of drifters can be observed between $y = 6 - 10$ km, which are marked by two lines of convergence. On the other hand, the drifters to the south, between $y = 0 - 4$ km, are not staying over a convergence zone, and are not tightly aligned, with many drifters
Figure 5.9: Horizontal divergence normalized by planetary vorticity at 2 h and 45 min of the 3 hours period, from regressions using the rational quadratic covariance function with $C_1$ (A), $C_{1b}$ (B), $C_2$ (C) and $C_3$ (D). Black vectors are the reconstructed velocities sampled every 40 grid points. Green vectors are the velocity observations used in the regression. The black line delineates the area inside which the estimated error from the posterior covariance is smaller than 0.05 m/s.
Figure 5.10: Horizontal divergence normalized by planetary vorticity at 2 h and 45 min of the 3 hours period, from reconstructions with $C_{1b}$ using as covariance function: SE (A), RQ (B), SE+SE (C) and M32 + M32 (D). Black vectors are the reconstructed velocities sampled every 40 grid points. Green vectors are the velocity observations used in the regression. The black line delineates the area inside which the estimated error from the posterior covariance is smaller than 0.05 m/s.
flowing parallel to each other. In the final two snapshots, the two lines of drifters to the north collapsed into one, and the front at this area displays an intensification in all kinematic quantities. The group of drifters to the south have aligned, and are now also staying over an area with intensified convergence, relative vorticity and strain rate. Beside these changes, the whole front was displaced eastward by the background flow during the period analyzed.

The velocity gradients obtained over the front are significantly stronger than the ones estimated in the previous chapter, with the convergence and strain rate reaching almost double their intensity. The steeper gradients were not caused by the differences in grid spacing, as in both cases the resolution was more than enough to resolve the scales in the covariance function. They were also not caused by the different methodology, as tests with similar configuration as the one used in the previous chapter ($C_3$ with $SE + SE$) provided velocity gradients as intense. It is likely that they reflect frontal features with different intensity. This is evidenced by the fact that in the observations used in the previous chapter, the drifters did not collapse into a line as the ones used in this chapter.

The front described above is the same feature observed by Rascle et al. (2017), which was discussed in the previous chapter. Velocity estimates from X-band radar measurements presented velocity gradients much weaker than ours, with divergence close to 2 f, and relative vorticity close to 5 f. On the other hand, the sea roughness observations, which resolution is 3 to 4 orders of magnitude higher than the X-band radar estimates, pointed to velocity gradients close to 80 f. As mentioned in the previous chapter, the difference between our results and the sea roughness estimates is probably due to limits in the drifter sampling.
Figure 5.11: Snapshots of horizontal divergence normalized by planetary vorticity from the velocity reconstruction using $C_{1b}$ with the rational quadratic covariance function. Black vectors are the reconstructed velocities sampled every 40 grid points and yellow vectors are the velocity observations used in the regression. The black line delineates the area inside which the estimated error from the posterior covariance is smaller than 0.05 m/s.
Figure 5.12: Same as fig. 5.11 for relative vorticity normalized by planetary vorticity.
Figure 5.13: Same as fig. 5.11 for lateral strain rate normalized by planetary vorticity.
5.4 Summary

In this chapter it was presented a GPR methodology to reconstruct the velocity field over a submesoscale front from drifters observations. The approach is based on removing the effect of the background flow on the observations, and on adjusting the coordinate system to the general alignment of the drifters. The premise of the approach is that the drifters align following cross-front convergences, so that their alignment indicate the position of a front. By adjusting the coordinate system to the disposition of the drifters, we expect the covariance function to be representative of the motion over a front.

An extensive evaluation of the adjusted coordinate system showed that this approach significantly improved the velocity estimates in comparison to the Cartesian coordinates. Comparisons between 14 different covariance functions indicated that the rational quadratic was the most appropriate for this problem. In spite of that, the velocity reconstructions were more sensitive to the type of coordinate system than to the covariance function. Maps of differential kinematic quantities presented a narrow and elongated area, following the line of drifters, with strong convergence, relative vorticity and strain rate, as expected under a submesoscale front.

The time evolution of the reconstructed velocity shows the advection of the front by a background flow, indicating that the applied methodology reproduced one multiscale aspect of the flow field. On the other hand, the part of the flow attributed to the eddy was only represented by the curvature of the frontal structure, which provided a general cyclonic circulation pattern for the reconstructed velocity. The composition of the prior mean velocity by both the large and meso scales should be subject of future research. The framework presented here could be applied as the last step in a hierarchical analysis from large to smaller scales. Another venue for future research is the application of data mining techniques to identify clusters of drifters over frontal
features, and to identify the general shape of the feature where a given cluster is embedded in order to implement an appropriate local coordinate system.
Chapter 6

Conclusion and Outlook

The present dissertation was motivated by the challenges of predicting the pathways of material transport in the ocean. Taking this motivation as a guide, two main goals were set. Given that it is impossible to have a perfect forecast for material transport, we tackle the task of obtaining a probabilistic measure of model predictions while accounting for the uncertainties in the model inputs. The second goal aimed to help deciphering the ocean circulation in the submesoscale range, which is considered to have an important role in the transport and dispersion of material. We tackled this problem by providing estimates of the velocity field from Lagrangian observations at these scales. For both problems, frameworks based on uncertain response reconstructions were elaborated.

An UQ framework is presented in chapter 3 to provide statistical assessment for oil spill simulations. The technique is based on propagating uncertainties from input parameters to model outputs using a non-intrusive PC expansion. The framework was tested in a simulation of the first 30 days of the Deepwater Horizon blowout, when a large amount of oil was released at 1500 meters depth. Oil droplet size and ocean currents were selected as uncertain inputs, and a tenth-order polynomial surrogate
for the model was used to sample the uncertain space. Probabilistic hazard maps of oil impact and sensitivity analysis were presented. Information about the real location of the oil slick from composites of satellite images was only available at the end of simulated period. Most of the observed oil slick was within the envelope of the predicted oil spill, although the model did not predict the oil impact west of the Mississippi delta. Lack of wind drag and stokes drift in the forcing fields were probably responsible for the poor prediction over this continental shelf area. Another limiting factor is the Lagrangian predictability, which is about 1.5-6 days in the oil spill area. The 30 days forecast is too uncertain; updating the oil slick position with observations and using shorter forecast periods would reduce this uncertainty. The UQ framework can be expanded to include more uncertain inputs such as unknown parameters in oil weathering parametrizations, and uncertainties in transport by wind drag and wave motion. An increase in the dimensionality of the uncertain input space raises exponentially the number of model samples needed to compute the PC expansion coefficients; the cost increase could be mitigated by resorting to adaptive quadrature techniques.

It is important to point out that the framework presented focuses primarily on quantifying uncertainties caused by model input data such as initial and boundary conditions, and uncertain coefficients. Other types of uncertainty present in model forecasts should also be considered and be the subject of research. One example is model uncertainties, which originate from discretizing the continuous mathematical laws, and include approximation errors and truncation errors. This type of uncertainty in oil transport simulations stem from unresolved scales of motion and is usually addressed with subgrid scale parametrizations. Another type of uncertainty known as structural, or model bias, derives from abstracting real systems with mathematical laws. The random walk subgrid parametrization used here is a source of this type
of uncertainty. Random walk generates a diffusive oil dispersion in the submesoscale range, which is not expected under an energetic submesoscale field as the one observed by Poje et al. (2014) in the vicinity of the Deepwater Horizon site. Another source for this type of uncertainty in our simulations is the shape of the oil droplet size distribution. The Rosin-Rammler function used here was based on previous laboratory and field experiments, although there is no evidence that the oil droplets released from the Deepwater Horizon blowout followed this type of distribution.

Chapters 4 and 5 were devoted to reconstruct the oceanic flow field sampled by surface drifters at scales not typically resolved by numerical models. As mentioned before, an active submesoscale flow field would play a crucial role in oceanic dispersion at short separation scales. Submesoscale processes are also thought to play an important role for the kinetic energy dissipation, as they can extract energy from the mesoscale flow field and transfer it down to smaller scales. Identifying the existence of these processes and characterizing their intensity is important for the development of appropriate subgrid scale parametrization for transport models as the one used in chapter 3, and for general ocean circulation models. The reconstructions of the velocity field from drifter data were carried out with GPR. A Bayesian optimization procedure was applied to estimate the space and time correlation scales in the covariance function, as well as the data noise, from the observations. The velocity reconstructions indicated the presence of vigorous submesoscale features, displaying significant ageostrophic circulation.

In chapter 4, data from 326 drifters were used in a 12 hours window to reconstruct the flow field inside a cyclonic frontal eddy. Given the large number of samples in a wide range of drifter pair separations, and that the behavior of the drifters presented evidence of convergent features inside the eddy, the GPR was set to capture two distinct scales in the flow field. This was done by defining the covariance function with
two scales per dimension. The Bayesian optimization procedure successfully identified two distinct scales; a dominant larger scale with spatial correlations ranging between 1.4 and 3.5 km and time correlation scale of 3 hours, which captured the cyclonic circulation inside the eddy, and a smaller scale, with spatial correlations of 400 m and time correlation of 39 minutes, which captured sharp velocity gradients from submesoscale fronts inside the eddy. The Eulerian velocity reconstructions indeed presented strong velocity gradients, with differential kinematic properties typical of submesoscale fronts. Horizontal convergence reached $8f$ over a few elongated structures, with lateral strain rate reaching up to $10f$ and cyclonic relative vorticity up to $13f$.

One limitation of the above application is that a stationary covariance function, that is expressed in Cartesian coordinates, can not appropriately represent spatial correlations over curved frontal features. As a consequence of this limitation, the smaller spatial correlation scales obtained from the optimization procedure, which represented the velocity gradients over fronts, were isotropic. The velocity correlations over these features should be larger along the frontal axis and smaller in the cross-front direction. A framework to retrieve the velocity field along a curvilinear submesoscale front was presented in chapter 5. One difference between the application in chapter 4 and the one in chapter 5 is that in the former there was a larger spatial coverage of the flow field, while in the latter the drifters aligned over a convergence zone, limiting their spatial coverage, and making it particularly challenging to estimate horizontal convergence. A curvilinear coordinate system following the general alignment of the drifters was defined, so that one dimension accounted for the frontal axis, and the other represented the cross-front direction. An extensive evaluation of the framework showed that it improved consistently the results when compared to reconstructions using Cartesian coordinates. Further, 14 different covariance functions
were tested. The best overall performance was obtained by the rational quadratic, which provided a sharp correlation decay at shorter separation scales in the cross-front direction, as expected in a front. The function then transitions to a smoother decay at larger separations. The reconstructed velocity field presented characteristics of a submesoscale front, displaying a narrow structure marked by strong velocity gradients, with convergence reaching up to $15f$, relative vorticity up to $17f$ and strain rate reaching $20f$.

The results of chapter 5 suggest that, in order to accurately reconstruct the sub-mesoscale velocity field from drifter data using GPR, the observations should be separated according to specific flow regimes (or submesoscale features), and the mapping of each regime should be carried out separately. A topic for future research is to implement data mining techniques to identify clusters of observations sampling specific flow regimes, and to adjust the coordinate system over different submesoscale structures. Moreover, the coordinate system adjustment as presented allows the covariance function to represent the local variability of a frontal feature for as long as the latter keeps its shape and orientation. If the front is deformed and/or rotated by a background flow, the correlation scales of the covariance function cease to be representative of the along and cross frontal directions, which would consequently deteriorate the quality of the velocity reconstruction. One way to overcome this difficulty would be to add a time dependence in the coordinate system itself, allowing it to change shape and orientation over time.

Another interesting topic for future research is the expansion of the framework presented in chapter 5 to reconstruct multi-scale flows. One could implement a hierarchical GPR analysis in a similar fashion as the sequential data assimilation technique from Li et al. (2015). Low passed filtered observations would be used to build a large scale flow field, which would be used as a prior mean for the mesoscale field, and
so on. The targeted GPR for the submesoscale features, with separate clusters of observations for different submesoscale regimes, would be the last layer of velocity reconstruction. This type of analysis could be applied to evaluate the ocean kinetic energy spectrum and its implications to ocean mixing and material transport.
References


Chassignet, E. P., et al., 2009: U.S. GODAE: global ocean prediction with the HYbrid Coordinate Ocean Model (HYCOM). *Oceanography*, **22** (2), 64–75. 21


