Uncertainty Quantification and Assimilation for Efficient Coastal Ocean Forecasting

Dissertation by
Adil Siripatana

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

King Abdullah University of Science and Technology
Thuwal, Kingdom of Saudi Arabia

April, 2019
The dissertation of Adil Siripatana is approved by the examination committee

Committee Chairperson: Ibrahim Hoteit
Committee Co-Chair: Omar Knio
Committee Members: Matthew McCabe, Clint Dawson, Olivier Le Maître
ABSTRACT

Uncertainty Quantification and Assimilation for Efficient Coastal Ocean Forecasting
Adil Siripatana

Bayesian inference is commonly used to quantify and reduce modeling uncertainties in coastal ocean models by computing the posterior probability distribution function (pdf) of some uncertain quantities to be estimated conditioned on available observations. The posterior can be computed either directly, using a Markov Chain Monte Carlo (MCMC) approach, or by sequentially processing the data following a data assimilation (DA) approach. The advantage of data assimilation schemes over MCMC-type methods arises from the ability to algorithmically accommodate a large number of uncertain quantities without a significant increase in the computational requirements. However, only approximate estimates are generally obtained by this approach often due to restricted Gaussian prior and noise assumptions.

This thesis aims to develop, implement and test novel efficient Bayesian inference techniques to quantify and reduce modeling and parameter uncertainties of coastal ocean models. Both state and parameter estimations will be addressed within the framework of a state-of-the-art coastal ocean model, the Advanced Circulation (ADCIRC) model. The first part of the thesis proposes efficient Bayesian inference techniques for uncertainty quantification (UQ) and state-parameters estimation. Based on a realistic framework of observation system simulation experiments (OSSEs), an ensemble Kalman filter (EnKF) is first evaluated against a Polynomial Chaos (PC)-surrogate MCMC method under identical scenarios. After demonstrating the relevance of the EnKF for parameters estimation, an iterative EnKF is introduced and validated for the estimation of a spatially varying Manning’s n coefficients field. Karhunen-Loève (KL) expansion is also tested for dimensionality reduction.
and conditioning of the parameter search space. To further enhance the performance of PC-MCMC for estimating spatially varying parameters, a coordinate transformation of a Gaussian process with parameterized prior covariance function is next incorporated into the Bayesian inference framework to account for the uncertainty in covariance model hyper-parameters. The second part of the thesis focuses on the use of UQ and DA on adaptive mesh models. We developed new approaches combining EnKF and multiresolution analysis, and demonstrated significant reduction in the cost of data assimilation compared to the traditional EnKF implemented on a non-adaptive mesh.
ACKNOWLEDGEMENTS

The work in this dissertation would not have been possible if it were not for the diligent guidance and support by my advisor, Professor Ibrahim Hoteit. Despite each of our regular meeting in the past six years being brief, all were extremely productive thanks to his vast knowledge and mastery in the art of supervising. His comments on my work were always on point, insightful, visionary, and cleverly strategic. Each of them was in its own way a revelation, which directed me, in the most efficient way, to complete my research in a timely manner. I want to thank him for splendidly handling me both on the day of my peak and worst performance. His right amount of constructive criticism and generous encouragement were exactly what I needed in the time of academic crisis. Work aside, It was an honor to have shared with him a superb sushi meal in Kobe, which is among a few social experiences together that I will always remember and cherish.

I wish to thank my co-advisor Professor Omar Knio, who has provided me a solid background on uncertainty quantification early on in my doctoral pursue. Indeed, it has become the main foundation in which this dissertation was built upon. His analogous approach in explaining the subject was such a powerful tool to convey complicated ideas to this younger mind. I have found myself learning from him philosophically as much as scientifically in each of our meetings. I am very grateful for his patience with me on the occasions I was struggling to meet his academic and professional standards.

I would like to thank Dr. Olivier Le Maître for assisting and guiding me during his many visits at KAUST. Despite his very busy schedule, he has always managed to find time for me and went through all mathematical derivations required for my work in minute details. I have learned a great deal by mimicking, adapting and integrating his research technique, perspective, and presentation styles into my own. I feel extremely privileged to
have the opportunity to work alongside and co-author the papers with such a seasoned and brilliant researcher.

Special thanks to Professor Matthew McCabe and Professor Clint Dawson. As the dissertation committee members, they provided important comments and suggestions, which greatly improved the thesis quality. I would also like to thank the rest of the co-authors of my papers, Dr. Talea Mayo, Dr. Ihab Sraj, Dr. Ahmad Kadoura, Dr. Loïc Giraldi and Professor Shuyu Sun for their tremendous cooperations and contributions. Additionally, I offer my sincere thanks to my colleagues, fellow students, and esteemed post-docs at KAUST: Peng, Boujemaa, Naila, Dias, George, Lily, John, Yixin, Samah and Habib for always sharing with me their expertise and laughter, making my Ph.D. journey much more enjoyable.

I would like to mention my father, Dr. Chairat Siripatana, who has inspired me to become a scientist. His life story, how he overcome poverty through his relentless and honest efforts to seek knowledge is nothing short of extraordinary. To complete my Ph.D. and continue my journey in the path of a scholar is the least I can do to honor his legacy. I would like to express the deepest gratitude to my mother, Solihah, for her utmost dedication and sacrifice to provide me plenty in every aspect of my life for the past 29 years. So selflessly, my prosperity has always been her focus and dream. I will continue to be your good son and make you a proud mother. I also thank all my family members, relatives, and friends back in Thailand for being the major driving force behind my pursuit of excellence and success.

Finally, I dedicate this thesis to my wife, Faten. Thank you for being such an ideal partner in every way and for giving me the greatest gift in life, our first daughter, Insha. No words can describe the deep appreciation I have for you. I simply love you Faten.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination Committee Page</td>
<td>2</td>
</tr>
<tr>
<td>Copyright</td>
<td>3</td>
</tr>
<tr>
<td>Abstract</td>
<td>4</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>6</td>
</tr>
<tr>
<td>List of Figures</td>
<td>12</td>
</tr>
<tr>
<td>List of Tables</td>
<td>19</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>20</td>
</tr>
<tr>
<td>1.1 The Importance of Coastal Ocean Modeling</td>
<td>20</td>
</tr>
<tr>
<td>1.2 Uncertainty Quantification and Reduction</td>
<td>21</td>
</tr>
<tr>
<td>1.2.1 Bayesian Inference in Coastal Oceans</td>
<td>21</td>
</tr>
<tr>
<td>1.2.2 Data Assimilation with Adaptive Mesh Models</td>
<td>24</td>
</tr>
<tr>
<td>1.3 Objectives and Contributions</td>
<td>25</td>
</tr>
<tr>
<td>1.4 Thesis Outline</td>
<td>27</td>
</tr>
<tr>
<td>1.4.1 Chapter 2: Assessing an Ensemble Kalman Filter Inference of</td>
<td>27</td>
</tr>
<tr>
<td>Manning’s $n$ coefficient of an Idealized Tidal Inlet against a Polynomial Chaos-Based MCMC</td>
<td>32</td>
</tr>
<tr>
<td>1.4.2 Chapter 3: Ensemble Kalman Filter Inference of Spatially-varying Manning’s $n$ coefficients in the Coastal Ocean</td>
<td>28</td>
</tr>
<tr>
<td>1.4.3 Chapter 4: Bayesian Inference of Spatially-Varying Manning’s $n$ Coefficients in the Coastal Ocean Using a Generalized Karhunen-Loève Expansion and Polynomial Chaos</td>
<td>29</td>
</tr>
<tr>
<td>1.4.4 Chapter 5: Combining Ensemble Kalman Filter and Multiresolution Analysis for Efficient Assimilation into Adaptive Mesh Models</td>
<td>30</td>
</tr>
<tr>
<td><strong>2 Assessing an Ensemble Kalman Filter Inference of Manning’s $n$ coefficient of an Idealized Tidal Inlet against a Polynomial Chaos-Based MCMC</strong></td>
<td>32</td>
</tr>
</tbody>
</table>
2.1 Introduction .......................................................... 32
2.2 Bayesian inference .................................................. 34
   2.2.1 Parameter estimation with MCMC ......................... 35
   2.2.2 Parameters estimation with an EnKF ....................... 35
   2.2.3 Polynomial chaos-based MCMC (PC-MCMC) .............. 39
2.3 Experimental setup ................................................. 42
   2.3.1 The ADvanced CIRCulation Model (ADCIRC) .......... 42
   2.3.2 The model discretization ................................... 43
   2.3.3 Parameterizing a field of Manning’s n coefficient .... 44
   2.3.4 Manning’s n coefficients classification ................. 45
   2.3.5 Observation System Simulation Experiments (OSSEs) ... 46
   2.3.6 Construction of the PC surrogate ......................... 49
2.4 Numerical Results .................................................. 50
   2.4.1 Manning’s n inference from uniform prior ............ 50
   2.4.2 Inference sensitivity to a more challenging Gaussian priors 56
   2.4.3 SEIK tuning .................................................. 63
2.5 Discussion and Conclusions ....................................... 71

3 Ensemble Kalman Filter Inference of Spatially-varying Manning’s n coefficients in the Coastal Ocean 79
   3.1 Introduction ...................................................... 79
   3.2 Problem formulation ............................................. 83
      3.2.1 ADvanced CIRCulation (ADCIRC) model .............. 83
   3.3 Parameter Estimation Framework ............................... 85
      3.3.1 Sequential simulation algorithm ......................... 85
      3.3.2 Karhunen-Loève (KL) expansion ......................... 87
      3.3.3 Joint-SEIK for parameter Inference .................... 89
      3.3.4 Iterative SEIK .............................................. 92
   3.4 Experimental setup .............................................. 93
      3.4.1 Generating synthetic multi-Gaussian Manning’s n fields .93
      3.4.2 Observation System Simulation Experiments (OSSEs) ....94
      3.4.3 KL basis construction .................................... 96
   3.5 Results and discussion ......................................... 96
      3.5.1 SEIK inference of nodally defined Manning’s n values .96
      3.5.2 SEIK inference in KL space ............................... 103
      3.5.3 SEIK-KL sensitivity to inaccurate covariance model ....109
3.5.4 Iterative SEIK (ISEIK) in the full and KL spaces .......................... 111
3.6 Conclusions .................................................................................. 114

4 Bayesian Inference of Spatially-Varying Manning’s $n$ Coefficients in the Coastal Ocean Using a Generalized Karhunen-Loève Expansion and Polynomial Chaos

4.1 Introduction .................................................................................. 118
4.2 Bayesian inference framework ......................................................... 122
  4.2.1 Karhunen-Loève (KL) expansion ............................................ 123
  4.2.2 Polynomial Chaos (PC) expansion ........................................... 125
4.3 Change of coordinates for uncertain covariance functions ............... 127
  4.3.1 Generalized Karhunen-Loève expansion .................................. 128
  4.3.2 PC surrogate with coordinate transformation .......................... 130
  4.3.3 Bayesian inference with the change of coordinates (COC) ........ 132
4.4 Experimental setup ....................................................................... 133
  4.4.1 The advanced circulation model (ADCIRC) ............................ 133
  4.4.2 The model discretization ......................................................... 134
  4.4.3 Prior covariance function and the reference basis ................... 134
  4.4.4 Observation simulation system experiments (OSSEs) ............... 137
4.5 Results and Discussions ................................................................ 140
  4.5.1 Offline step constructions results .......................................... 140
  4.5.2 1st case: inference of a random Manning’s $n$ field drawn from a
         Gaussian process .................................................................... 146
  4.5.3 2nd case: Inference of a 2D parametrized zonal Manning’s $n$ field 153
4.6 Conclusions ............................................................................... 162

5 Combining Ensemble Kalman Filter and Multiresolution Analysis for Efficient Assimilation into Adaptive Mesh Models

5.1 Introduction ............................................................................... 165
5.2 Background ................................................................................ 168
  5.2.1 Multiresolution analysis ....................................................... 168
  5.2.2 Ensemble Kalman filter ......................................................... 172
5.3 MRA-EnKF methods ................................................................... 174
  5.3.1 EnKF update in the MRA settings ....................................... 174
  5.3.2 MRAEnKF ........................................................................... 176
  5.3.3 Forecast Mean Space Projection (FMSP) EnKF ....................... 176
  5.3.4 Analysis Mean Space Projection (AMSP) EnKF ..................... 177
  5.3.5 Correction Projection (CrP) EnKF ........................................ 178
5.3.6 Constant Space Projection (CnSP) EnKF ........................................ 179
5.4 Computational test problem ............................................................... 179
  5.4.1 Model ...................................................................................... 179
  5.4.2 Roe flux calculation ................................................................. 181
  5.4.3 Multiresolution Roe solver ....................................................... 182
5.5 Numerical experiments .................................................................... 184
  5.5.1 Twin experiments ................................................................... 184
  5.5.2 MRA-EnKFs vs. Reference ....................................................... 185
  5.5.3 Assessment of MRA-EnKFs performances ................................. 187
5.6 Conclusions .................................................................................... 197

6 Summary and Concluding Remarks ..................................................... 199
  6.1 Contribution of the Thesis ............................................................... 199
  6.2 Future Research Work .................................................................. 203

References .............................................................................................. 206

Appendices .............................................................................................. 223
LIST OF FIGURES

2.1 Idealized inlet with ebb shoal domain. The discretization of the domain is represented. Observation stations are marked with red dots. The color bar represents the bathymetry of the domain measuring down from the geoid (m). 43

2.2 A schematic representation of one of the true constant Manning's $n$ coefficients ($n = 0.017$) and the different Gaussian priors (with initial guesses from remaining classes as the means) we use in PC-MCMC and SEIK. 48

2.3 Left panels: Distribution of water elevation generated by random Manning's $n$ coefficients at the observed locations at several time steps $t$, as indicated. Different scales are used in each panels to facilitate comparison of the different curves. $p$ denotes the highest polynomial degree in the truncated expansion. In each case, the PC expansion was sampled $10^6$ times to generate the distribution curve. Right panels: PC coefficients at observed locations at several time steps $t$. 51

2.4 Posterior distributions of the constant Manning's $n$ coefficients as estimated by PC-MCMC. The green dashed line represents the true value and the blue dashed line the estimated Manning's $n$ coefficients, as the mode of the posterior pdf (red curve). Note that the estimates do not depend on the choice of initialized value of the MCMC chains if the MCMC chains are sufficiently long. 53

2.5 Posterior distributions of the piece-wise constant Manning's $n$ coefficients as estimated by PC-MCMC. The green dashed line represents the true value and the blue dashed line the estimated Manning's $n$ coefficients, as the mode of the posterior pdf (red curve). Note that the estimates do not depend on the choice of initialized value of the MCMC chains if the MCMC chains are sufficiently long. 54

2.6 Posterior distributions of the constant Manning's $n$ coefficients as estimated by SEIK with 10 ensemble members compared to PC-MCMC. The prior is $\mathcal{U}(0.005, 0.2)$ for both schemes. The green dashed line represents the true value. The black curve is the posterior pdf as estimated by SEIK and the red curve the posterior pdf as estimated by PC-MCMC. 55
2.7 Posterior distributions of the two constants Manning’s $n$ coefficients as estimated by SEIK with 10 ensemble members compared to PC-MCMC. The prior is $\mathcal{U}(0.005, 0.2)$ for $\alpha$ and $\beta$ for both schemes. The green dashed line represents the true value. The black curve is the posterior pdf as estimated by SEIK and the red curve the posterior pdf as estimated by PC-MCMC.

2.8 Posterior distributions of the constant Manning’s $n$ coefficients as estimated by PC-MCMC initialized with Gaussian priors with different means. Black dashed lines represent the true values. Colored dashed lines are the estimated Manning’s $n$ coefficients, or the modes of the posterior distributions. Gray lines and curves are respectively the posterior distributions from uniform prior and their modes. The estimates are not sensitive to the choice of the prior if the MCMC chains are sufficiently long. The spread of the posterior distribution is clearly sensitive to the choice of the prior mean.

2.9 Posterior distributions of the two constants Manning’s $n$ coefficients as estimated by PC-MCMC initialized with Gaussian priors with different means. Black dashed lines represent the true values. Colored dashed lines are the estimated Manning’s $n$ coefficients, or the modes of the posterior distributions. Gray lines and curves are respectively the posterior distributions from uniform prior and their modes. The estimates are not sensitive to the choice of the prior when the MCMC chains are sufficiently long.

2.10 Estimates of various Manning’s $n$ fields parameterized by a single value using 10 ensemble members. From top to bottom, left to right the true Manning’s $n$ value is 0.015, 0.035, 0.06, 0.105 and 0.17. The uncertainty of the estimates are represented by opaque color band (95th percentile) around the mean.

2.11 Estimates of two constants piecewise Manning’s $n$ coefficients, $\alpha$ (left) and $\beta$ (right). The uncertainty of the estimates are represented by opaque color band (95th percentile) around the mean.

2.12 Comparison of the posterior pdfs of the smallest and the largest class of constant Manning’s $n$ coefficients as estimated by SEIK with 10 ensemble members (at the end of the assimilation period) and MCMC.

2.13 Comparison of the posterior pdfs of two constants Manning’s $n$ coefficients as estimated by SEIK (at the end of the assimilation period) and MCMC. Distributions of $\alpha$ are plotted in the left panels and those of $\beta$ in the right panels.
2.14 Comparison of the posterior pdfs as estimated by SEIK for constant Manning’s $n$ coefficient with three different ensemble sizes and initial guess = 0.06. Left panels: the 95$^{th}$ percentile over time. Right panels: the posterior pdfs at the end of the assimilation period. 

2.15 Top row: sensitivity of the posterior distribution as it results from SEIK for the case of piecewise Manning’s $n$ coefficient in the open ocean ($\alpha$) to different ensemble sizes. Left column: the 95$^{th}$ percentile over time, Right panels: the pdf of the posterior at the end of the simulations. Bottom row: comparison of the posterior distribution as it results from SEIK for piecewise Manning’s $n$ coefficient in the landlocked area ($\beta$) with different ensemble sizes. Left panels: the 95$^{th}$ percentile over time, Right panels: the pdf of the posterior at the end of the simulations. All experiments start from the initial guess $\alpha = 0.1$ and $\beta = 0.1$.

2.16 Comparison of the posterior distribution as it results using SEIK for constant Manning’s $n$ coefficient between SEIK without and with inflation factor (1.1, 1.2, 1.5 and 2.0 respectively), initial guess = 0.015. Left panels: the 95$^{th}$ percentile over time, Right panels: the pdf of the posterior at the end of the simulations.

2.17 Comparison of the posterior distribution as it results using SEIK for two constant piecewise Manning’s $n$ coefficients between SEIK with and without inflation in the open ocean area. Left panels: the 95$^{th}$ percentile over time, Right panels: the posterior pdfs at the end of the assimilation period.

2.18 Comparison of the posterior pdfs as it results using SEIK for two constant piecewise Manning’s $n$ coefficients between SEIK with and without inflation in the land-locked area. Left panels: the 95$^{th}$ percentile over time, Right panels: the posterior pdfs at the end of the assimilation period.

2.19 Comparison of the posterior pdfs of classes of constant Manning’s $n$ coefficients as estimated by SEIK with different inflation (at the end of assimilation period) and MCMC.

2.20 Comparison of the posterior pdfs of two constant piecewise Manning’s $n$ coefficients as estimated by SEIK (at the end of the assimilation period) and MCMC. Right panels: the distributions of $\alpha$, Left panels: the distribution of $\beta$. 

67

69

72

73

74

75

76
3.1 Idealized inlet with ebb shoal domain. The discretization of the domain is represented. The first 15 observation stations used in the experiment are marked with red dots, and the additional 9 observation stations added later are marked with white. The color bar represents the bathymetry of the domain measuring down from the geoid (m).

3.2 A few realizations of Manning’s $n$ fields generated by the sequential simulation algorithm. The 101st realization is taken as the reference field that the inference results are compared against.

3.3 The reconstruction of a Manning’s $n$ field using the truncated KL expansion for various retained KL modes. The top six subplots: spatial plots of the reconstructed Manning’s $n$ field. The bottommost subplot: the accumulative sum of the eigenvalues obtained for the KL decomposition.

3.4 The results of SEIK inference with 10 ensemble members and localization, (a) the true field, (b) the initial ensemble mean, (c) the final analysis after 108 updates, and (d) the ratio between the final error and the initial error.

3.5 Time series of the RMSEs of the Manning’s $n$ fields after each analysis step with respect to the true Manning’s $n$ field for varying localization distance (ld).

3.6 Time series of the RMSEs of the Manning’s $n$ field after every analysis step. Top panel: the RMSEs for varying localization distances (ld) with 100 members. Bottom panel: the RMSEs for various ensemble sizes (N) with the same localization.

3.7 The results of SEIK inference with 100 ensemble members, (a) the true field, (b) the initial ensemble mean, (c) the final analysis with 15 observation points, (d) the ratio between the final error and the initial error, (e) the final analysis with 24 observation points and 468 assimilation cycles, and (f) the ratio between the final error and the initial error, (g) the absolute error between the estimate and the truth for 24 observation points case, and (h) standard deviation of the ensembles at the final analysis step.

3.8 Time series of the RMSEs of the Manning’s $n$ fields after each analysis step for 100 ensemble members, different localizations (ld) and different number of observations with respect to the true Manning’s $n$ field.
3.9 Time series of the RMSEs as results from SEIK and SEIK-KL using varying ensemble sizes and numbers of retained KL terms. Left column: each figure represents a fixed ensemble size but varying number of retained KL terms. Right column: each figure represents a fixed number of retained KL terms but varying ensemble size. 108

3.10 Manning’s $n$ field estimates as inferred by different ensemble sizes and number of retained KL modes. 110

3.11 Time series of the RMSEs of the Manning’s $n$ fields inferred from initial ensembles generated from various variogram models. 112

3.12 Inferred Manning’s $n$ fields when various variogram models are used to generate the initial ensembles in KL space. 113

3.13 Inferred Manning’s $n$ fields. 1st row: spatial plots of regular SEIK inference, 2nd row: spatial plots of ISEIK inference with 5 iterations, 3rd row: spatial plots of ISEIK-KL inference with 5 iterations. 115

3.14 Time-evolution of the RMSE of Manning’s $n$ field as inferred by the regular SEIK (in black) and iterative SEIK with various stopping criteria. 116

4.1 Idealized inlet with ebb shoal domain shown with the finite element discretization. Observation stations are marked with white dots. The color bar represents the bathymetry of the domain measuring down from the geoid (m). 135

4.2 The true Manning’s $n$ field of the two test cases. Truths were used to generate the synthetic water elevation observations for the inference. Left: $M^{ran}$, generated from sampling the KL coordinates and $l$ of the squared exponential covariance model. Right: $M^{zone}$ generated by setting $\alpha = 0.005$ and $\beta = 0.1$ in Equation (4.29). 139

4.3 Resulting basis from KL decomposition of $\tilde{C}$ for the first test case, $M^{ran}$. First three panels: the first six eigenmodes (from left to right, top to bottom), bottom: the spectrum of $\tilde{C}$. 141

4.4 Resulting basis from KL decomposition of $\tilde{C}$ for the second test case, $M^{zone}$. First three panels: the first six eigenmodes (from left to right, top to bottom), bottom: the spectrum of $\tilde{C}$. 142

4.5 Approximation of $M^{ran}$, using the reference basis and change of coordinates. Top-left: $M_K$, the true profile, top-right: $M_K^{PC}$, the approximation of $M_K$ in the reference basis, and bottom: $M_K^{PC}$, the approximation of $M_K$ in the reference basis with the PC approximation of the change of coordinates, $\mathbf{B}(\mathbf{q})$. 144
4.6 Time-evolution of the water elevation as produced by the full model (blue) and the surrogate model (red). Each plot shows the water elevation at different observation stations and different quadrature points (i.e., different realization of $\xi$), which characterizes the Manning’s $n$ field.

4.7 Comparison between the priors and marginal posteriors of the 3 KL coordinates $\eta_k$, the length-scale hyper-parameter $l^{\text{norm}}$, and the noise variance hyper-parameter $\sigma^2$ (only the posteriors are shown for $l^{\text{norm}}$ and $\sigma^2$) for the inference of $M^{\text{true}}$ using COC method. The maximum a posteriori probability (MAP) estimate for each parameters are also presented using dashed blue line. The true $l^{\text{norm}}$ used to generate the true field is plotted in dashed green.

4.8 Comparison between the priors, marginal posteriors of the 3 KL coordinates $\eta_k$ and the true KL coordinates of $M^{\text{true}}$ in the $q$-averaged reference subspace ($\xi$). In each plot, the prior is the black curve, the posterior is in blue, the MAP estimate in dashed blue, and the truth in dashed green.

4.9 Spatial maps of several inferred Manning’s $n$ fields compared to the truth:
- (top-left) the true Manning’s $n$ field,
- (top-right) the inferred Manning’s $n$ field using a fixed correlation length scale,
- (bottom-left) the inferred Manning’s $n$ field based on the basis derived from $C$ without using COC, and
- (bottom-right) the inferred Manning’s $n$ field from using PC-MCMC with COC.

4.10 Spatial maps of the percent error with respect to the truth for several inferred Manning’s $n$ fields:
- (top-left) the inferred Manning’s $n$ field using a fixed correlation length-scale,
- (top-right) the inferred Manning’s $n$ field based on the basis derived from $C$ without using COC, and
- (bottom) the inferred Manning’s $n$ field from using PC-MCMC with COC.

4.11 Time-evolution of the water elevation in the final 12 hours of the ADCIRC simulations subjected to different Manning’s $n$ field.

4.12 Comparison between the priors and marginal posteriors of the 4 KL coordinates for the inference of $M^{\text{true}}$ using $q$-averaged covariance basis. The maximum a posteriori probability (MAP) estimate for each parameters are also presented using dashed blue line.
4.13 Comparison between the priors and marginal posteriors of the 4 KL coordinates and the length-scale hyper-parameter $l$ (only the posterior is shown) for the inference of $M_{\text{zone}}$ using COC method. The maximum a posteriori probability (MAP) estimate for each parameters are also presented using dashed blue line. .................................................. 157

4.14 Spatial maps of several inferred Manning’s $n$ fields compared to the truth: (top) the true Manning’s $n$ field, (middle-left) the inferred Manning’s $n$ field based on the basis derived from $\mathcal{C}$ without using COC, and (middle-right) the inferred Manning’s $n$ field from using PC-MCMC with COC. Bottom panel, spatial maps of the percent error with respect to the truth: (bottom-left) the inferred Manning’s $n$ field based on the basis derived from $\mathcal{C}$ without using COC, and (bottom-right) the inferred Manning’s $n$ field from using PC-MCMC with COC. .................................................. 159

4.15 Cross section plot at $y = 0$ (midway between north and south) of the inferred Manning’s $n$ fields compared to the true field $M_{\text{zone}}$ (black). The inferred field based on the basis derived from $\mathcal{C}$ without using COC and with COC are plotted in red and blue, respectively. .................................................. 160

4.16 Time-evolution of the water elevation in the final 12 hours of the ADCIRC simulations subjected to different Manning’s $n$ field. .................................................. 161

5.1 Reference solution vs EnKF solution. .................................................. 185
5.2 Reference vs. MRA-EnKFs (ensemble mean estimates). .................................................. 187
5.3 Normalized L2 error versus assimilation step. .................................................. 189
5.4 RMSE versus assimilation step. .................................................. 189
5.5 RMSE versus assimilation step. For each scheme, curves are generated for different tolerances, as indicated. .................................................. 191
5.6 time-integrated RMSE versus normalized complexity. For each scheme, curves are generated for different ensemble size, as indicated. .................................................. 193
5.7 time-integrated RMSE versus normalized complexity. For each ensemble size, curves are generated for different MRA-EnKF scheme, as indicated. .................................................. 195
5.8 time-integrated RMSE versus tolerance. For each scheme, curves are generated for different ensemble size, as indicated. .................................................. 196
LIST OF TABLES

2.1 Mannings $n$ coefficients in idealized inlet with ebb shoal classified by amplitude of corresponding tidal data at station 1. .......... 45
2.2 The random input parameters for ADCIRC. ................. 47
2.3 RMSEs between the true model and the surrogate model (water elevations) for the whole simulation window at a specific observation station and varying Manning’s $n$ coefficients. ................. 50
2.4 Final estimates, RMSEs between the estimates and the truths and pdfs variance of constant Manning’s $n$ coefficients as estimated by PC-MCMC. ....... 56
2.5 Final estimates, RMSEs between the estimates and the truths and pdfs variance of piecewise Manning’s $n$ coefficients used to parameterize two constants Manning’s field as estimated by PC-MCMC. $n_{\alpha,B}$ represents the field parameterized by $\alpha$ and $\beta$. .......... 58
2.6 The computational complexity of different inference methods. The PC-MCMC and filtering codes are written in matlab. The codes are executed on Intel Xeon(R) CPU X5550 @2.67GHz × 8 processor and the memory of 23.5 GiB. .......... 66
2.7 Final estimates, RMSEs between the estimates and the truths and pdfs variance of constant Manning’s $n$ coefficients as estimated by SEIK with 100 members starting from initial guess = 0.06. .......... 68
2.8 Final estimates, RMSEs between the estimates and the truths and pdfs variance of piecewise Manning’s $n$ coefficients used to parameterize two constants Manning’s field as estimated by SEIK with 100 members. $n_{\alpha,B}$ represents the field parameterized by $\alpha$ and $\beta$. .......... 70
5.1 Summary of EnKF and its multiresolution variants. .......... 180
Chapter 1

Introduction

1.1 The Importance of Coastal Ocean Modeling

Coastal and estuarine systems are home to more than half of the human population \(^\text{[1]}\). Understanding and forecasting the dynamics of these systems is critical for human resources sustenance around the world, both economically and ecologically \(^\text{[2, 3, 4]}\). Coastal inundation during extreme events, such as hurricanes and tsunamis, is a major cause of destruction to human lives and their habitats. Consequently, coastal ocean modeling has been increasingly recognized as a valuable research area \(^\text{[5]}\). The information gained from coastal ocean simulations and forecasts is of great relevance for the authorities to make informed decisions and develop smart and sustainable policies to exploit the ocean resources.

One of the most important applications of coastal ocean modeling is the prediction of storm surge, defined as the abnormal rise of coastal water forced by the wind. Storm surges often induce coastal flash floods, and sometimes act as a pathway for short surface or wind waves to move and break farther inland \(^\text{[6]}\), causing further environmental and infrastructural destruction. A robust water elevation forecasting system based on state-of-the-art coastal ocean modeling would play a major role for decision making by the national and local governments, to formulate efficient and cost-effective measures in order to mitigate the losses caused by the aforementioned devastation. Even in normal conditions, accurately forecasting coastal ocean events (e.g., tides and coastal flows) is important for many marine activities. These include, but are not limited to, fisheries, nearshore construction, tourism, recreational activities, extreme event mitigation, contaminant transport management, and
Aside from the wind forcing, storm surge is also nonlinearly influenced by tides [7]. Accurately predicting the tides is fundamental for assuring the quality of storm surge forecast as well as other coastal ocean phenomena. The performance of the coastal ocean modeling to simulate the tidal currents in a particular region heavily relies on the accurate physical parameterization of the domain of interest. Especially, in the shallow water, the parameters associated with the bottom stress becomes increasingly significant [8]. In the real world, the model parameters represent the physical characteristics of the domain, which are typically heterogeneous and highly variable in space. Furthermore, these parameters cannot be measured directly. One must therefore infer these parameters based on a limited amount of indirect observations. Nevertheless, the parameters estimates will always contain a degree of uncertainty, which can originate from many sources of error [9].

1.2 Uncertainty Quantification and Reduction

Numerical simulations of physical problems are always subjected to uncertainty due to limited knowledge of the systems under consideration [10]. These include the approximation of physics, boundary and initial conditions, model parameters and imperfect discretization schemes [11]. The sensitivity of the model solution to such errors must (generally) be identified, quantified and reduced to improve the model predictions.

1.2.1 Bayesian Inference in Coastal Oceans

Coastal ocean modeling is widely utilized to simulate shallow water circulations for the purposes of conservation, contaminant transport modeling, development of coastal structures (e.g., bridges, damps, and breakwaters), and emergency and economic planning. State-of-the-art coastal ocean models are generally based on the shallow water equations (SWEs), which are derived by depth-integrating the Navier-Stokes equations, assuming hydrostatic pressure and horizontal length scales that are large in comparison to the vertical length
scales \[12\]. Even with the simplifications, such models contain many parameters, carrying uncertainties on which the accuracy of model prediction strongly depends \[1\]. Moreover, many of the parameters cannot be directly measured, or in some situations, the relevant data to estimate these parameters are difficult to collect \[13\]. The values of these parameters must, therefore, be inferred from available data \[4\]. Quantifying and reducing uncertainties in the model outputs associated with uncertainties in the parameters is essential for reliable and robust coastal ocean simulations and predictions (e.g., \[5, 14, 15\]). Of particular importance is the Manning’s \(n\) coefficient of roughness, introduced in the SWEs through the bottom stress components in the momentum equation \[6, 16\].

A general framework for inferring model parameters is the Bayesian inversion approach, \[17, 18, 19, 20\] where the parameters uncertainties are represented with probability distributions (pdfs) conditioned on available data. The parameter inference problem is then viewed as the transformation of a prior pdf to a posterior pdf by incorporating the likelihood of the observations \[21\]. The posterior is rarely explicit and often needs to be sampled as a collection of realizations that are consistent with data and prior information \[22\]. The most popular implementation of this method is the Markov Chain Monte Carlo (MCMC) method (e.g., \[23, 24\]), which has become more practical in recent years with the tremendous increases in computational power. The primary advantage of MCMC is the ability to produce a full approximation of the posterior distribution. As a result, MCMC is often treated as the benchmark to evaluate the performance of other parameter inference methods \[25, 26, 27\]. In order to obtain good resolution of the posterior distribution, a large number of samples is required \[15, 28\]. This makes MCMC computationally very demanding, as each MCMC iteration requires a model evaluation in order to evaluate the likelihood. As a result, using MCMC for parameter estimation is often too costly for a realistic large scale inference problem. Even with model reduction techniques, e.g., Polynomial chaos, KL expansions, etc., parameter estimation in MCMC may still be quite computationally demanding \[29\].

Bayesian inference can also be cast as a filtering problem in which the posterior distri-
bution is updated sequentially as data becomes available [30], an approach known as data assimilation [31, 32]. A Bayesian filter operates as a succession of forecast steps to propagate the pdf of the unknowns forward in time, and analysis steps to update the forecast every time new observations become available. For parameter estimation, filtering schemes usually apply the standard augmented state-parameter technique [33, 34, 35], which allow the state and parameters of the system to be estimated concurrently. Currently, the most popular approach for data assimilation into ocean models is the Ensemble Kalman Filter (EnKF) [36, 30] and its deterministic versions ([37, 34, 38, 39, 40], to cite but a few). An EnKF follows a Monte Carlo framework to integrate an ensemble of model realizations in the forecast step and then applies a linear Kalman correction in the update step [41]. The stochastic EnKF assimilates perturbed observations and this was shown to induce noise in the final solution when the filter is implemented with small ensembles [40, 42]. Deterministic EnKFs, which avoid observations perturbations, mainly differ in the way they sample the new analysis ensemble after the filter update step. Various deterministic EnKFs were compared with a realistic setting of ADvanced CIRCulation (ADCIRC) model in the Gulf of Mexico [42], showing that, with enough tuning, these filters performed closely well, all outperforming the stochastic EnKF. With an adequate straightforward approximation, [40] were lately able to derive a stochastic EnKF-like scheme with deterministic error observations sampling with at least comparable performances to the deterministic EnKFs.

The primary advantage of EnKF-type techniques over MCMC is the algorithmic ability to directly accommodate the estimation of large dimensional state-parameter vectors [43, 44, 45, 46, 47, 48]. Furthermore, these methods are non-intrusive, i.e. their implementation is straightforward and requires no modifications to the model code. Despite their empirical Gaussian framework [49, 50], EnKF methods have been found to be efficient in terms of performance, computational cost, and robustness in handling ocean state estimation problems (e.g., [51, 52, 53, 54, 55]). There is now increasing interest in the coastal ocean community to apply EnKF methods to parameter estimation applications. [5] and
have demonstrated that the EnKFs are able to provide very good estimates of low-dimensional parameterizations of Manning’s $n$ coefficients in the SWEs.

### 1.2.2 Data Assimilation with Adaptive Mesh Models

In hydrological simulations, the loss of solution accuracy due to limited spatial resolution constitutes a major source of uncertainties. The naive approach in reducing this uncertainty is often to employ a fixed spatial mesh with a sufficiently fine resolution for the simulation to capture both large- and small-scale dynamics. However, this approach can be computationally very demanding. To this end, adaptive mesh refinement (AMR) methods have been developed to allow the computational mesh to locally adapt its resolution, according to the features of the model solution \([56, 57, 58, 59]\). By limiting the fine resolution to regions where it is required (or of interest), AMR-based simulations significantly reduce the computational complexity and accordingly enhance the computational performances. Particularly, AMR can be applied to adjust the details of the model solutions according to their shapes, i.e., assigning high mesh resolution to the areas where the solution exhibits large spatial variability, and coarsening the mesh in regions with small spatial variability. This makes AMR particularly attractive due to its ability to capture sharp fronts and shock formations, which are also the natural occurrences observed in the ocean waves \([60, 61]\).

In recent years, the idea of applying data assimilation on adaptively discretized models has increasingly become more practical in the hydrodynamics modeling community \([57, 62, 63, 64, 65]\, [66]\ and \([67]\) applied variational data assimilation methods, i.e. 3D-VAR and 4D-VAR, to adaptive mesh ocean and meteorological models, respectively. Regarding the sequential data assimilation methods such as the EnKF, the difficulties in computing the mean and error covariance required in the analysis step with to the mesh difference for each ensemble member must first be addressed. The recent contribution by \([63]\) demonstrates the first successful implementation of EnKF with multi-level AMR for data assimilation into large-scale coastal ocean models. In their work, the EnKF updates were performed by
projecting each ensemble member to the union of all the meshes computed from all ensemble members. Nevertheless, presently, only very few studies have investigated developing combined AMR-EnKF schemes to address this problem. Therefore, uncertainty reduction in the coastal ocean through the combined data assimilation and adaptive mesh modeling is a subject that remains wide open for further investigation and improvement.

1.3 Objectives and Contributions

This dissertation focuses on studying, quantifying, assessing and reducing the uncertainties associated with coastal ocean models simulations. The main objectives of this work will be to advance the present UQ and DA techniques for Bayesian inference and consolidate both fields under the same framework. Despite our focus being the estimation of spatially varying Manning’s $n$ coefficients of a coastal ocean, the developed inference framework formulated in this thesis could also more generally be used for other types of spatially-varying parameters. Observation System Simulation Experiments (OSSEs) with different modeling settings will be used as the experimental platform to test and validate the proposed methods.

The first part of the dissertation aims is to conduct an in-depth investigation of the Bayesian inference framework for state-parameters estimation in the context of coastal ocean through advanced uncertainty quantification (UQ) and data assimilation (DA) approaches, exploiting the merits and addressing the limitations of each approach, namely, the UQ approach ability to provide full posterior while being subject to the ‘curse of dimensionality’ [68], and DA ability to accommodate large state-parameters vector but with the restricted Gaussian noise assumptions. A comprehensive framework to quantify and reduce the uncertainty of coastal ocean models predictions via state-parameters estimation using UQ and DA techniques, e.g., MCMC, Polynomial chaos, Karhunen-Loève (KL) expansion, and EnKFs, will be formulated. Great efforts will be spent on developing robust and innovative methods to overcome the limitations of these approaches and simultane-
ously improve the accuracy of the inferred solution with respect to the existing conventional
methods. Effective model reduction techniques will be employed and enhanced to speed
up the sampling process of MCMC. The uncertainty of the hyperparameters characterizing
the stochastic process representing the parameter field will be efficiently incorporated into
the inference formulation. Fine-tuning methods for efficient implementation of the EnKFs,
such as localization and inflation, etc., will be applied to establish a suitable framework for
data assimilation in the coastal ocean. Furthermore, we will assess these inference tech-
niques, and against one another, to evaluate their performances and weaknesses, and also to
gain more insights into their inter-relations through statistical and mathematical analysis.

In the second part of this dissertation, a new approach for data assimilation with adap-
tive mesh models will be introduced to improve the computational efficiency of the EnKF
in the context of coastal ocean forecasting. Here, we extend the thesis beyond the scope of
reducing model uncertainties by the mean of parameter estimation but also seek to tackle
the uncertainty that arises from the poorly resolved mesh. We aim at combining the EnKF
with Multiresolution analysis to build cost-effective coarsening and projection strategies
for the ensemble members, allowing each ensemble member to adapt its grid according
to its variability, and as such improving the numerical approximation of the model solu-
tion and also greatly reducing the computational complexity. This approach could also be
presented as a data-driven strategy from adaptive mesh refinement in the coastal ocean.

The improvements in state-parameter estimation with MCMC and EnKF and the de-
velopment of MRA-EnKF methods for adaptive mesh models serve as the foundation of
our contributions to the advancement of the framework of UQ and DA for coastal ocean
forecast models.
1.4 Thesis Outline

1.4.1 Chapter 2: Assessing an Ensemble Kalman Filter Inference of Manning’s $n$ coefficient of an Idealized Tidal Inlet against a Polynomial Chaos-Based MCMC

This chapter demonstrates the relevance of the EnKF for parameter uncertainty quantification in a coastal ocean by comparing the EnKF inference against the gold-standard MCMC. This study was published as an article in *Ocean Dynamics* (doi: 10.1007/s10236-017-1074-z).

Bayesian estimation/inversion is commonly used to quantify and reduce modeling uncertainties in coastal ocean model, especially in the framework of parameter estimation. Based on Bayes rule, the posterior probability distribution function (pdf) of the estimated quantities is obtained conditioned on available data. It can be computed either directly, using a Markov Chain Monte Carlo (MCMC) approach, or by sequentially processing the data following a data assimilation approach, which is heavily exploited in large dimensional state estimation problems. The advantage of data assimilation schemes over MCMC-type methods arises from the ability to algorithmically accommodate a large number of uncertain quantities without a significant increase in the computational requirements. However, only approximate estimates are generally obtained by this approach due to the restricted Gaussian prior and noise assumptions that are generally imposed in these methods. This contribution aims at evaluating the effectiveness of utilizing an ensemble Kalman-based data assimilation method for parameter estimation of a coastal ocean model against an MCMC Polynomial Chaos (PC)-based scheme. We focus on quantifying the uncertainties of a coastal ocean ADCIRC model with respect to the Manning’s $n$ coefficients.

Based on a realistic framework of observation system simulation experiments (OSSEs), we apply an ensemble Kalman filter and the MCMC method employing a surrogate of ADCIRC constructed by a non-intrusive PC expansion for evaluating the likelihood and test
both approaches under identical scenarios. We study the sensitivity of the estimated posteriors with respect to the parameters of the inference methods, including ensemble size, inflation factor, and PC order. A full analysis of both methods, in the context of coastal ocean model, suggests that an ensemble Kalman filter with appropriate ensemble size and well-tuned inflation provides reliable mean estimates and uncertainties of Manning’s $n$ coefficients compared to the full posterior distributions inferred by MCMC.

### 1.4.2 Chapter 3: Ensemble Kalman Filter Inference of Spatially-varying Manning’s $n$ coefficients in the Coastal Ocean

This chapter presents an efficient approach to apply the Joint-EnKF for inferring high dimensional spatially-varying parameter, both in the full physical space and in the enhanced KL space. This chapter was published as an article in *Journal of Hydrology* (doi: [10.1016/j.jhydrol.2018.05.021]).

Ensemble Kalman (EnKF) filtering is an established framework for large scale state estimation problems. EnKFs can also be used for state-parameter estimation, using the so-called “Joint-EnKF” approach. The idea is simply to augment the state vector with the parameters to be estimated and assign invariant dynamics for the time evolution of the parameters. In this contribution, we investigate the efficiency of the Joint-EnKF for estimating spatially-varying Manning’s $n$ coefficients used to define the bottom roughness in the Shallow Water Equations (SWEs) of a coastal ocean model.

Observation System Simulation Experiments (OSSEs) are conducted using the Advanced CIRCulation (ADCIRC) model, which solves a modified form of the Shallow Water Equations. A deterministic EnKF, the Singular Evolutive Interpolated Kalman (SEIK) filter, is used to estimate a vector of Manning’s $n$ coefficients defined at the model nodal points by assimilating synthetic water elevation data. It is found that with reasonable ensemble size ($O(10)$), the filter’s estimate converges to the reference Manning’s field. To enhance performance, we have further reduced the dimension of the parameter search space
through a Karhunen-Loève (KL) expansion. We have also iterated on the filter update step to better account for the nonlinearity of the parameter estimation problem. We study the sensitivity of the system to the ensemble size, localization scale, dimension of retained KL modes, and the number of iterations. The performance of the proposed framework in term of estimation accuracy suggests that a well-tuned Joint-EnKF provides a promising robust approach to infer spatially varying seabed roughness parameters in the context of coastal ocean modeling.

1.4.3 Chapter 4: Bayesian Inference of Spatially-Varying Manning’s $n$
Coefficients in the Coastal Ocean Using a Generalized Karhunen-Loève Expansion and Polynomial Chaos

This chapter demonstrates how incorporating uncertainty in the prior covariance hyper-parameters of a Gaussian process may greatly enhance the performance of the Bayesian inference of spatially-varying coastal ocean parameters. This chapter of the thesis was submitted to the special issue in Ocean Dynamics as an article and is now under review.

Bayesian inference with coordinate transformations and polynomial chaos for a Gaussian process with a parametrized prior covariance model was introduced in [69] to enable and infer uncertainties in a parameterized prior field. The feasibility of the method was successfully demonstrated on a simple transient diffusion equation. In this work, we adopt a similar approach to infer a spatially varying Manning’s $n$ field in a coastal ocean model. The idea is to view the prior on the Manning’s $n$ field as a stochastic Gaussian field, expressed through a covariance function with uncertain hyper-parameters. A generalized Karhunen-Loève (KL) expansion, which incorporates the construction of a reference basis of spatial modes and a coordinate transformation, is then applied to the prior field. A polynomial chaos expansion with a similar coordinate transformation is then exploited to build a cheap surrogate of the large-scale advanced circulation (ADCIRC) numerical model, which is used to accelerate the Bayesian inference process using a Markov chain Monte Carlo algo-
algorithm. Water elevation data are inverted within an observing system simulation experiment framework, based on a realistic ADCIRC model, to infer the KL coordinates and hyper-parameters of a reference 2D Manning’s field. Our results demonstrate the efficiency of the proposed approach and suggest that including the hyper-parameter uncertainties greatly enhances the inferred Manning’s \( n \) field, compared to the case of covariance with fixed hyper-parameters.

1.4.4 Chapter 5: Combining Ensemble Kalman Filter and Multiresolution Analysis for Efficient Assimilation into Adaptive Mesh Models

In this chapter, we expand the scope of the thesis by introducing a new approach combining the EnKF with efficient AMR algorithm, with the aim to further advance the science of DA and UQ for hydrological and ocean simulations. This chapter of the thesis was submitted to *Computational Geosciences* as an article and is now under review.

A new approach is developed for efficient data assimilation into adaptive mesh models with the ensemble Kalman filter (EnKF). The EnKF is combined with a wavelet-based multi-resolution analysis (MRA) scheme to enable robust and efficient assimilation in the context of reduced-complexity adaptive spatial discretization. The wavelet representation of the solution enables the use of different meshes that are individually adapted to the corresponding members of the EnKF ensemble. The analysis step of the EnKF is then performed by involving coarsening, refinement, and projection operations on its ensemble members. Depending on the choice of these operations, five variants of the MRA-EnKF are introduced and tested on the one dimensional Burgers equation with periodic boundary condition. The numerical results suggest that, given an appropriate tolerance value for the coarsening operation, four out of the five proposed schemes significantly reduce the computational complexity of the assimilation system, with marginal accuracy loss with respect to the reference, full resolution, EnKF solution. Overall, the proposed framework
offers the possibility of capitalizing on the advantages of adaptive mesh techniques, and the flexibility of choosing suitable context-oriented criteria for efficient data assimilation.
Chapter 2

Assessing an Ensemble Kalman Filter Inference of Manning’s $n$
coefficient of an Idealized Tidal Inlet against a Polynomial Chaos-Based
MCMC

In this chapter, the relevance of EnKFs for parameters estimation is assessed in the context
of realistic coastal ocean modeling. We consider the Advanced Circulation model (AD-
CIRC) [12] and conduct OSSEs to attempt to recover the Manning’s $n$ coefficient from a
set of initial guesses and available data. To evaluate the EnKF performance for parameter
estimation, we compared its posterior distributions against those resulting from MCMC.
Since the implementation of MCMC is too costly with realistic coastal ocean models, we
resort to a polynomial chaos (PC) method to sample the MCMC posterior distribution with
a computationally efficient surrogate model.

2.1 Introduction

There is currently a growing interest in using EnKF methods for parameter estimation of
ocean models, as they have been found to be efficient and robust in many coastal ocean
state estimation applications (e.g., [51, 52, 55, 70]). This is particularly needed as the
predictability of coastal models may in certain situations depend more on the system pa-
rameters than the initial state. Parameter estimation within an EnKF framework is usually
achieved through state space augmentation [33, 34]. In this approach, parameters are con-
sidered to be part of the system state, and they are updated via the cross-covariance between
the parameters and the observations as sampled by the ensemble members, similar to the
way state variables are updated. However, it has been suggested that parameter estimation using EnKF may not perform well with strongly nonlinear systems [71]. This is because EnKFs are formulated based on Gaussian prior and noise assumptions [41, 72, 73, 74], which may not hold for such problems.

Few recent studies have evaluated ensemble Kalman-based techniques against the more sophisticated but computationally demanding Markov Chain Monte Carlo (MCMC) methods. The comparison of the EnKF for state estimation of the two-dimensional incompressible Navier-Stokes equations in a periodic box against the gold-standard MCMC in [25] suggests that the bias in the Gaussian approximation may become significant and the covariance may misrepresent the dependence between the parameters and the data when the dynamics are sufficiently nonlinear. [26] also tested parameter estimation in the context of an idealized 1D deep convection model, comparing an ensemble transform Kalman smoother (ETKS) with a full Markov Chain Monte Carlo (MCMC) method. Similar conclusions were drawn; with proper parameter choices, evaluated filters perform well in producing the distribution’s mean but not an accurate covariance because the model is strongly nonlinear.

To assess the relevance of EnKFs for parameters estimation in the context of realistic coastal ocean modeling, we consider the Advanced Circulation model (ADCIRC) [12] and conduct observation system simulation experiments (OSSEs) to attempt to recover the Manning’s $n$ coefficient from a set of initial guesses and available data. To evaluate the EnKF performance for parameter estimation, we compare the produced posterior distributions against those resulting from MCMC. Since the implementation of MCMC is too costly with realistic coastal ocean models, we resort to a polynomial chaos (PC) method to sample the MCMC posterior distribution with a computationally efficient surrogate model.

PC methods have become one of the standard approaches to propagate and quantify uncertainties in various disciplines, including mechanical systems (e.g., [75, 76, 77]), chemical systems (e.g., [78, 79, 80]), and more recently large scale oceanic systems (e.g.,
These methods represent the model parameters in terms of a spectral expansion in an orthogonal polynomial basis according to their probabilistic distributions. This enables the construction of extremely cost effective surrogate models that can then be used to efficiently sample the statistical properties of some Quantities of Interest (QoIs) of the model outputs \[68\ 15\ 29\]. The advantage of using PC based inference over an EnKF is in its ability to provide the full probability distribution function of the estimated parameters, Gaussian or not. Nonetheless, PC suffers from ‘the curse of dimensionality’, which makes it difficult to implement in problems with large number of stochastic parameters \[68\]. Here we show that with good tuning of the filter parameters, one can accurately recover the reference parameter, even when the posterior pdf is parameterized Gaussian. We further perform sensitivity experiments to various parameters (e.g., ensemble size and inflation factor) and study their impact on the posterior pdfs.

The rest of this paper is organized as follows. Section 2.2 establishes the Bayesian framework for parameter estimation, comprising the parameter inference methodology with the EnKF and PC-MCMC. Section 2.3 elaborates in detail the experimental setup. Inference results are presented and analyzed in section 2.4. Summary and conclusions are offered in section 2.5.

### 2.2 Bayesian inference

Bayes’ rule states that the probability distribution function (pdf) of the unknown parameters \(\theta\) conditioned on available data \(y\), can be computed as \[82\ 83\]

\[
p(\theta \mid y) = \frac{p(y \mid \theta)p_{pr}(\theta)}{p_Y(y)}, \tag{2.1}
\]

where \(p(\theta \mid y)\) is also known as the posterior pdf, \(p(y \mid \theta)\) is the likelihood function of obtaining data \(y\) given a parameter value \(\theta\), \(p_{pr}(\theta)\) is the prior distribution which defines the prior knowledge of the parameters, and \(p_Y(y)\) is a scaling factor, which is the density
function of all possible measurements.

2.2.1 Parameter estimation with MCMC

The most straightforward way to sample the posterior distribution is to use an MCMC method \cite{84, 24}. These methods iteratively generate a Markov chain in which the parameters vector $\theta_i$ at a given iteration $i$ depends only on the previous one $\theta_{i-1}$. Suppose the current value of the chain is $\theta_i$, one draws a proposal parameter $\theta'$, for instance from a Gaussian function centered at the current parameter: $Q(\theta_i) \sim N(\theta_i, \beta^2 I)$, and then calculates the ratio of the likelihood evaluated for the proposed sample $\theta'$ and the previous sample $\theta_i$:

$$r = \frac{L(\theta')}{L(\theta_i)},$$

where $L(\theta) = \pi(\theta \mid y)$. The new parameter $\theta_{i+1}$ is then chosen according to the following rule:

$$\theta_{i+1} = \begin{cases} \theta' & \text{if } \alpha < r, \text{ ACCEPT,} \\ \theta_i & \text{if Otherwise, REJECT.} \end{cases}$$

Here $\alpha \sim U(0, 1)$ and $\beta^2$ is the variance in the proposal drawing step. $\beta^2$ needs to be well-tuned in order to produce a well-mixed chain \cite{23, 85}. In this study, the value of $\beta^2$ was set by trial and error, but more sophisticated techniques could be also used such as adaptive MCMC \cite{86, 87} to achieve an average acceptance ratio between 0.40 and 0.50, which has been shown to be the optimal acceptance rate for inferring a 1-D Gaussian distribution \cite{88}.

2.2.2 Parameters estimation with an EnKF

Another mean of estimating $\pi(\theta \mid y)$ is a filtering approach in which the data are sequentially assimilated as they become available. The posterior is then computed conditioned on data available up to the estimation time only, and is updated with the Bayes’ rule every
time a new observation is available, using the most recent estimate as the prior [17].

The Kalman filter (KF) is optimal for state estimation when the system is linear and its noise is Gaussian [89]. The parameter estimation problem often arises, however, with a nonlinear system. One popular approach that has been developed in oceanography and meteorology to cope with the nonlinearity is the Ensemble Kalman filter (EnKF) and its variants (e.g., [34, 37, 38, 43, 90]). The EnKF uses an ensemble of realizations of the state vector, from which the first two moments are forecasted with the nonlinear model. The parameters inference is then performed recursively every time new observations become available, assuming Gaussian distributions.

EnKF methods can be stochastic, updating each ensemble member with perturbed observations, or deterministic, not requiring observations perturbations [36, 39, 40, 70]. The Singular Evolutive Interpolated Kalman (SEIK) filter belongs to the class of deterministic EnKFs and updates the sample mean and covariance of the forecast ensemble as in the KF. The SEIK filter is implemented in three steps: (1) a sampling step, (2) a forecast step, and (3) an analysis step. The sampling step generates an ensemble using a second-order resampling scheme that exactly matches the analyzed state and error covariance. The forecast step uses the nonlinear model to integrate the sampled ensemble members forward in time. The analysis step updates the forecast ensemble mean and covariance with incoming observations exactly as in the KF.

The filter starts from an estimate of the model state, $x_{k-1}^a$ and a low-rank ($r << n$) error covariance, decomposed as $P_{k-1}^a = L_{k-1}U_{k-1}L_{k-1}^T$, where $L_{k-1}$ is $n \times r$, and $U_{k-1}$ is an $r \times r$ positive definite matrix. The SEIK filter then samples the analysis ensemble as

$$ x_{k-1}^{a,i} = x_{k-1}^a + \sqrt{r + 1}L_{k-1}(\Omega_{k-1}C_{k-1}^{-1})^T, \quad i = 1, ..., r + 1, $$

where $r + 1$ is the ensemble size and $\Omega_{k-1}$ is $(r + 1) \times r$ matrix with orthonormal columns and zero column sums generated using Householder matrices [38]. The analysis ensemble
members are then integrated with the nonlinear model to obtain the forecast ensemble, \( \mathbf{x}_{k}^{f,i} \).

Taking the average \( \overline{\mathbf{x}}_k^{f,i} \) of the \( \mathbf{x}_k^{f,i} \) as the forecast state, and their sample covariance as the forecast error covariance, the latter can be decomposed as

\[
P_k^f = L_k U_{k-1} L_k^T,
\]

where

\[
L_k = \begin{bmatrix} \mathbf{x}_k^{f,1} & \cdots & \mathbf{x}_k^{f,r+1} \end{bmatrix},
\]

and

\[
U_k^{-1} = \rho^{-1} \left[ (r+1) T^T T \right]^{-1} + \left( L_k^T L_k \right)^{-1} L_k^T Q_k L_k \left( L_k^T L_k \right)^{-1} - 1 + \left[ (HL)_k \right]^T R_k^{-1} (HL)_k.
\]

Here, \( T \) is an \((r+1) \times r\) full rank matrix with zero column sums. \( R_k \) and \( Q_k \) are the observation and model covariance matrices, respectively. \( H \) is the observation matrix and \( \rho \) is an inflation factor, which increases the background error covariance and improves the forecast error when small ensemble sizes are used \([38, 90]\). The term \( \left( L_k^T L_k \right)^{-1} L_k^T Q_k L_k \left( L_k^T L_k \right)^{-1} \) represents the model error projection on the forecast ensemble. One may also account for the model error directly by perturbing the forecast model during the integration of the ensemble members \([91]\).

When an observation \( \mathbf{y}_k^i \) becomes available, the observation operator \( \mathbf{H}_k \) is applied to each \( \mathbf{x}_k^{f,i} \) in (2.4) to compute \( (HL)_k \). The forecast state is then updated to the analysis state as

\[
\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k (\mathbf{y}_k^i - \mathbf{H}_k \mathbf{x}_k^f),
\]

(2.6)
where $K_k$ is the Kalman gain computed as

$$K_k = L_k U_k (HL)_k^T R_k^{-1},$$

The analysis error covariance can be expressed as

$$P_k^a = L_k U_k L_k^T.$$  \hfill (2.7)

The SEIK filter has recently been successfully applied to state estimation in storm surge forecasting [70, 92], providing good estimates of state and error statistics of the system using small ensembles ($O(10)$).

[34] suggested that an EnKF could also be used for parameters estimation through joint state-parameters estimation, also commonly referred to as state augmentation [93, 94]. In this method, a vector of model parameters, $w$, is appended to the system state, to form the joint state-parameter vector

$$\tilde{x}_k = \begin{bmatrix} x_k \\ w \end{bmatrix}. $$

Given that the evolution of the model parameters is a stationary process, a time evolution system for the state-parameter augmented state vector is thus constructed as

$$\tilde{x}_k = \begin{bmatrix} x_k \\ w_k \end{bmatrix} = \begin{bmatrix} M(x_{k-1}) \\ w_{k-1} \end{bmatrix}, $$  \hfill (2.8)

The model parameters are not observed, so the corresponding observation operator is defined as
\[
\tilde{H}_k \tilde{x}_k^f = \begin{bmatrix} H_k x_k^f \\ 0 \end{bmatrix},
\]

where \( \tilde{H}_k := [H_k \ 0] \), is the modified observation operator. Following the new augmented system one may directly apply the SEIK filter steps to jointly estimate both the model state and unknown parameters, and the associated error statistics.

Joint EnKF estimation has several desirable features. Its implementation is straightforward and its computational cost is very reasonable. In addition, it provides information about the uncertainty of the parameter’s estimates. However, several issues in its implementation should be considered in practice. A specific issue concerns the assumption of stationary parameters. While the variance in the state variable increases during each forecast step, the variance in the parameter gradually decreases with the assimilation cycles. This may cause the ensembles to eventually collapse. The common practice to mitigate this problem is to use an inflation factor. In this approach, the posterior’s standard deviation is inflated by some factor to maintain a finite variance in the ensemble members. [93] suggested that inflation may not be always effective, depending on the characteristics of the parameter being estimated. Another inherent issue in the joint estimation method is that the parameters are only observed through the model state. Thus, the model state tends to converge to the true state much more rapidly than the estimated parameters. This suggests that increasing number of assimilation cycles may be required to obtain satisfactory estimates of the parameters [93]. Despite these issues, the joint-state EnKF approach has been successfully implemented and used in a wide range of problems [34, 93, 95, 96, 97].

2.2.3 Polynomial chaos-based MCMC (PC-MCMC)

MCMC is computationally prohibitive for realistic coastal ocean models, and can therefore only be implemented with some model reduction techniques. The polynomial chaos (PC)
method is an efficient approach in this regard [77, 81]. Uncertainty quantification based on a PC framework could be implemented following different techniques [98]. Due to the complexity of our coastal ocean model, we adopt a non-intrusive approach that requires no modification to the model code.

The PC method assumes that the model output $X$ to admit a spectral expansion of the form

\[ X(\xi) = \sum_{k=0}^{\infty} c_k \Psi_k(\xi) \approx \sum_{k=0}^{P} c_k \Psi_k(\xi), \tag{2.9} \]

where $\xi \in \Omega^* \subseteq \mathbb{R}^d$ are independent random variables, with density $F_\xi : \mathbb{R}^d \rightarrow \mathbb{R}^+$, parameterizing the random inputs, $\Psi_k$ are d-variate orthogonal polynomials and $c_k$ are the PC coefficients. In this work we use $\xi \overset{\text{iid}}{\sim} \mathcal{U}(-1,1)^d$ so the $\Psi_k$ are Legendre polynomials.

The expansion (2.9) is truncated to a certain polynomial order $m$, such that

\[ P = \frac{(d + m)!}{d!m!} - 1. \tag{2.10} \]

Since $\{\Psi_k\}_{0}^{P}$ forms an orthogonal system,

\[ \langle X, \Psi_k \rangle = c_k \langle \Psi_k, \Psi_k \rangle, \tag{2.11} \]

with

\[ c_k = \frac{\langle X, \Psi_k \rangle}{\langle \Psi_k^2 \rangle}. \tag{2.12} \]

Here the moments $\langle \Psi_k^2 \rangle$ of the multivariate Legendre polynomials in (2.12) can be evalu-
ated analytically [77] while \( \langle X, \Psi_k \rangle \) is obtained from the projection

\[
\langle X, \Psi_k \rangle = \int_{\Omega^*} X(s)\Psi_k(s)dF_\xi(s), \quad k = 0, \ldots, P.
\]

Thus evaluating \( c_k \) involves computing the values of a set of \( P + 1 \) integrals over \( \Omega^* \subseteq \mathbb{R}^d \), where \( \Omega^* \) is the image of the sample space over \( \xi \), which can be discretized as finite sums using an appropriate quadrature formula:

\[
\int_{\Omega^*} X(s)\Psi_k(s)dF_\xi(s) = \sum_{j=1}^{N_q} w_jX(\xi_j)\Psi_k(\xi_j).
\]

(2.13)

Here, \( \xi_j \in \Omega^* \) and \( w_j \) are the nodes and weights of the corresponding \( X \), and \( N_q \) is the number of quadrature nodes. The set of integration quadrature nodes comprises the so-called non-intrusive spectral projection (NISP) sample denoted by

\[
\mathcal{S} = \{\xi_j\}_{j=1}^{N_q} \subseteq \Omega^*.
\]

(2.14)

Thus, to evaluate (2.13), one needs to compute \( X(\xi_q) \) for all \( \xi_q \in \mathcal{S} \). Let \( \Pi \in \mathbb{R}^{(P+1) \times N_q} \) be the NISP projection matrix

\[
\Pi_{k,j} = \frac{w_j \Psi_k(\xi_j)}{\langle \Psi^2_k \rangle}, \quad k = 0, \ldots, P, \quad j = 1, \ldots, N_q,
\]

and \( \zeta \) be the vector with coordinates \( \zeta_j = X(\xi_j) \). Then the vector of PC coefficients \( c \) can be expressed as \( \Pi \zeta \), or in the coordinate form

\[
c_k = \sum_{j=1}^{N_q} \Pi_{k,j} \zeta_j = \sum_{j=1}^{N_q} \Pi_{k,j}X(\xi_j), \quad k = 0, \ldots, P.
\]

(2.15)

The complexity of NISP scales with \( N_q \) and can increase greatly with the number \( d \) of canonical random variables. Thus the application of this approach is usually computationally restricted to a limited number of parameters. For a more detailed mathematical
formulation of the NISP method the reader is referred to [68].

2.3 Experimental setup

2.3.1 The ADvanced CIRCulation Model (ADCIRC)

The coastal ocean model ADCIRC [99] is used in this study. This model is commonly implemented for simulating coastal and estuarine systems, as well as for analyses of water elevations and currents [5, 12]. The model solves a modified version of the shallow water equations (SWEs); the continuity equation is replaced by the second-order, hyperbolic generalized wave continuity equation (GWCE). This prevents spurious oscillations that often arise from the numerical solution of the original form [100, 101]. Together, the reformulation of the continuity equation and the momentum equation form the governing equations solved by ADCIRC. This system of equations are spatially discretized on unstructured, triangular elements and solved using a first-order continuous Galerkin finite element scheme. The time derivatives are approximated using centered finite differences in the GWCE and forward differences in the momentum equations.

A variety of physical domain scales and complex bathymetry structures can be represented in ADCIRC, including the range of scales necessary to represent the deep ocean basins and continental shelves, and coastal inland areas [92]. The minimum required inputs for ADCIRC are the description of the finite element mesh and tidal forcing parameters [102].

ADCIRC has been implemented and validated in many hindcast studies of several storms, including hurricanes Betsy (1965), Ivan (2004), Dennis (2004), Katrina (2005), Rita (2005) [2, 3, 103], Gustav (2008) [4] and Ike (2008) [16]. The model accuracy heavily depends upon the accuracy of model inputs and parameters. In this study we focus on the characterization of the bottom friction parameterized via the Manning’s $n$ coefficient.
2.3.2 The model discretization

We adopt the experiment of [5], which considers an idealized inlet with an ebb shoal domain as shown in Fig. 2.1. The structure of the domain includes a bay that is connected to the open ocean on the west side through an inlet with twin jetties. This model domain is a simplified version of a real ebb shoal, which is a common natural feature of coastal inlets. The domain contains 1,518 grid nodes and 2,828 elements, and is 4500 m wide and 3000 m long. The bathymetry of the domain is measured from the geoid toward the ocean floor. The depth of the ocean floor increases linearly from 3.8 m at the left-most boundary to 1 m at the mouth of the inlet on the west side of the domain. The landlocked area of the domain has a constant bathymetry of 1 m. The diameter of the ebb shoal is 750 m and the maximum height at the center of the ebb shoal is 4 m. On the open ocean boundary, ADCIRC is forced with the principal lunar semi-diurnal $M_2$ tidal constituent with an amplitude of 0.25 m relative to the geoid, and 2 s time step.
2.3.3 Parameterizing a field of Manning’s \(n\) coefficient

Instead of estimating the Manning’s \(n\) coefficient on each and every node of the domain, we model the field with low-dimensional parameterizations. This allows us to reduce the dimension of the parameter estimation problem. Here, we conduct the experiments in two Manning’s \(n\) coefficients settings: (1) The estimation of fields of constant Manning’s \(n\) coefficients. In this case, we assign a single constant value of Manning’s \(n\) coefficient to every node in the domain, as commonly implemented in coastal ocean models. (2) The estimation of two parameters, \(\alpha\) and \(\beta\), which parametrically define a 2-D field of Manning’s \(n\) coefficients. In this case, we parameterize the field with two piecewise constants of Manning values in the open ocean (\(\alpha\)) and the bay (\(\beta\)). The two constants field is parameterized in such a way that all nodes in the open ocean and land-locked area are assigned constant Manning’s \(n\) coefficients, and these increase or decrease linearly for the nodes within the inlet. This is expressed as

\[
w = \alpha E_1 + \beta E_2 + (\beta - \alpha) E_3.
\]  

Here \(w\) is the model parameter vector containing node-wise Manning’s \(n\) coefficients. \(E_1\), \(E_2\) and \(E_3\), of the same dimension as \(w\), are the parameterization coefficients vectors with values ranging between 0 and 1. These vectors are referred to as the basis functions in [5]. The parameterized two constants Manning’s field is denoted by \(n_{\alpha,\beta}\). Note that if \(\alpha = \beta\), the two constants Manning’s \(n\) field reduces to a constant Manning’s \(n\) field.

These two settings of Manning’s \(n\) fields are idealized and only reflect bathymetry surface characteristics of the artificial inlet with an ebb shoal. For a more realistic setting where such parameters vary in space, some data (e.g., bottom surface characteristics) need to be collected. More sophisticated static data integration techniques, such as kriging [104] and Gaussian process regression (GPR) [105], can then be used to generate a highly varying 2-D Manning’s \(n\) field conditioned on the data.
2.3.4 Manning’s $n$ coefficients classification

The model sensitivity greatly varies with the value of the Manning’s $n$ coefficients [5]. To study the sensitivity of the inverted solution to the choice of the prior distribution, we adopted a systematic Manning’s $n$ coefficients classification approach proposed in [5].

For the case of estimating single constant Manning’s $n$ coefficients, we classify Manning’s $n$ coefficients into five classes. The classes are defined by first running the ADCIRC model with the smallest Mannings $n$ coefficient, 0.005, as smaller friction coefficients generate tides with larger amplitudes. The largest mean amplitude of the tides generated at several locations throughout the domain is computed, and the Mannings $n$ coefficients that generate tides with mean amplitudes less than 20%, 40%, 60%, 80%, and 100% of this value are then divided into five classes, each class represents a set of Mannings $n$ coefficients that generate comparable water elevation data. These classes and their produced mean water elevation (e.g., at observation station 1) are summarized in Table 2.1.

Table 2.1: Mannings $n$ coefficients in idealized inlet with ebb shoal classified by amplitude of corresponding tidal data at station 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mean of tidal amplitude (m)</th>
<th>Mannings $n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.267989-0.214391</td>
<td>0.005-0.025</td>
</tr>
<tr>
<td>B</td>
<td>0.214391-0.160793</td>
<td>0.03-0.04</td>
</tr>
<tr>
<td>C</td>
<td>0.160793-0.107195</td>
<td>0.045-0.07</td>
</tr>
<tr>
<td>D</td>
<td>0.107195-0.053598</td>
<td>0.075-0.135</td>
</tr>
<tr>
<td>E</td>
<td>0.053598-0</td>
<td>0.14-0.2</td>
</tr>
</tbody>
</table>

For the two parameters case, $\alpha$ and $\beta$, the classification of Manning’s $n$ coefficients cannot be archived as was done in the single coefficient case without substantial computation since it is not obvious which pair of $\alpha$ and $\beta$ yields the largest mean tidal amplitude. Furthermore, different combinations of $\alpha$ and $\beta$ may also result in similar maximum water elevations, and the solution of the ADCIRC is non-monotonic with respect to the two constants Manning’s $n$ coefficients. Instead, we select a few pairs of $\alpha$ and $\beta$ values that produce maximum tidal amplitudes that are within 20% of those amplitudes that result from
ADCIRC when it is integrated with the true two constants parameterized Manning’s $n$ coefficients as the initial guesses. We chose a single true field $n_{0.005,0.1}$ in which the coefficient is set to 0.005 in the deep water of the open ocean and 0.1 in the shallow bay area. The selected true Manning’s $n$ coefficients represent suitable bottom stress of the deep water ($\alpha = 0.005$) and shallow water ($\beta = 0.1$) that are commonly used in coastal ocean models, e.g. [2][106].

### 2.3.5 Observation System Simulation Experiments (OSSEs)

Observation system simulation experiments (OSSEs) are conducted in this study. Synthetic observations are obtained by running ADCIRC simulations with specified initial conditions and Manning’s $n$ coefficients. The specified Manning’s $n$ coefficients are considered to be the true parameters, and we attempt to recover them from an inaccurate specification based on the extracted data. We initiate the model simulation with some given estimate of Manning’s $n$ coefficients, and then apply the SEIK filter and test for convergence of the updated parameters toward the true parameters after the subsequent assimilation cycles. We also apply the MCMC method using a PC surrogate of ADCIRC in the same OSSE setting to directly sample the posterior pdf of the Manning’s $n$ coefficients, and use its results as the reference solution to evaluate the SEIK estimates. We are interested not only in the parameter estimates, but also in their estimated uncertainties, and more generally in the posterior pdfs. We further investigate the OSSEs results with uniform and (more challenging) Gaussian priors.

#### Uniform prior

Here the prior $p_{pr}(\theta)$ in (2.1) follows a uniform distribution $\mathcal{U}(0.005,0.2)$, corresponding to the allowed range of Manning’s $n$ coefficients of ADCIRC. The parameter inference schemes (i.e. MCMC and the SEIK filter) are initialized by sampling this prior.

For the single coefficient case, we set the true Manning’s $n$ coefficient to a value from
the middle of each of the classes as described in the previous section. The values we attempt
to estimate are 0.015, 0.035, 0.06, 0.105 and 0.17. For each target Manning’s $n$ coefficient,
we first generate the synthetic water elevation data at 15 pre-defined locations representing
the 15 observation stations and perturb them with Gaussian noise of variance 0.01 m. The
observation error covariance matrix $R = \delta^2 \text{Id}$, with $\delta^2 = 0.01$ and $\text{Id}$ is the identity matrix.
For PC-MCMC inference, we initiate the MCMC chain from the middle value of Manning’s
$n$ coefficient $n = 0.1025$ and sample directly from the uniform prior $\mathcal{U}(0.005, 0.2)$. For the
EnKF inference, we initiate the model using $n = 0.1025$ and the initial ensemble members
are drawn from the same uniform prior as PC-MCMC. Assimilation starts after a 12 hour
ramp up period. We assume the data are available every hour over a 5 days period.

In the two coefficients case, $\alpha$ and $\beta$, the OSSEs are performed in the same manner as
in the single coefficient case described above. We generate synthetic water elevation data
from a given setting of Manning’s $n$ coefficients and consider these data as the truth. We
then try to recover these values starting from different initial guesses. The inferences are
initialized with uniform priors of the two constants Manning’s $n$ coefficients $\alpha$ and $\beta$ as
outlined in Table 2.2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$</td>
<td>Manning’s $n$ coefficient in the open ocean</td>
<td>$\mathcal{U}(0.005, 0.2)$</td>
</tr>
<tr>
<td>$n_2$</td>
<td>Manning’s $n$ coefficient in the landlocked area</td>
<td>$\mathcal{U}(0.005, 0.2)$</td>
</tr>
</tbody>
</table>

**Gaussian prior**

In the Gaussian prior setting, which is more common in a filtering/assimilation framework,
we invert the same perturbed data generated by the true Manning’s $n$ coefficients as in the
previous subsection. We follow [5] and initialize the assimilation from Gaussian priors
with the initial guesses as means and a small standard deviation equal to 0.01. This choice
of prior imitates the situation where the distribution and the support of the prior are not
well-known and one has to initialize the inference from some initial guesses. In this case, we allow more flexibility for the choice of prior and attempt to improve the accuracy of the estimation by online tuning of the inference parameters during assimilation.

In the single coefficient case, to estimate a true Manning’s $n$ coefficients of a specific class, we initialize the model using a value from any of the four remaining classes as the initial guess. For instance, to recover the true constant Manning’s coefficient of $0.17$, we may initialize the model with Manning’s $n$ coefficients = $0.015$, $0.035$, $0.06$ or $0.105$, respectively. This means 4 experiments for every single true constant coefficient we attempt to estimate, for a total of 20 experiments for the 5 considered true Manning’s $n$ values, as schematized in Fig.2.2.

For the two parameters case, we synthesize the observations of water elevation by running ADCIRC with $n_{0.005,0.1}$ and attempt to recover this true parameter field while initializing our simulation with different initial guesses. The selected initial guesses are $n_{0.005,0.005}$, $n_{0.1,0.1}$, $n_{0.06,0.06}$ and $n_{0.1,0.005}$. Thus we evaluate the results from incorrect initial guesses of either $\alpha$, $\beta$, or $\alpha$ and $\beta$. 

Figure 2.2: A schematic representation of one of the true constant Manning’s $n$ coefficients ($n = 0.017$) and the different Gaussian priors (with initial guesses from remaining classes as the means) we use in PC-MCMC and SEIK.
2.3.6 Construction of the PC surrogate

To cast ADCIRC uncertainties as stochastic variables, we let \( \mathbf{n} = (n_1, n_2)^T \) be the vector of random model inputs having uniform distribution as specified in Table 2.2. The inputs \( n_i \) are parameterized by \( \xi_i \sim \mathcal{U}(-1, 1) \) through

\[
n_i(\xi) = \mu_i + \sigma_i \xi_i, \quad i = 1, 2
\]

(2.17)

where \( \mu_i = (1/2)(a_i + b_i) \), and \( \sigma_i = (1/2)(b_i - a_i) \), so that \( p_i \sim \mathcal{U}(a_i, b_i) \) as in Table 2.2. Here \( a_i \) and \( b_i \) represent the pre-determined minimum and maximum values of a particular parameter, respectively. At a given time \( t \) and a point \( x \in \mathcal{G} \) and given a vector of random inputs \( \mathbf{n}(\xi) \), the model output is represented as

\[
\mathbf{X}(t, x, \xi) = \mathbf{A}(t, x; \mathbf{n}(\xi)),
\]

(2.18)

where \( \mathbf{A}(t, x; \mathbf{n}(\xi)) \) is the output of a deterministic ADCIRC simulation solved at time \( t \) and point \( x \) with the input parameters \( \mathbf{n}(\xi) \). Here, \( \mathbf{X} \) may correspond to any QoIs (e.g., water height at observation station).

We use NISP with 6 points Gaussian quadrature rule to compute the spectral expansion of the model output in the PC basis. With full tensorization in two stochastic dimensions (i.e. piece-wise Manning’s \( n \) coefficients), the construction of the PC model requires 36 model runs. To confirm that the quadrature ensures good representation of ADCIRC in this specific case, we generate the PC approximation of the model for different PC order 1 to 6 and then assess the convergence of these representations. This is shown in Fig. 2.3 along with the convergence to the PC coefficients at a selected station and time. According to these results, the distribution seems to level off at the PC order level \( m = 6 \), suggesting that a sixth-order expansion is sufficient. In addition, we also compute the RMSE between the water elevation from ADCIRC and its PC-surrogate. The RMSE is less than 5% of the
wave amplitude for all varying Manning’s $n$ coefficients. The error values are summarized in Table 2.3. Note that when using PC methods, it is useful to ensure that the posterior distribution is not in a low probability region of the prior, as the quality of the surrogate may deteriorate in such region. This is not an issue in our case because the prior is uniform, and the approximation is suitable for the entire range of the prior.

Table 2.3: RMSEs between the true model and the surrogate model (water elevations) for the whole simulation window at a specific observation station and varying Manning’s $n$ coefficients.

<table>
<thead>
<tr>
<th>Manning’s $n$ coefficients</th>
<th>RMSE (at station 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant Manning’s $n = 0.005$</td>
<td>0.0053</td>
</tr>
<tr>
<td>two constants Manning’s $\alpha = 0.063$ and $\beta = 0.030$</td>
<td>0.0011</td>
</tr>
<tr>
<td>two constants Manning’s $\alpha = 0.175$ and $\beta = 0.063$</td>
<td>0.00036</td>
</tr>
<tr>
<td>two constants Manning’s $\alpha = 0.030$ and $\beta = 0.142$</td>
<td>0.00052</td>
</tr>
<tr>
<td>two constants Manning’s $\alpha = 0.142$ and $\beta = 0.175$</td>
<td>0.00054</td>
</tr>
<tr>
<td>Constant Manning’s $n = 0.2$</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

2.4 Numerical Results

We present the PC-MCMC and SEIK inference results for the two different prior settings; uniform $\mathcal{U}(0.005, 0.2)$ and Gaussian with shifted means and a small variance. For both settings, we start by assuming the results of the PC-MCMC inference and then use these as the reference to evaluate those of SEIK.

2.4.1 Manning’s $n$ inference from uniform prior

PC-based MCMC Inference with uniform prior

We first infer the Manning’s $n$ coefficients using PC-surrogate ADCIRC with MCMC. The posterior distribution of estimated Manning’s $n$ coefficients are generated using 200,000 MCMC iterations. The results of estimating single constant Manning’s $n$ coefficients are presented in Fig. 2.4. It is shown that MCMC accurately recovers the true Manning’s $n$
Figure 2.3: Left panels: Distribution of water elevation generated by random Manning’s $n$ coefficients at the observed locations at several time steps $t$, as indicated. Different scales are used in each panels to facilitate comparison of the different curves. $p$ denotes the highest polynomial degree in the truncated expansion. In each case, the PC expansion was sampled $10^6$ times to generate the distribution curve. Right panels: PC coefficients at observed locations at several time steps $t$. 
coefficient (Table 2.4). The mode of the pdf well represents the true Manning’s value. The posterior pdf as sampled by MCMC exhibits a shape close to a single mode Gaussian pdf. The true Manning’s $n$ coefficients falls within the 95% confidence intervals as inferred from these pdfs.

The inference of two constants piece-wise Manning’s $n$ coefficients are shown in Fig. 2.5. As with the single coefficient case, PC-MCMC accurately estimates the true Manning’s $n$ coefficients for all tested values of $\alpha$ and $\beta$, as outlined in Table 2.5. The estimated posterior for $\alpha$ is not Gaussian but positively skewed. The mass of the distribution is concentrated near $\alpha = 0.005$ in all cases. This shape results from the construction of PC surrogate which assumes a uniform prior with a minimum value of 0.005. In all test cases, the estimated posteriors for $\beta$ approximately exhibits Gaussian shape with the true Manning’s $n$ coefficients falling within the 95% confidence intervals of the inferred pdfs, with exceptions of some cases where the pdf exhibits bimodal distribution (e.g., when the initial guess for $\beta$ is 0.1, which is the true $\beta$). For both the single coefficient and two coefficients cases, the estimates produced by PC-MCMC is clearly not sensitive to the choice of initial guesses, which is expected for long enough MCMC chains.

**SEIK Inference with uniform prior**

We infer the same set of true Manning’s $n$ coefficients as in the previous section using the SEIK filter with initial ensembles sampled from the uniform prior $\mathcal{U}(0.005, 0.2)$. After the model ramp up, these initial ensembles are updated following the SEIK filter procedure as described in section 2.2.2. Ensembles of 10 members were used in these experiments. The results of the SEIK filter inference compared to PC-MCMC for the single and two constants coefficients cases are shown in Fig. 2.6 and 2.7 respectively.

For the single coefficient case, the SEIK filter accurately recovers the true Manning’s $n$ coefficients for all classes, except for class D (i.e. $n = 0.105$) where the SEIK filter slightly underestimates the truth. The filter estimate is however within class D and is therefore
Figure 2.4: Posterior distributions of the constant Manning’s $n$ coefficients as estimated by PC-MCMC. The green dashed line represents the true value and the blue dashed line the estimated Manning’s $n$ coefficients, as the mode of the posterior pdf (red curve). Note that the estimates do not depend on the choice of initialized value of the MCMC chains if the MCMC chains are sufficiently long.
Figure 2.5: Posterior distributions of the piece-wise constant Manning’s $n$ coefficients as estimated by PC-MCMC. The green dashed line represents the true value and the blue dashed line the estimated Manning’s $n$ coefficients, as the mode of the posterior pdf (red curve). Note that the estimates do not depend on the choice of initialized value of the MCMC chains if the MCMC chains are sufficiently long.

considered acceptable. The posterior pdfs inferred by the SEIK filter are narrow compared to those computed by PC-MCMC. This suggests that the SEIK inference underestimates the uncertainty about the inferred parameter compared to the gold-standard MCMC, which is optimal given the same prior.

The inference of the two constant coefficients case shows similar results. $\alpha = 0.005$ and $\beta = 0.1$ are accurately recovered, but with a narrow posterior pdfs compared to the posterior pdfs resulting from PC-MCMC. In addition, because of the Gaussian assumption, the SEIK inferred posterior fails to capture the MCMC posterior’s heavy tail and further assigns positive probability for $\alpha < 0.005$.

In summary, this section shows that initializing the SEIK filter with uniform priors that cover the true Manning’s $n$ coefficients allows to accurately recovering the true value of the Manning’s $n$ coefficient in the single and two coefficients cases using only 10 ensemble members. The SEIK filter however underestimates the posteriors spread in all cases (as shown in Tables 2.4 and 2.5). The underestimation of the filter solution uncertainty will be addressed in section 2.4.3.
Figure 2.6: Posterior distributions of the constant Manning’s $n$ coefficients as estimated by SEIK with 10 ensemble members compared to PC-MCMC. The prior is $\mathcal{U}(0.005, 0.2)$ for both schemes. The green dashed line represents the true value. The black curve is the posterior pdf as estimated by SEIK and the red curve the posterior pdf as estimated by PC-MCMC.
Table 2.4: Final estimates, RMSEs between the estimates and the truths and pdfs variance of constant Manning’s $n$ coefficients as estimated by PC-MCMC.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Truth</th>
<th>Inference method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PC-MCMC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.015</td>
</tr>
<tr>
<td>Final estimate</td>
<td>0.035</td>
<td>0.035344</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.061107</td>
</tr>
<tr>
<td></td>
<td>0.105</td>
<td>0.105986</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.171362</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.015</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.035</td>
<td>0.000456</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.001092</td>
</tr>
<tr>
<td></td>
<td>0.105</td>
<td>0.001013</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.002637</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.015</td>
</tr>
<tr>
<td>Variance</td>
<td>0.035</td>
<td>3.6691e-05</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.000165</td>
</tr>
<tr>
<td></td>
<td>0.105</td>
<td>0.000140</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.000555</td>
</tr>
</tbody>
</table>

2.4.2 Inference sensitivity to a more challenging Gaussian priors

The OSSEs are conducted here as described in section 2.3.5 For each true Manning’s $n$ coefficient, we first generate ADCIRC outputs from which the observations are extracted. These data are then used both in the PC-MCMC and the SEIK filter to infer the truth starting from inaccurate initial guesses.

PC-based MCMC inference with Gaussian priors

We apply PC-MCMC to infer the same set of true Manning’s $n$ coefficients as in section 2.4.1 but with Gaussian priors of variance 0.0001. The posterior pdfs of the single and two coefficients cases are presented in Fig. 2.8 and 2.9, respectively.

For the single coefficient case, PC-MCMC successfully recovers the truth starting from any initial guess (i.e. prior’s mean) and regardless of the class of the true Manning’s $n$
Figure 2.7: Posterior distributions of the two constants Manning’s $n$ coefficients as estimated by SEIK with 10 ensemble members compared to PC-MCMC. The prior is $\mathcal{U}(0.005, 0.2)$ for $\alpha$ and $\beta$ for both schemes. The green dashed line represents the true value. The black curve is the posterior pdf as estimated by SEIK and the red curve the posterior pdf as estimated by PC-MCMC.

coefficient. Our results suggest that considering a sufficiently long MCMC chain (and enough prior variance), PC-MCMC with Gaussian prior converges to the true solution. The posteriors spread, though small, depends on how far the prior mean is from the truth. This suggests that MCMC is sensitive to the choice of the Gaussian prior. In other experiments, not shown here, when the prior is far from the truth and the variance of the prior is very small, MCMC failed to recover the truth. Larger inference window and more iterations may improve MCMC results in this case. We also found that, although the truth is accurately estimated for all classes, the posterior spread is larger when the true parameters belong to the large classes (for instance, the posterior spread of estimating the truth = 0.17 are much larger than those of truth = 0.015). This is due to the different sensitivity of the model to the class of constant Manning’s $n$ coefficients as discussed in subsection 2.3.4.

Similar results are obtained for the two parameters cases (Fig. 2.9). The truth $n_{0.005,0.1}$ is accurately estimated with all initial guesses. However, in this case, the posterior spreads are approximately the same. This is because the initial guesses are selected from a set of Manning’s $n$ coefficients that produce similar average maximum water elevation as the truth (i.e. within 20% of the truth). As a result, the effect of the sensitivity of PC-MCMC to
Table 2.5: Final estimates, RMSEs between the estimates and the truths and pdfs variance of piecewise Manning’s \(n\) coefficients used to parameterize two constants Manning’s field as estimated by PC-MCMC. \(n_{\alpha,\beta}\) represents the field parameterized by \(\alpha\) and \(\beta\).

<table>
<thead>
<tr>
<th>Measures</th>
<th>Truth</th>
<th>Initial guess</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC-MCMC</td>
<td>SEIK</td>
</tr>
<tr>
<td>Final estimate</td>
<td>(\alpha = 0.005)</td>
<td>0.005505</td>
</tr>
<tr>
<td></td>
<td>(\beta = 0.1)</td>
<td>0.098813</td>
</tr>
<tr>
<td>RMSE</td>
<td>(\alpha = 0.005)</td>
<td>0.000605</td>
</tr>
<tr>
<td></td>
<td>(\beta = 0.1)</td>
<td>0.000186</td>
</tr>
<tr>
<td>Variance</td>
<td>(\alpha = 0.005)</td>
<td>7.3667e-06</td>
</tr>
<tr>
<td></td>
<td>(\beta = 0.1)</td>
<td>4.8089e-05</td>
</tr>
</tbody>
</table>

The prior in the two coefficients case is not as pronounced as in the single coefficient case.

**SEIK inference with Gaussian priors**

We apply the SEIK filter to estimate the Manning’s \(n\) coefficients starting from different initial guesses using 10 ensemble members. The results of the OSSEs are shown in Fig. 2.10. The SEIK filter accurately recovers the Manning’s \(n\) coefficient from the correct class in the cases where the truth lies in any of the three smallest classes, A, B, and C, regardless of the class of the initial guesses. Values in class D are only recovered when the initial guesses are from class C and E. For the estimation of a value from class E, the SEIK filter is able to recover this value only when the initial guess is from the neighboring class (i.e. class D).

In SEIK, the initial ensemble is generated by adding Gaussian noise to the initial guess. Then the ensemble is updated after every analysis step. By construction, the filter ensemble represents, approximately and based on Gaussian assumptions, the pdf of the parameter conditional on the available observation. We plot the evolution of the 95% percentile of Manning’s \(n\) coefficient along with their filter estimates, taken as the mean of the ensemble members (Fig. 2.10). It is clear that the uncertainties in the parameters decrease as the filter
Figure 2.8: Posterior distributions of the constant Manning’s $n$ coefficients as estimated by PC-MCMC initialized with Gaussian priors with different means. Black dashed lines represent the true values. Colored dashed lines are the estimated Manning’s $n$ coefficients, or the modes of the posterior distributions. Gray lines and curves are respectively the posterior distributions from uniform prior and their modes. The estimates are not sensitive to the choice of the prior if the MCMC chains are sufficiently long. The spread of the posterior distribution is clearly sensitive to the choice of the prior mean.
Figure 2.9: Posterior distributions of the two constants Manning’s $n$ coefficients as estimated by PC-MCMC initialized with Gaussian priors with different means. Black dashed lines represent the true values. Colored dashed lines are the estimated Manning’s $n$ coefficients, or the modes of the posterior distributions. Gray lines and curves are respectively the posterior distributions from uniform prior and their modes. The estimates are not sensitive to the choice of the prior when the MCMC chains are sufficiently long.

assimilates more data in time, as one may expect. The truth falls within the 95% estimated confidence interval for low Manning’s $n$ coefficients (e.g., 0.015, 0.035 and 0.06). To estimate the parameter in larger classes, a larger ensemble and/or longer simulation window may be required, which we will examine later.

In the second set of OSSEs we estimate a two constants parameterized field of Manning’s $n$ coefficients. The point estimation of the two constants parameterized Manning’s $n$ coefficients and their uncertainties are shown in Fig. 2.11. We obtain accurate estimates of the parameters for all initial guesses except for the initial guess $n_{0.005,0.005}$. In this case $\beta = 0.1$ is significantly underestimated. This is similar to the constant coefficient case where the SEIK filter is not very successful at recovering the large Manning’s value starting from the smallest guesses. However, for each initial estimate, the initial errors with respect to the truth are still reduced by at least 80%. In contrast to PC-MCMC, the posterior pdf produced by SEIK assigns some probability for $\alpha$ values below 0.005 due to Gaussian assumption on the posterior. However, we observe zero probability for negative values, since the pdfs spread around the mean estimates are small. For the two constants parameterized case, the choice of the initial guesses hardly affect the final estimates even when $\alpha$ and/or
Figure 2.10: Estimates of various Manning’s $n$ fields parameterized by a single value using 10 ensemble members. From top to bottom, left to right the true Manning’s $n$ value is 0.015, 0.035, 0.06, 0.105 and 0.17. The uncertainty of the estimates are represented by opaque color band (95$^{th}$ percentile) around the mean.
Figure 2.11: Estimates of two constants piecewise Manning’s $n$ coefficients, $\alpha$ (left) and $\beta$ (right). The uncertainty of the estimates are represented by opaque color band (95th percentile) around the mean.

$\beta$ for each guesses are vastly differ. We found no specific relation between the estimation accuracy and the choice of initial guess for the two constants coefficients case. This is because all initial guesses produce an average maximum water elevation within 20% of those obtained when the model is integrated with the true parameters. In this case, one may consider all initial guesses to be within the same class as the truth in term of produced water elevation, in analogy to the classification of the single constant coefficients OSSEs.

The accuracy of the estimated water elevation (by MCMC and SEIK) in the single coefficient case is directly related to the accuracy of the inferred roughness parameters. This is due to the monotonic dependence of the water levels on the Mannings $n$ coefficients \cite{5}. Accurate estimates of the Mannings $n$ coefficient produce accurate estimates of water elevations, and vice versa (not shown). Likewise, we noticed that accurate estimates of the water elevation in the two coefficients OSSEs were only obtained in the cases where the Mannings $n$ coefficients $\alpha$ and $\beta$ were well recovered. Similar findings were reported in \cite{5} to which we refer the reader for more detailed discussion and analysis.

Fig. 2.12 and 2.13 compare the posterior pdfs of the Manning’s $n$ coefficient as estimated by SEIK with 10 ensemble members and MCMC for a constant field and two constants parameterized field, respectively. The pdfs produced by SEIK are much narrower in comparison to those estimated by MCMC. SEIK also underestimates the truth in
most cases compared to MCMC which experiences no difficulty in recovering the true parameters. Furthermore, one can clearly observe the dominant Gaussian shape of the pdfs estimated by SEIK, as one might expect. The MCMC posterior distributions cover the wide range of Manning’s $n$ coefficients and their means are clearly centered at the true values. In this specific setting, the SEIK filter with small ensemble sizes does not provide a good description of the uncertainty of the estimated parameters. It also estimates values below the range of possible values as specified by the prior due to its Gaussian assumption, even though we do not observe likely probability for negative values.

### 2.4.3 SEIK tuning

#### Effect of increasing ensemble size

The SEIK filter should benefit from increased ensemble size [38, 107]. However, increasing the number of ensemble members means increased computational cost; for instance doubling the size of the ensemble means doubling the number of required model runs, which in the case of ADCIRC significantly increases the computational load. This section examines the SEIK filter performance with respect to the ensemble size.

For the constant Manning’s $n$ coefficient case, we use the SEIK filter with ensemble sizes of 10, 20, 100, 200, 300 and 400, and compare the parameter posterior focusing on both the mean of the ensemble and the uncertainty as estimated from the spread of the ensemble and represented by the 95th percentile of the parameter. The posterior distributions estimated with different ensemble sizes and initial guess equal to 0.06 are plotted in Fig. 2.14.

With 100 ensemble members, the filter performance significantly improves in comparison to using only 10 members. This can be for example clearly seen in the case where we choose the initial guess from class C to estimate the truth in class E (Fig. 2.14). The final estimate has increased from 0.1510 with 10 members to 0.1645 with 100 members, which has less than 3.5% relative error (Table 2.7). However, with 100 members, the truth
Figure 2.12: Comparison of the posterior pdfs of the smallest and the largest class of constant Manning’s $n$ coefficients as estimated by SEIK with 10 ensemble members (at the end of the assimilation period) and MCMC.
Figure 2.13: Comparison of the posterior pdfs of two constants Manning’s $n$ coefficients as estimated by SEIK (at the end of the assimilation period) and MCMC. Distributions of $\alpha$ are plotted in the left panels and those of $\beta$ in the right panels.
in class E still lies outside the 95% confidence interval of the pdf as estimated from class C. An ensemble size of 300 members or more is needed in order for the estimate to fall within the 95% confidence interval of the estimate (when the initial guess is 0.06). No other significant improvement is observed for increasing ensemble size beyond 300 members. In the case of estimating small Manning’s $n$ coefficients with the guess from a larger class, increasing the ensemble improves the estimate. Nevertheless, using 100 ensemble members is sufficient to recover the truth in a smaller class. Clearly, recovering the truth in larger classes from smaller classes is more accurate with increased ensemble size. We also notice faster convergence toward the truth with increased ensemble sizes. In term of computational complexity, using 100 members costs approximately ten times the execution time of running the SEIK filter using 10 ensemble members. This could be consequent for a large scale model such as ADCIRC. Thus, increasing the ensemble size does not seem to be the most efficient strategy to improve the estimation of low dimensional representations of Manning’s $n$ coefficients. Also, the computational time of PC-MCMC with 200,000 iterations is approximately twice the computational time of SEIK with 10 members. In more realistic applications with a spatially varying Mannings $n$ coefficients, the PC method may become unpractical, and an EnKF-like inference scheme will be more attractive in this case. A summary of the execution time of SEIK inference with 10 and 100 members, and PC-MCMC is provided in Table 2.6.

Table 2.6: The computational complexity of different inference methods. The PC-MCMC and filtering codes are written in matlab. The codes are executed on Intel Xeon(R) CPU X5550 @2.67GHz × 8 processor and the memory of 23.5 GiB.

<table>
<thead>
<tr>
<th>Inference method</th>
<th>Execution time</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC-MCMC</td>
<td>21,414.58 sec</td>
</tr>
<tr>
<td>SEIK (10 members)</td>
<td>11,176.25 sec</td>
</tr>
<tr>
<td>SEIK (100 members)</td>
<td>103,611.67 sec</td>
</tr>
</tbody>
</table>

For the case of two constants parameterized Manning’s $n$ field, we compare the esti-
Figure 2.14: Comparison of the posterior pdfs as estimated by SEIK for constant Manning’s $n$ coefficient with three different ensemble sizes and initial guess = 0.06. Left panels: the 95th percentile over time. Right panels: the posterior pdfs at the end of the assimilation period.
Table 2.7: Final estimates, RMSEs between the estimates and the truths and pdfs variance of constant Manning’s \( n \) coefficients as estimated by SEIK with 100 members starting from initial guess = 0.06.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Initial guess</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.015</td>
</tr>
<tr>
<td>Final estimate</td>
<td>0.015003</td>
</tr>
<tr>
<td>RMSE</td>
<td>3.3684e-06</td>
</tr>
<tr>
<td>Variance</td>
<td>6.3893e-09</td>
</tr>
</tbody>
</table>

As expected, we found clear improvement in estimating the parameters, especially \( \beta \) when using 100 ensembles, in comparison to using only 10 ensembles, which underestimates the truth for all initial guesses (Table 2.8). However, the spread of the pdfs in the case of 100 members does not significantly differ from those obtained in the 10 members case. Both estimate a variance of the same order of magnitude (e.g., \( O(10^{-8}) \) for \( \alpha \) and \( O(10^{-7}) \) for \( \beta \)). PC-MCMC posterior pdfs exhibit much larger variances (\( O(10^{-5}) \)). Compared to PC-MCMC, the SEIK filter accurately recovers the truth, but underestimates the associated uncertainty by as much as two orders of magnitude. This indicates that the SEIK filter fails to accurately describe the uncertainty of the inferred parameters. This likely explains the cause of SEIK’s inability to recover some parameter, e.g., Manning’s \( n \) values from class E. Covariance inflation is expected to improve the SEIK filter performance in that regard. This will be investigated in the next section.

The effect of inflation

In the previous section we showed that increasing the ensemble size is effective, however large ensembles are needed to accurately recover the true parameter, resulting in increased computational cost. Another way to enhance the robustness of an EnKF is to inflate the forecast error covariance [34,108]. Covariance inflation is now considered to be an impor-
Figure 2.15: Top row: sensitivity of the posterior distribution as it results from SEIK for the case of piecewise Manning’s $n$ coefficient in the open ocean ($\alpha$) to different ensemble sizes. Left column: the 95th percentile over time, Right panels: the pdf of the posterior at the end of the simulations. Bottom row: comparison of the posterior distribution as it results from SEIK for piecewise Manning’s $n$ coefficient in the landlocked area ($\beta$) with different ensemble sizes. Left panels: the 95th percentile over time, Right panels: the pdf of the posterior at the end of the simulations. All experiments start from the initial guess $\alpha = 0.1$ and $\beta = 0.1$. 
Table 2.8: Final estimates, RMSEs between the estimates and the truths and pdfs variance of piecewise Manning’s $n$ coefficients used to parameterize two constants Manning’s field as estimated by SEIK with 100 members. $n_{\alpha,\beta}$ represents the field parameterized by $\alpha$ and $\beta$.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Truth</th>
<th>Initial guess</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$n_{0.005,0.005}$</td>
</tr>
<tr>
<td>Final estimate</td>
<td>$\alpha = 0.005$</td>
<td>0.005481</td>
</tr>
<tr>
<td></td>
<td>$\beta = 0.1$</td>
<td>0.093857</td>
</tr>
<tr>
<td>RMSE</td>
<td>$\alpha = 0.005$</td>
<td>0.000481</td>
</tr>
<tr>
<td></td>
<td>$\beta = 0.1$</td>
<td>0.088857</td>
</tr>
<tr>
<td>Variance</td>
<td>$\alpha = 0.005$</td>
<td>3.9019e-08</td>
</tr>
<tr>
<td></td>
<td>$\beta = 0.1$</td>
<td>2.4972e-07</td>
</tr>
</tbody>
</table>

 tant auxiliary technique for successful implementation of an EnKF.

We perform the same OSSEs as in the previous section for the idealized inlet ebb shoal using 10 ensemble members, with inflation factors, $\rho = 1.1, 1.2, 1.5$ and 2. Fig. 2.16 shows the performance of the SEIK filter in recovering the true constant Manning’s $n$ coefficient 0.17 from an initial guess of 0.015 with several inflation factors. Significant improvement is observed when using inflation and the largest class of Manning’s $n$ coefficient are now accurately recovered from an initial guess from the smallest class. With inflation, the estimates further converge much faster toward the true values; convergence is reached before the 60th assimilation cycle in all cases except for the case with the smallest inflation, $\rho = 1.1$. Inflating the covariance also widens the ensemble spread around the mean, as expected. The range of the posterior pdf remains positive, which is consistent with the range of the Manning’s $n$ coefficients in all tested cases. On top of accurately recovering the true Manning’s $n$ coefficient, the SEIK filter with a well-tuned inflation significantly improves the posterior pdf. The variances of the posterior pdfs produced by the SEIK filter with inflation are approximately the same order of magnitude as those produced by PC-MCMC. An inflation factor of 1.2 is sufficient to recover the posterior pdf, i.e. for the truth to fall within the 95% confidence interval of the estimated posterior.
The results of estimating the two constants parameterized Manning’s \( n \) field using the SEIK filter with inflation are shown compared to using SEIK without inflation in Fig. 2.17 and 2.18. We see similar improvements as in the constant coefficient cases. Fig. 2.19 and 2.20 compare the posterior pdfs of the Manning’s \( n \) coefficients as estimated by the SEIK filter with an inflation factor \( \rho = 1.2, 1.5 \) and 2.0 for the single and two constants parameterized fields. An inflation factor of 1.2 is sufficient for the SEIK filter to accurately recover the true Manning’s coefficients from all classes. However, an inflation factor of 2.0 is required in most cases to obtain posterior’s variance that is of the same order of magnitude as those produced by PC-MCMC. Using inflation imposes larger ensemble spread and also improves the mean estimates in all cases. However, in the two constant coefficients case, the SEIK filter with inflation assigns probability to negative values of \( \alpha \), because of its Gaussian framework.

### 2.5 Discussion and Conclusions

We investigate the effectiveness of estimating the Manning’s \( n \) coefficients in the coastal ocean model, ADCIRC, using an ensemble Kalman filter, namely the SEIK filter. The filter performance was evaluated by comparing its estimated posterior distributions to those resulting from an MCMC method. To enable the implementation of MCMC with an idealized ADCIRC coastal ocean model, we resorted to the polynomial chaos (PC) approach to construct a surrogate model of ADCIRC. PC allows likelihoods to be evaluated at a fraction of the time required for a full evaluation with ADCIRC. A non-intrusive spectral projection scheme was used to derive the PC representation of the water elevation of ADCIRC. Using the PC expansion, we created a surrogate model for an idealized inlet with ebb shoal configuration. A degree 6 PC model was found to be adequate for producing the desired model response to changes in the Manning’s \( n \) coefficients. PC-MCMC avoids the restriction of the Gaussian posteriors required in ensemble Kalman filters.

We conduct OSSEs in which synthetic water elevation data is generated from various
Figure 2.16: Comparison of the posterior distribution as it results using SEIK for constant Manning’s $n$ coefficient between SEIK without and with inflation factor (1.1, 1.2, 1.5 and 2.0 respectively), initial guess = 0.015. Left panels: the 95th percentile over time, Right panels: the pdf of the posterior at the end of the simulations.
Figure 2.17: Comparison of the posterior distribution as it results using SEIK for two constant piecewise Manning’s $n$ coefficients between SEIK with and without inflation in the open ocean area. Left panels: the 95th percentile over time, Right panels: the posterior pdfs at the end of the assimilation period.
Figure 2.18: Comparison of the posterior pdfs as it results using SEIK for two constant piecewise Manning’s $n$ coefficients between SEIK with and without inflation in the land-locked area. Left panels: the 95th percentile over time, Right panels: the posterior pdfs at the end of the assimilation period.
Figure 2.19: Comparison of the posterior pdfs of classes of constant Manning’s \( n \) coefficients as estimated by SEIK with different inflation (at the end of assimilation period) and MCMC.
Figure 2.20: Comparison of the posterior pdfs of two constant piecewise Manning’s $n$ coefficients as estimated by SEIK (at the end of the assimilation period) and MCMC. Right panels: the distributions of $\alpha$, Left panels: the distribution of $\beta$. 
fields of Manning’s $n$ coefficients. We then attempt to recover these fields from incorrect initial guesses using both PC-MCMC and EnKF methods. We first investigate the sensitivity of the inference methods to the choice of prior. For the uniform prior case with a support covering the range of possible Manning’s $n$ coefficients, both PC-MCMC and SEIK successfully recover the true coefficients in the single and two constants Manning’s settings. The SEIK filter however produces much smaller posterior’s spread compared to PC-MCMC, implying an underestimation of the parameter’s uncertainty.

We further explore the sensitivity of the inference to more challenging Gaussian priors. In this case, PC-MCMC still accurately recover the truth in all tested scenarios. However, the posterior’s variance varies with the priors, depending on how far the prior’s mean is from the truth. The MCMC chains required larger number of iterations to converge when the prior was further away from the truth. For the constant field, the SEIK filter has difficulty recovering the Manning’s field of large values from small initial guesses. We show that increasing the number of ensemble members improves the filter results. The accuracy of the filter estimates increases by more than 200% when the ensemble size is increased from 10 to 100 and the truth lies within the 95% confidence interval of the estimate when we use 300 members (or more). For the two constants parameterized field of both domains, the use of a small ensemble size is enough to recover the true parameters in every test case. Since the increase of ensemble size requires more computation, we also investigate the impact of inflation of the background error covariance on the filter performance. In this case, the estimates converge to the truth much faster, even when the filter is implemented with 10 members. The estimates are also accurate regardless of the choice of the initial guesses. The use of inflation widens the posterior distribution of the parameter produced by the SEIK filter and makes it more comparable to that of PC-MCMC.

The SEIK filter shows good capability to accurately recover the true parameters, consistent with the results of [25]. Using inflation improves the estimation, convergence and distribution spread. However, over-inflation can cause the SEIK filter to enable negative
Manning’s $n$ coefficients due to the Gaussian assumption, and thus the filter needs to be well tuned. In contrast to the study of [26], who estimated 10 parameters of a simple deep convection model using 10,000 ensemble members, we successfully obtained good approximations of posterior pdfs using small ensemble sizes with SEIK (only 10 members) and reasonable inflation.

To conclude, while the PC-based approach coupled with MCMC is well-suited for parameter estimation problems with low dimension stochastic systems, its efficiency may still be limited to real-time coastal ocean forecasting, which often includes a large number of parameters. Sequential data assimilation schemes could provide a robust alternative to compute the pdfs. Despite their inherent Gaussian assumption, our results suggest that the SEIK filter with enough tuning of the ensemble size and the inflation factor may produce reliable estimates of the pdf mean and variance that are close to those of MCMC. This supports the idea of using an ensemble Kalman-type filter for parameter estimation in the context of large-scale realistic coastal ocean models, and for providing useful information about the associated uncertainties at reasonable computational cost. The parameter estimation problem of large dimensional spatially varying parameters with the SEIK filter will be explored. We will assess the relative merits of this approach by contrasting with more elaborate PC-based methods [69, 109].
Chapter 3

Ensemble Kalman Filter Inference of Spatially-varying Manning’s $n$ coefficients in the Coastal Ocean

In the preceding chapter, we validate the EnKF framework for parameters estimation on the coastal ocean. In this chapter, the Singular Evolutive Interpolated Kalman (SEIK) filter, a deterministic square-root version of EnKF [38, 90, 92], is used for inferring a 2D spatially varying Manning’s $n$ coefficient. The approach we follow resembles that of [48], which consists of a sequence of methods to formulate the inference of parameters, including a statistical parameterization of the parameter search space, the construction of a synthetic parameter field, the generation of an initial (prior) ensemble, the implementation of a model reduction technique, and finally the application of a parameter inference method. We test the above procedures using OSSEs with ADCIRC. The parameter inference is conducted on both the full parameter space and truncated Karhunen-Loéve (KL) subspaces. The inference results in the full space provides us the primary information regarding the effectiveness of SEIK to estimate spatially-varying ocean bottom stress, while the inference of the same parameter in the KL space offers enhancement of the parameter search space and important model reduction with acceptable error trade-off.

3.1 Introduction

The shallow water equations (SWEs), derived from depth-integrating the Navier-Stokes equations, have been widely used in coastal ocean modeling. They assume that the horizontal length scale of the problem domain is much larger than the vertical length scale
under hydrostatic pressure \[7, 12\]. In real world applications, the numerical solution of the SWEs is subject to various sources of uncertainty, such as modeling errors, numerical discretization, inputs uncertainty, etc. In particular, the uncertainty associated with the poor characterization of the model parameters is considered a major source of error \[1, 110, 111\]. A number of recent studies have therefore focused on quantifying and reducing the uncertainties associated with input parameters, aiming to achieve more reliable forecasts in fluid flow modeling (e.g. \[14, 15, 28, 112\]). In coastal ocean modeling, the specification of a parameter called “the Manning’s \(n\) coefficient of roughness”, used to define the bottom stress components in the SWEs, is particularly important \[5, 27, 113\].

The Manning’s \(n\) coefficient is an empirically derived parameter, defined as the resistance to water flow due to bottom surface characteristics (e.g., sands, rocks and reefs etc.). It is used to describe multiple types of resistance, e.g. friction resistance, form resistance, wave resistance, and resistance of flow instability \[9, 106\]. It enters the SWEs via the momentum equations, and the amplitude of the water column at a given point in the model domain can be highly sensitive to its value \[5\]. The Manning’s \(n\) coefficient cannot be measured directly \[13\] and often exhibits spatially heterogeneous variability. It also depends on the ocean bottom surface characteristics; changes in the ocean floor during extreme events (such as storm surges and tsunamis) may further alter the near-shore Manning’s \(n\) field. In such hazardous scenarios, it is critical that changes in ocean bottom stress be detected and updated to accurately predict water height. Unfortunately, the acquisition of the complete knowledge of Manning’s \(n\) coefficients in realistic settings is not feasible. Parameter identification by trial-and-error, e.g. comparing the SWEs solution produced by different Manning’s fields to observations, is tedious and impractical \[114\]. As a consequence, parameter specification methods, based on established look-up tables for each land cover type and roughness, have been commonly used to parameterize Manning’s \(n\) fields in large scale coastal ocean models \[2, 12, 115\]. A more advanced specification method based on a random forest model was also proposed in \[116\]. In this paper, we resort to a well-established
inverse modeling approach \cite{5} to infer spatially-varying Manning’s $n$ coefficients.

A number of approaches have been developed to solve parameter estimation problems in the context of meteorology and oceanography (e.g. \cite{26, 42, 97, 117}). Many are originally motivated by optimal control theory \cite{118}, and are based on the minimization of a cost function penalizing discrepancies between model outputs and observations \cite{119, 120, 121, 122}. However, this approach can be computationally demanding and typically requires the development of an adjoint model \cite{123, 124, 125}. Another popular approach for parameter inference is through the Bayesian framework \cite{17, 18}, where the parameters are represented with probability density functions (pdfs) conditioned on available data. The parameter inference problem is then viewed as the transformation of a prior pdf to a posterior pdf by incorporating the likelihood of the observations \cite{21}. The posterior is rarely explicit and often needs to be sampled as a collection of realizations that are consistent with data and prior information \cite{22}. The most popular implementation of this method is the Markov Chain Monte Carlo (MCMC) method (e.g., \cite{23, 24}), which has become more practical in recent years with increases in computational power. The primary advantage of MCMC is the ability to produce a full approximation of the posterior distribution. As a result, MCMC is often treated as the benchmark to evaluate the performance of other parameter inference methods \cite{25, 26, 27}.

Bayesian inference can also be cast as a filtering problem in which the posterior distribution is updated sequentially as data becomes available \cite{30}, an approach known as data assimilation. A Bayesian filter operates as a succession of forecast steps to propagate the pdf of the unknowns forward in time, and update steps to incorporate data every time new observations become available. For parameter estimation, filtering schemes usually apply the standard augmented state-parameter technique \cite{33, 34, 35}, that allow the state and parameters of the system to be estimated concurrently. Currently, the most popular approach for data assimilation into ocean models is the Ensemble Kalman Filter (EnKF) \cite{30, 36} and its deterministic versions (\cite{34, 37, 38, 39, 40}, to cite but a few). An EnKF follows a
Monte Carlo framework to integrate an ensemble of model realizations in the forecast step and then applies a linear Kalman correction in the update step [41]. The stochastic EnKF assimilates perturbed observations and this was shown to induce noise in the final solution when the filter is implemented with small ensembles [40, 42]. Deterministic EnKFs, which avoid observations perturbations, mainly differ in the way they sample the new analysis ensemble after the filter update step. Various deterministic EnKFs were compared with a realistic setting of ADvanced CIRCulation (ADCIRC) model in the Gulf of Mexico [42], showing that, with enough tuning, these filters performed closely well, all outperforming the stochastic EnKF.

In this study, we are interested in the inference problem of a 2D spatially varying Manning’s $n$ coefficient. The approach we follow resembles that of [48], which consists of a sequence of methods to formulate the inference of parameters, including a statistical parameterization of the parameter search space, the construction of a synthetic parameter field, the generation of an initial (prior) ensemble, the implementation of a model reduction technique, and finally the application of a parameter inference method. We generate realizations of 2D spatial maps of Manning’s $n$ coefficients subjected to a few synthetic observations based on the sequential simulation algorithm of multi-Gaussian fields [126]. A reference field and an initial ensemble are then selected from these realizations. Next, we apply the Singular Evolutive Interpolated Kalman (SEIK) filter, a deterministic EnKF [38, 90, 92], to estimate the reference Manning’s $n$ field using the Joint-EnKF. Localization [127, 128] is also applied to enable efficient implementation of the SEIK with reasonable ensemble sizes and to remove any spurious correlations between distant points. To limit the parameter search space, and impose some regularization on the inferred model, a truncated Karhunen-Loéve (KL) series is constructed by applying a singular value decomposition on the covariance matrix of various realizations of Manning’s $n$ coefficients. The parameters are then updated through their coordinates in the reduced KL basis, instead of the large nodally defined parameter vector. The representation of the ensemble members in the KL
basis is expected to better preserve the geostatistical characteristics of the parameter field in the filter update steps [129, 130]. Finally, to enhance the filter’s performance and better deal with the nonlinear parameter estimation problem, we introduce iterations to the SEIK update steps as in [131] and [132]. Numerical experiments are conducted to evaluate the performance of the iterative SEIK against the EnKF in a realistic coastal configuration using the ADCIRC model.

The rest of this paper is organized as follows. The problem formulation is described in section 3.2. Section 3.3 summarizes the techniques used in our inference framework, including the sampling of multi-Gaussian realizations of the parameter field, the KL expansion, and the SEIK filter. Section 3.4 describes the details of the experimental setup. The experimental results, its significance, and implications are presented and discussed in section 3.5. A summary of the work and conclusions are given in section 3.6.

3.2 Problem formulation

3.2.1 ADvanced CIRCulation (ADCIRC) model

We use the ADvanced CIRCulation (ADCIRC) model, which solves the SWEs derived from the depth integration of the incompressible Navier-Stokes equations:

$$\frac{\partial H}{\partial t} + \frac{\partial}{\partial x}(Q_x) + \frac{\partial}{\partial y}(Q_y) = 0,$$

(3.1)
\[
\begin{align*}
\frac{\partial Q_x}{\partial t} + \frac{\partial U Q_x}{\partial x} + \frac{\partial V Q_x}{\partial y} - f Q_y &= -g H \frac{\partial [\zeta + P_x/g \rho_0 - \alpha \eta]}{\partial x} \\
& \quad + \frac{\tau_{sx}}{\rho_0} - \frac{\tau_{bx}}{\rho_0} + M_x - D_x - B_x, \\
\frac{\partial Q_y}{\partial t} + \frac{\partial U Q_y}{\partial x} + \frac{\partial V Q_y}{\partial y} - f Q_x &= -g H \frac{\partial [\zeta + P_x/g \rho_0 - \alpha \eta]}{\partial y} \\
& \quad + \frac{\tau_{sy}}{\rho_0} - \frac{\tau_{by}}{\rho_0} + M_y - D_y - B_y.
\end{align*}
\]

Here, \( \zeta \) is the free-surface elevation relative to the geoid, \( h \) is the bathymetric depth relative to geoid, \( H = \zeta + h \) is the water depth, \( U \) and \( V \) are the depth-averaged horizontal velocity components, \( Q_x = U H \) and \( Q_y = V H \) are the flux per unit width in the \( x \) and \( y \) directions, \( f \) is the Coriolis parameter, \( g \) is acceleration due to gravity, \( P_x \) is the atmospheric pressure at the free surface, \( \rho_0 \) is the reference density of water, \( \alpha \) is the Earth elasticity factor, \( \eta \) in the Newtonian equilibrium tide potential, \( \tau_{sx} \) and \( \tau_{sy} \) are the applied free surface stresses, \( \tau_{bx} \) and \( \tau_{by} \) are the bottom friction components, \( M_x \) and \( M_y \) are the vertically-integrated lateral stress gradients, \( D_x \) and \( D_y \) are the momentum dispersion, and \( B_x \) and \( B_y \) are the vertically-integrated baroclinic pressure gradients. In ADCIRC, the continuity equation is replaced by the second-order, hyperbolic generalized wave continuity equation (GWCE) to reduce spurious oscillations that occur in the original form. Manning’s \( n \) coefficients arise in the bottom friction terms of \( (3.2) \). The explicit expression of the bottom friction components are \( \frac{\tau_{bx}}{\rho_0} = \frac{K_{slip} Q_x}{H} \) and \( \frac{\tau_{by}}{\rho_0} = \frac{K_{slip} Q_y}{H} \). The coefficient \( K_{slip} = c_f |u| \), where \( c_f = \frac{g \eta^2}{H^{1/3}} \), represents a quadratic drag law. The scalar value, \( n \), is the Manning’s \( n \) coefficient. Since the Manning’s \( n \) coefficients spatially vary, they are defined node-wise within the discretized physical domain, and are a piece-wise linear representation of the continuous bottom friction field.

The SWEs in ADCIRC are discretized spatially using a first-order continuous Galerkin finite element method with unstructured triangular elements. The time derivatives in the GWCE are approximated with centered finite differences, and forward differences are used
for the time derivatives in the momentum equations. ADCIRC has been successfully implemented in many coastal ocean studies (e.g. [2, 3, 4, 16, 92, 103]).

To simulate tides in an estuarine system, we adopted the same domain as that of [5] and [27]. This selected domain is an idealized coastal inlet with an ebb shoal, with an open ocean boundary on the left and a reflective boundary (representing the wall along the coastline) on the right as shown in Figure 3.1. The domain is discretized into 1,518 grid nodes and 2,828 elements. Its dimension is 4500 m in the x-direction and 3000 m in the y-direction. Bathymetry is measured downward from the geoid to the ocean floor. The bathymetric depth increases linearly from 3.8 m at the open ocean boundary to 1 m at the mouth of the inlet on the west side of the domain. The shallowest area of the domain is on the mound in front of the west entrance of the inlet with a depth of 0.5 m below the geoid. The landlocked area has a constant bathymetry of 1 m. The diameter of the ebb shoal is 750 m. This configuration is considered to be a simplified version of a real-world ebb shoal system, which is a natural feature of many coastal ocean regions. We force ADCIRC by the $M_2$ tidal constituent with an amplitude of 0.25 m (relative to the geoid) and a 2 s time step.

3.3 Parameter Estimation Framework

This section describes the techniques that are used in our parameters inference framework. These include: 3.3.1 a sampling scheme and a search space representation (sequential simulation algorithm), 3.3.2 reduction of the search space (Karhunen-Loève (KL) expansion), 3.3.3 an ensemble filtering inference scheme (Joint-SEIK for parameter inference), and 3.3.4 an iterative technique in the filter update step (iterative SEIK).

3.3.1 Sequential simulation algorithm

The generation of spatially-dependent fields of various variables is useful for the numerical simulation of many problems in geophysical fluid dynamics [126]. Since the collected
Figure 3.1: Idealized inlet with ebb shoal domain. The discretization of the domain is represented. The first 15 observation stations used in the experiment are marked with red dots, and the additional 9 observation stations added later are marked with white. The color bar represents the bathymetry of the domain measuring down from the geoid (m).
data is often limited, one must resort to algorithms capable of generating realizations of a full variable field, subject to available data and a suitable covariance model. One of the well-established techniques to generate spatially variant maps is the so-called ‘sequential simulation algorithm’ \[133\]. This method recursively draws realizations of variables from a multivariate pdf modeled from series of univariate conditional pdfs that are constrained by available data. For variables following joint Gaussian distributions, the prescribed covariance model, mean, and variance of the field are needed in order to solve for a set of coefficients in a simple kriging system \[134\]. These are then used to calculate the mean and variance that characterize the conditional density function of each variable, given the set of conditioning data. The covariance model is given by \( \text{Cov}(h) = c - g(h) \) where \( h \) is the variable, \( g(h) \) is the corresponding semi-variogram model and \( c \) is its sill. In this study, the Manning’s field is assumed Gaussian and anisotropic, which can be sampled from Gaussian semi-variogram of the form

\[
g(h) = c \cdot \left(1 - \exp \left(-h^2\right)\right). \tag{3.3}
\]

Here \( h = \sqrt{(h_x/a_x)^2 + (h_y/a_y)^2} \), where \( h_i, i = x, y \), is the lag distance between two locations in the \( i \) direction and \( a_i, i = x, y \), is an appropriate range in the \( i \) direction.

### 3.3.2 Karhunen-Loève (KL) expansion

The KL expansion \[135, 136\], is a classical method for expressing stochastic processes as an orthonormal set of deterministic functions. It follows the result of Mercer’s theorem \[137\], which states that a symmetric positive definite matrix \( C(x_1, x_2) \) admits the spectral decomposition

\[
C(x_1, x_2) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x_1) \psi_k(x_2), \tag{3.4}
\]
where $\lambda_k > 0$ are the eigenvalues of $C$ and $\psi_k$ are the corresponding eigenvectors, i.e. the terms in (3.4) must satisfy

$$\int_{\Omega} C(x_1, x_2) \psi_k(x_2) dx_2 = \lambda_k \psi_k(x_1), \quad k = 1, 2, \ldots$$  \hspace{1cm} (3.5)$$

The sequential simulation algorithm described in subsection 3.3.1 produces realizations of a variable field with mean $\mu(x)$ and a discretization $C(x_1, x_2)$ of the covariance function $\text{Cov}(h)$, where $x \in \mathbb{R}^d$ is a vector of length $d$ of the nodes of a discretized domain. The covariance function is then decomposed according to Mercer’s theorem. Let $K(x, \xi)$ be a stochastic function of a coordinate vector $x$ and a random variable $\xi$. Every realization of $K$ can then be expressed as

$$K(x, \xi) = \mu(x) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k \psi_k(x).$$  \hspace{1cm} (3.6)$$

In the case of a multi-Gaussian field generated by the sequential simulation algorithm, $\xi_k$ is a Gaussian independent identically distributed random variable with zero mean and unit variance. The function $K$ is fully characterized by a set of $\xi_k$ when the basis $\psi_k$ is known. Given a realization of $K$, together with a known covariance matrix, the $\xi_k$ can be obtained by evaluating the integral

$$\xi_k = \int (K(x, \xi) - \mu(x)) \psi_k(x) dx.$$  \hspace{1cm} (3.7)$$

KL expansions represent highly spatially variant parameters as realizations of a stochastic process with only a few dominant modes by truncating the infinite series in (3.6) with a finite number of $N$ terms. The size $N$ essentially depends on the desired energy percentage to be retained by the KL modes $\sum_{k=1}^{N} \sqrt{\lambda_k} / \sum_{k=1}^{\infty} \sqrt{\lambda_k}$. This notion of “optimal” truncation is particularly useful for large scale parameter inference problems in order to alleviate
computational burdens while retaining the essential features of the inference space \[48\].

### 3.3.3 Joint-SEIK for parameter Inference

The Joint EnKF approach is widely used for parameter estimation by the subsurface modeling community (e.g. \[35\] \[44\] \[46\] \[97\] \[138\]). In the most general form, a vector of model parameters to be estimated, \( \mathbf{w} \), is appended to the system state vector \( \mathbf{x}_k \), to form the joint state-parameter vector

\[
\mathbf{z}_k = \begin{bmatrix} \mathbf{x}_k \\ \mathbf{w} \end{bmatrix}.
\]

Assuming stationary dynamics for the parameters, the augmented state-space model is then written as

\[
\mathbf{z}_k = \mathbf{M}(\mathbf{x}_{k-1}) \mathbf{w}_{k-1} + \mathbf{0},
\]

where \( \mathbf{M} \) is the dynamical operator describing the time evolution of the state vector from time \( k - 1 \) to time \( k \), and \( \mathbf{w}_k \) is the model error with Gaussian of mean zero and covariance matrix \( Q_k \). The observation \( \mathbf{y}_k \) is then related to the augmented state vector as

\[
\mathbf{y}_k = \mathbf{H}_k^T(\mathbf{z}_k) + \mathbf{e}_k = \mathcal{H}_k(\mathbf{x}_k) + \mathbf{e}_k,
\]

where \( \mathbf{H}_k \) is the linearized observation operator and \( \mathbf{e}_k \) the measurement noise.

Some studies pointed to some difficulties in estimating the model parameters with the ensemble Kalman filter \[139\] \[140\] \[141\], but many more presented quite successful implementations, e.g. \[27\] \[93\] \[96\] \[97\] \[142\] just to cite a few. Among the most reported issues were related to strong nonlinear relations between the observations and the estimated parameters \[71\] \[143\] \[144\], the relevance of the assimilated information \[34\], and the size of
the problem \[145\]. These were however not problematic in our particular setting and the filter performances were deemed quite satisfactory in our numerical experiments presented in section 3.5.

Here we follow \[5\] and \[52\], and implement the Singular Evolutive Interpolated Kalman (SEIK) filter, which was found to be particularly efficient at enhancing the predictive capabilities of ADCIRC \[52\] \[42\] and also for parameters estimation \[5\] \[27\]. Compared to the other deterministic EnKFs, SEIK involves a stochastic rotation in the resampling step to randomly spread the error variance in the ensemble space \[38\], which is suitable for strongly nonlinear dynamics that often arise during storm surges and was later suggested for the Ensemble Transform Kalman Filter (ETKF) \[146\] \[147\]. SEIK algorithm can be split in three steps; given an initial ensemble \((z_{a,i}^{0}, i = 1, ..., N)\).

**Forecast step**

The forecast step integrates the analyzed ensemble members, \(z_{a,i}^{k-1}\), with the model \(3.9\) to compute the forecast ensemble members, \(z_{f,i}^{k}\). One then takes the average of the \(z_{f,i}^{k}\) as the forecast state vector, \(z_{f}^{k}\), and their sample covariance as the forecast error covariance, \(P_{f}^{k}\). Assuming a perfect model \((Q_{k} = 0)\), one can decompose

\[
P_{f}^{k} = L_{k} U_{k-1} L_{k}^{T}, \tag{3.11}
\]

with

\[
L_{k} = \begin{bmatrix} z_{f,1}^{k} - z_{f}^{k} & \cdots & z_{f,r+1}^{k} - z_{f}^{k} \end{bmatrix} T, \tag{3.12}
\]

and

\[
U_{k-1} = [(r - 1)T^{T}T]^{-1}. \tag{3.13}
\]
Here \( T \) is an \((r+1) \times r\) full rank orthogonal matrix with zero column sums. When the model error is not neglected, SEIK accommodates the model error by adding its covariance matrix to the right hand side of (3.11). Its algorithm remains mostly unchanged. However in this case, \( P^f_k \) will not remain of low-rank \( r \), and re-approximating the forecast covariance matrix \( P^f_k \) will be required (91).

**Analysis step**

When a new observation \( y_k \) becomes available, The forecast state is updated to obtain the analysis state

\[
z^a_k = z^f_k + K_k(y_k^o - H_k z^f_k),
\]

(3.14)

where \( K_k \) is the Kalman gain

\[
K_k = L_k U_k (H L)^T_k R_k^{-1}.
\]

(3.15)

\((HL)_k\) is computed by applying \( H_k \) to the ensemble perturbations \( z^{f,i}_k - z^f_k \),

\[
(HL)_k = \begin{bmatrix}
H_k(z^{f,1}_k - z^f_k) & \cdots & H_k(z^{f,r+1}_k - z^f_k)
\end{bmatrix} T,
\]

(3.16)

and

\[
U_k^{-1} = \frac{1}{\rho} U_{k-1}^{-1} + (HL)_k^T R_k^{-1} (HL)_k.
\]

(3.17)

The inflation factor, \( \rho \), is used to inflate the forecast error covariance as a way to account for various sources of uncertainties in the system, e.g. model error, small ensembles, Gaussian assumption, etc [34, 38]. The analysis error covariance can be expressed as \( P^a_k = L_k U_k L_k^T \), but this is not needed for the filter algorithm.
Resampling step

New ensemble members need to be generated to start the next forecast cycle. These are sampled from the analysis mean and the covariance as

\[
\mathbf{z}_{k-1}^{a,i} = \mathbf{z}_{k-1}^a + \sqrt{N} \mathbf{L}_{k-1} (\mathbf{\Omega}_{k-1}^{-1} \mathbf{C}_{k-1}^{-1})^T, \quad i = 1, ..., N
\] (3.18)

where \( \mathbf{\Omega}_{k-1} \) is an \((r+1) \times r\) matrix with orthonormal columns and zero column sums generated using Householder matrices [38, 90]. In this study, we are only interested in estimating the parameters, i.e. \( \mathbf{w}_k \).

3.3.4 Iterative SEIK

Parameter estimation with an EnKF can suffer from strong nonlinearities between the observed state and the parameters [148]. Iterating on the parameter update step has been shown to improve the accuracy of the filter estimates [132, 149].

Let \( \mathbf{x}_{a,j}^k \) be the analyzed state at timestep \( k \) and iteration \( j \). The iterative SEIK (ISEIK) seeks the solution of the nonlinear least squares problem:

\[
\mathbf{x}_{a,j+1}^k = \mathbf{x}_{a,j}^k + \mathbf{K}_k (\mathbf{y}^o_k - \mathbf{H}(\mathbf{x}_{a,j}^k))
\] (3.20)

We see that this equation is the iterative form of (3.14). As the iterations advance, the inbreeding problem may cause the filter to increasingly

\[
\arg\min_{\mathbf{x}_{a,j+1}^k} \| \mathbf{y}^o_k - \mathbf{H}(\mathbf{x}_{a,j+1}^k) - \mathbf{J}_k^i (\mathbf{x}_{a,j+1}^k - \mathbf{x}_{a,j}^k) \|^2 + \| \mathbf{x}_{a,j+1}^k - \mathbf{x}_{a,j}^k \|^2 \mathbf{C}_k
\] (3.19)

where \( \mathbf{J} \) is the Jacobian matrix of \( \mathbf{H} \). The term \( \mathbf{H}(\mathbf{x}) - \mathbf{J}_k^i (\mathbf{x}_{a,j+1}^k - \mathbf{x}_{a,j}^k) \) is the first-order Taylor approximation of \( \mathbf{H}(\mathbf{x}_{a,j+1}^k) \) and \( \mathbf{C}_k \) is a symmetric, positive semidefinite matrix. The solution \( \mathbf{x}_{a,j+1}^k \) of (3.19) is derived in [150] as

\[
\mathbf{x}_{a,j+1}^k = \mathbf{x}_{a,j}^k + \mathbf{K}_k (\mathbf{y}^o_k - \mathbf{H}(\mathbf{x}_{a,j}^k))
\] (3.20)
underestimate the ensemble variance, ultimately degrading the filter’s performance [151].

This problem is more pronounced when the parameter and the state are strongly nonlinearly related. To this end, we adopt a strategy that limits the size of the update term in the later iterations via a damping factor $\omega_j$, as suggested by [45]. The factor $\omega_j$ takes values between 0 to 1 and multiplies the update term (increment to the forecast) in (3.20). This helps to smooth the perturbation of Manning’s $n$ coefficients, which alleviates the impact of the state-parameter nonlinear relation and sampling errors.

The iterative scheme (3.20) can be directly applied to SEIK with minor modifications. In particular, if $\mathcal{H}$ is linear (as in this study) and $\mathbf{C}_k$ is taken as the covariance matrix $\mathbf{P}_k$, then one only needs to iterate on $\left(\mathbf{H}\mathcal{L}\right)_k$ to derive the iterative SEIK scheme. Moreover, since here we only update the parameters and not the state, we only need to iterate on the parameter ensemble mean in (3.20), while maintaining the ensemble variance during each assimilation cycle $k$. In this study, the iterations are stopped when the updates become small or a maximum iteration number is reached. For more sophisticated stopping criteria, readers may refer to [152].

3.4 Experimental setup

3.4.1 Generating synthetic multi-Gaussian Manning’s $n$ fields

Synthetic data of the Manning’s $n$ coefficients are first generated by taking a small number of samples from the uniform distribution $U(0.005, 0.2)$ to simulate a scenario where a few point-wise Manning’s $n$ coefficients data are collected (or inferred from point-wise bottom surface characteristics). These data are assumed collected at 24 locations representing the observations stations as illustrated in Figure 3.1. The synthetic Manning’s $n$ coefficient data are then integrated with the public domain ANSI-C code ‘GCOSIM3D’ developed in [126], to generate multi-Gaussian 2D Manning’s fields for our idealized ebb shoal domain, based on the sequential simulation algorithm (section 3.3.1). From this, any number of
Manning’s field realizations can be generated once the properties of the semi-variogram are set. We first generate 1000 realizations of nodally-defined parameter fields following the Gaussian semi-variogram as in (3.3), with a mean of 0.1025, a variance of 0.0002, and a correlation range of 180 m in the x-direction and 30 m in the y-direction. The variance is properly scaled so that the realizations of 2D multi-Gaussian fields fall within an appropriate range of Manning’s $n$ coefficients (0.005 - 0.2). The maximum and minimum Manning’s $n$ values of these realizations are 0.1879 and 0.0177, respectively. Examples of realizations of Manning’s $n$ fields generated by the sequential simulation algorithm are shown in Figure 3.2. These realizations are used to select the initial ensemble, compute the KL modes, and define the reference field.

### 3.4.2 Observation System Simulation Experiments (OSSEs)

We let the 101st realization generated by GCOSIM3D code in section 3.4.1 be the reference Manning’s $n$ field, which we seek to infer (Figure 3.2). Synthetic observations of water elevation are generated by ADCIRC integrated using the reference Manning’s $n$ field. The dimension of the observations is the number of observed locations multiplied by the number of assimilation time steps. Initially, we use 15 observation stations as shown in Figure 3.1, and 108 assimilation timesteps (4.5 days with incoming data every 1 hour). Later, we also increase the number of observation stations to 24 and total assimilation time to 20 days in some experiments to study the impact of the number of observations on parameter inference. Two-hundred Manning’s $n$ field realizations, excluding the reference realization, are taken as the initial members. We first let the simulation ramp-up for 12 days using the mean of the initial members before the first assimilation cycle starts. The observations are then assimilated by SEIK to infer the reference Manning’s field. We test the filter with four different settings: 1) nodally-defined Manning’s $n$ values, 2) Manning’s $n$ field parameterized by the KL-reduced space, 3) Manning’s $n$ field inference in the KL space with perturbed variogram models, and 4) iterative SEIK filter inferring the Manning’s field
Figure 3.2: A few realizations of Manning’s $n$ fields generated by the sequential simulation algorithm. The 101$^{st}$ realization is taken as the reference field that the inference results are compared against.
in both the full and reduced space.

### 3.4.3 KL basis construction

The sample covariance of the 1000 realizations generated in subsection 3.4.1 is decomposed as in (3.4) to obtain the set of eigenpairs and the KL expansion of the parameter (i.e., the 2D Manning’s $n$ coefficients) as in (3.6). The cumulative sum of eigenvalues, which indicates the total variance retained by the KL expansion is plotted at the bottommost of Figure 3.3. It shows that retaining 10 and 20 KL terms respectively preserve more than 83% and 98% of the total variance of the realizations. By increasing the truncation to 30 KL terms, more than 99% of the total variance is retained.

Figure 3.3 also shows an example of the reconstruction of a Manning’s field using a truncated KL expansion. The top row of the figure shows the mean of the 1000-realizations of the Manning’s $n$ field and the 101$^{st}$ realization, respectively. The remaining subfigures are the reconstructions of the 101$^{st}$ realization as we increase the number of KL terms in (3.6). One can observe that with a small number of KL terms, for example, 3 KL modes, the reconstructed field resembles the mean, as the mean field dominates the modes. As we increase the number of KL terms, the reconstruction starts converging toward the target realization (the 101$^{st}$ realization in this figure).

### 3.5 Results and discussion

#### 3.5.1 SEIK inference of nodally defined Manning’s $n$ values

Figure 3.4(a)-(d) present the results of the Manning’s $n$ field SEIK inference using only 10 ensemble members compared to the reference field. The impact of the SEIK updates is clear from the final analysis, as the filter solution more accurately represents the reference Manning’s $n$ field. The 2D plot of the ratio between the final error and the initial error suggests reasonably small errors in most locations, except those where the Manning’s $n$
Figure 3.3: The reconstruction of a Manning’s $n$ field using the truncated KL expansion for various retained KL modes. The top six subplots: spatial plots of the reconstructed Manning’s $n$ field. The bottommost subplot: the accumulative sum of the eigenvalues obtained for the KL decomposition.
values of the initial ensemble vastly differ from the true values. With only 10 ensemble members, this set of results is considered as a preliminary test, upon which we make efforts to improve.

We then applied Local Analysis (LA) in an attempt to improve the SEIK filter performance [153], later referred to as ‘local SEIK’. This technique provides a straightforward way to cut the spurious long-range correlations in the covariance matrix of the filter’s analysis step. In Figure 3.5, we show the time-series of Root Mean Square Errors (RMSEs) of the analysis with respect to the reference field for various localization distances (in meters). Localization enhances the filter’s performance in most cases, although the filter’s behavior with 10 members is quite sensitive to the choice of the localization distances (ld); the smallest RMSE at the end of the assimilation window is attained using ld = 1500 m.

In Figure 3.4(e) and (f), we show the impacts of applying localization to the SEIK filter. There is a clear improvement compared to those obtained without localization. The recovery of the Manning’s n coefficients around the right side of the inlet and the top left corner of the domain is notably improved. The ratio between the final error and the initial error are close to zero in most areas, except the areas where the difference between the Manning’s n values of the initial ensemble and the reference values were sizable.

Sensitivity to ensemble size

Increasing the ensemble size is generally expected to enhance the performance of an EnKF [38, 107, 125]. In [27], increasing the ensemble size from 10 to 100 drastically improved the estimation of a 1D Manning’s n coefficient in ADCIRC. However, this raises the issue of determining a good trade-off between filter performance and computational costs. Doubling the size of the ensemble means twice as many model runs are required. In the case of a complex model such as ADCIRC, this can result in a tremendous increase in computational time. Our first aim is to determine the ensemble size that yields satisfying filter performance with reasonable computational cost.
Figure 3.4: The results of SEIK inference with 10 ensemble members and localization, (a) the true field, (b) the initial ensemble mean, (c) the final analysis after 108 updates, and (d) the ratio between the final error and the initial error.
Figure 3.5: Time series of the RMSEs of the Manning’s $n$ fields after each analysis step with respect to the true Manning’s $n$ field for varying localization distance ($ld$).

We thus assess the filter’s performance with increasing numbers of ensemble members: 10, 50, 100, and 200, respectively, using the same localization distance ($ld = 1500$ m). The best localization distance in term of reducing the RMSE can be dependent on the ensemble size, however, we found that the localization length scale of 1500 m provides the lowest RMSE for most of our experiments (the top panel of Figure 3.6). More in-depth discussions on the choice of the localization length scale can be found in [153]. The time-series RMSE results of these runs, including the runs from subsection 3.5.1, are shown in the bottom panel of Figure 3.6. Here we see that increasing the number of ensemble members to 100 greatly reduces the discrepancy between the estimates and the reference field. However, increasing the ensemble size beyond 100 members does not significantly boost the performance of the filter, although it drastically increases the computational cost.

Figure 3.7 (c) and (d) summarize the results obtained by implementing local SEIK with
Figure 3.6: Time series of the RMSEs of the Manning’s $n$ field after every analysis step. Top panel: the RMSEs for varying localization distances (ld) with 100 members. Bottom panel: the RMSEs for various ensemble sizes (N) with the same localization.
100 ensemble members. Improvements over the case with 10 ensemble members are clear. The analysis at the end of the assimilation window accurately recovers the 2D Manning’s \( n \) coefficients at the right side of the inlet. The pattern of small Manning’s \( n \) values around the upper-right corner of the domain is also well recovered compared to the case with only 10 members. With this improvement, the ratios of the final to the initial error are close to zero and less than one in most areas, indicating that the local SEIK solution converges toward the reference field at almost every point in the domain.

**Sensitivity to the number of observations and assimilation cycle**

In general, it is preferable to assimilate as many observations as possible to compute reliable estimates. Here we explore the behavior of the system with an increasing number of observations, both spatially and temporally. We first introduce 9 additional observation stations (indicated with white dots in Figure 3.1) to the domain, increasing the total number of observation stations from 15 to 24. The observations locations are sampled to evenly span the spatial domain. We also increase the simulation time to 20 days, which equates to 468 total assimilation cycles.

Figure 3.7(e) and (f) outline the results of this experiment, where 100 ensemble members are used. The SEIK filter successfully recovers most of the features of the Manning’s \( n \) coefficients shown in the reference field. The node-wise ratios between the final error and the initial error are small and close to zero in most areas. The pattern of low Manning’s \( n \) coefficients in the right land-locked area to the left area near the open ocean area is almost fully recovered. The only area where there is difficulty recovering the Manning’s \( n \) features is the bottom-right corner of the domain. This can be attributed to the absence of observations in this area. In addition, we analyze the misfit between the filter estimate and the truth in Figure 3.7(g) in relation to the predicted variance of the error as estimated by the ensemble standard deviation (STD) in Figure 3.7(h), as resulting from the filter. Overall, both statistics are of the same order despite relatively larger STDs along eastern and
northern boundaries. The plots further reveal similar spatial structures, e.g., large error and STD values at the bottom-right corner of the domain (highlighted in red) where the observations are scarce, contrasting with small errors and STD around the center of the open ocean area (highlighted in dark blue), where the observations are more abundant. Similar consistencies between the final (misfit between the truth and final estimate) and predicted (filter error variance) estimation errors were obtained in the rest of our experiments, indicating that with large enough ensembles and good tuning of the localization radius, the estimation of Manning’s $n$ coefficients with the EnKF does not suffer from any divergence problem in our particular setting.

In Figure 3.8 we see the time-evolution of the RMSE of the estimates with respect to the reference field, based on three different implementations of SEIK with 100 ensemble members: regular SEIK, local SEIK, and local SEIK with additional observations and assimilation cycles, respectively. The discrepancy between the estimate and the truth visibly decreases as more observations are assimilated into the system, with a decreasing RMSE trend that suggests further improvements might be obtained with more assimilation cycles. Another conclusion one can draw from the time-evolution of the RMSE is that the filter does not really benefit in terms of estimation accuracy from localization when implemented with 100 ensemble members. Hereafter, we will consider the regular SEIK solution with 100 members as a reference to evaluate the performance of various tested filtering schemes.

### 3.5.2 SEIK inference in KL space

Instead of using SEIK to update the nodally-defined parameter, here we update the KL coefficients, $\xi$, that represent a specific realization of the Manning’s $n$ field in the KL space, using the same filtering procedure for parameter estimation described in subsection 3.3.3. The number of KL coefficients to be updated by the filter is the number of terms retained in the KL expansion. Here, we study the sensitivity of the performance of SEIK for parameter estimation in the KL space, later referred to as SEIK-KL, to both the number
Figure 3.7: The results of SEIK inference with 100 ensemble members, (a) the true field, (b) the initial ensemble mean, (c) the final analysis with 15 observation points, (d) the ratio between the final error and the initial error, (e) the final analysis with 24 observation points and 468 assimilation cycles, and (f) the ratio between the final error and the initial error, (g) the absolute error between the estimate and the truth for 24 observation points case, and (h) standard deviation of the ensembles at the final analysis step.
Figure 3.8: Time series of the RMSEs of the Manning’s $n$ fields after each analysis step for 100 ensemble members, different localizations (ld) and different number of observations with respect to the true Manning’s $n$ field.
of retained terms and the ensemble size.

In Figure 3.9, we plot the time-evolution of the RMSE of the analyzed Manning’s $n$ field with varying ensemble size and the number of KL terms. Each individual curve represents the RMSE of a single SEIK run with a fixed ensemble size and a specific number of preserved KL terms. The left column of subfigures (Figure 3.9(a), (b) and (c)) show the RMSE of SEIK-KL using different numbers of KL terms for a specified ensemble size. Conversely, the right column of subfigures (Figure 3.9(d), (e) and (f)) show the RMSE for a specified number of KL terms and varying ensemble sizes. First, we examine the results of SEIK inference with 10 ensemble members (Figure 3.9(a)). In all cases, SEIK-KL efficiently reduces the RMSE over time and leads to better final estimates than the regular SEIK. This suggests that the ensembles in the KL space exploit the statistical information retained by the KL modes to better span the parameter search space as compared to the full space spanned by limited ensembles.

The convergence rate of SEIK-KL estimates to the truth is sensitive to the number of retained KL terms. For instance, when using 10 KL terms, the analysis converges rapidly toward the solution but quickly levels off after a few assimilation cycles. Increasing the simulation time does not improve the estimates when the ensemble size and number of retained KL terms are small. This is because a few KL terms are not enough to completely describe the variability of the search space. The filter stops improving after a few assimilation cycles due to the limited search directions, as also observed in [129]. Increasing the ensemble size in this case does not help much as a relatively small ensemble (often suggested to be of rank equal to the search space [90]) should be enough for efficient filtering. When the number of KL terms is increased (e.g., to 20 and 30 terms), the stagnant RMSEs pattern in the previous case is less pronounced, and the analysis starts to converge gradually, but slowly, toward the reference solution. Including more KL terms enables more search directions in the parameter space to be explored. As a result, more assimilation cycles may help SEIK-KL recover the reference field. Given a sufficiently large assimilation
window, the SEIK-KL with 20 and 30 KL terms outperforms that of 10 KL terms.

Increasing the ensemble size (Figure 3.9(b) and (c)), further reduces the RMSEs. However, the difference is not significant in the case of 10 KL terms. SEIK-KL inference with larger numbers of retained KL terms outperforms the cases with fewer KL terms for larger ensemble sizes. Figure 3.9(b) shows that using 50 ensemble members, SEIK-KL with 30 KL terms starts to outperform the 10 KL-terms case at the end of the assimilation window. When 100 members are used (Figure 3.9(c)), SEIK-KL with 20 and 30 KL terms leads to notably better estimates than those obtained using 10 KL terms. Due to less inherent variability, SEIK-KL with 20 KL terms performs poorer than the SEIK-KL with 30 KL terms for all tested ensemble sizes (Figure 3.9(a),(b) and (c)). Also, SEIK with the full, nodally-defined parameter vector outperforms the SEIK-KL with 10-KL terms. Applying regular SEIK using 100 members leads approximately to the same level of RMSE as that of the best KL case (i.e., the 30-KL-terms case).

The sensitivity of the performance of SEIK-KL inference to the ensemble size is presented in the three plots in the right column of Figure 3.9. In general, we see that as the ensemble size increases, the RMSE decreases, with an exception of the 10-KL terms case shown in Figure 3.9(d); the RMSE produced by SEIK-KL using 50 members is smaller than that using 100 members. Again, this is the effect of using a few KL terms, which insufficiently describe the search space. This observation is consistent with [129] and [48], who found that using 40 KL modes, small ensembles initially performed better, but are eventually outperformed by larger ensembles later in the simulation, as larger ensembles provide more exhaustive search directions.

In Figure 3.10, we show the spatial plots of the inferred Manning’s $n$ coefficients using SEIK-KL with varying ensemble sizes and numbers of KL modes. The top row depicts the spatial structure of the true Manning’s $n$ field (Figure 3.10(a)) and the initial guess (Figure 3.10(b)), respectively. From the second row downward, each column represents an ensemble size and each row represents a number of retained KL modes. We notice that when
Figure 3.9: Time series of the RMSEs as results from SEIK and SEIK-KL using varying ensemble sizes and numbers of retained KL terms. Left column: each figure represents a fixed ensemble size but varying number of retained KL terms. Right column: each figure represents a fixed number of retained KL terms but varying ensemble size.
using 10 ensemble members, the SEIK filter faces difficulty in recovering the reference field, even with a large number of KL terms; the best result is obtained using 10 KL modes (Figure 3.10(d)). When using 50 ensemble members, all SEIK-KL inferences are better than the regular SEIK in recovering the Manning’s $n$ field. This is particularly clear in the area of low Manning’s $n$ values (cooler colors). When using 100 ensemble members, the filter’s estimate is more accurate in all cases. The main Manning’s $n$ structures of the true parameter field are clearly recovered.

### 3.5.3 SEIK-KL sensitivity to inaccurate covariance model

The initial ensembles of the SEIK and the SEIK-KL have thus far been constructed based on the same covariance model from which the reference Manning’s $n$ field was generated. In many real-world applications, however, the initial covariance model might be poorly known. Here we examine the sensitivity of the performance of SEIK and SEIK-KL to perturbations in the covariance model used to generate both the initial ensembles of Manning’s $n$ coefficients and KL modes and compare the results against those obtained using the true (unperturbed) covariance model.

We first generate new realizations of Manning’s $n$ coefficients from different variogram models by perturbing some parameters in GCOSIM3D. 1) we use a Gaussian variogram with a range of 100 m in the x and y-directions (a perfect circle), 2) a Gaussian variogram with a range of 250 m in the x-direction and 15 m in the y-direction (i.e. the reference variogram is stretched in the x-direction and shrunk in y-direction), 3) a Gaussian variogram with range of 110 m in the x-direction and 45 m in the y-direction (i.e. the reference variogram is shrunk in x-direction and stretched in y-direction), and 4) an Exponential-type variogram.

In Figure 3.11 we plot the time-series of the RMSE of the analyzed Manning’s $n$ field as estimated by SEIK and SEIK-KL (with 30 KL modes) from the covariance models described above using 100 ensemble members. The first observation we make is that the
Figure 3.10: Manning’s $n$ field estimates as inferred by different ensemble sizes and number of retained KL modes
performances of SEIK and SEIK-KL degrade when using any of the perturbed variograms. However, the degree at which the final RMSEs of the perturbed covariance cases differ from the reference case depends on the form of the perturbed covariance. For instance, using an Exponential variogram instead of Gaussian variogram (with the same mean, variance and correlation length) does not significantly alter the structure of the variogram, and as a result, the final RMSEs for the Exponential variogram cases are close to the reference case. The same can be said for the case in which we stretch or shrink the correlation range in the x- and y-directions (cases (2) and (3)). For the case where the perturbed variogram vastly differs from the reference case (i.e. case (1)), the resulting final RMSE is considerably larger than that of the reference case. The second observation is that in most cases with perturbed covariance models, the SEIK performs better than SEIK-KL, with the exception of case 3. The overall better performance obtained by SEIK is consistent with the results of section 3.5.2, where the SEIK-KL was shown to outperform SEIK with small ensembles only (N < 50).

In Figure 3.12, we show the spatial plots of inferred Manning’s n fields as estimated by the SEIK and SEIK-KL for different covariance models. The results suggest that for all tested covariance models, the filter successfully recovers the main patterns of the true Manning’s n coefficients over the studied domain, even for the perfect circle variogram case (1) where the inferred field exhibits the largest RMSE compared to the other cases. With sufficiently large ensembles, assimilated observations, and retained KL terms, SEIK-KL is capable of successfully capturing the main spatial structures of the reference Manning’s n field, even when its reduced basis is constructed with imperfect KL-modes.

3.5.4 Iterative SEIK (ISEIK) in the full and KL spaces

Based on the above results, ISEIK is implemented using 100 members, 24 observation stations, and 5 days of simulation. The damping factor $\omega$ is chosen such that $\omega_{j+1} = \omega_j / 2$, where $j$ is the iteration number, and $\omega_0 = 1$. We start by studying the sensitivity of ISEIK to
Figure 3.11: Time series of the RMSEs of the Manning’s $n$ fields inferred from initial ensembles generated from various variogram models.

The number of iterations by performing 3, 5 and 7 iterations. The results of this experiment in terms of RMSE are presented in Figure 3.14. ISEIK outperforms SEIK in all cases, notably reducing the RMSE for both the full-vector and the KL cases. The lowest RMSEs were obtained with 5 iterations.

In Figure 3.13, we show the spatial plots of the estimates obtained using ISEIK. We notice particularly improved parameter recovery in the area of low Manning’s $n$ values compared to the regular SEIK. ISEIK also greatly reduces the ratio between the final error and the initial error in the right land-locked area. ISEIK in the full or KL spaces performs equally well in terms of reducing RMSE in all cases (Figure 3.14). However, ISEIK-KL seems to better recover the spatial patterns of the reference Manning’s field. This can be clearly observed in the area of low Manning’s $n$ values, colored in green (Figure 3.13(e)): the recovered Manning’s $n$ structure as estimated by ISEIK in the KL space is more consistent with the reference than those produced in the full space. The computational cost
Figure 3.12: Inferred Manning’s $n$ fields when various variogram models are used to generate the initial ensembles in KL space.
of ISEIK is approximately the same as the regular SEIK. Thus, only modest increases in computational cost, ISEIK with a well-tuned damping factor performs comparably to the regular SEIK when using a much larger simulation window; with only 5 days of assimilation time, the final RMSE produced by ISEIK is as small as that obtained with the SEIK over 20 days of assimilation window.

3.6 Conclusions

We proposed a sequential data assimilation framework to estimate a 2D field of spatially varying Manning’s $n$ coefficients in the context of coastal ocean modeling. The proposed framework combines a deterministic ensemble Kalman filter (called SEIK), KL decomposition, and an iterative update scheme to improve the accuracy of estimation over an unaltered/baseline SEIK filter. Multi-Gaussian initial realizations of the Manning’s $n$ coefficients field are generated using a sequential simulation algorithm. An empirical covariance matrix is computed from a sufficiently large number of realizations of Manning’s $n$ fields and used to construct KL coordinates representing the parameter in a reduced KL space. The KL expansion enhances the parameter search space and helps preserve the geostatistical characteristics of the parameter in the filter updates when the filter is implemented with a small number of ensembles.

Observation System Simulation Experiments (OSSEs) are conducted to evaluate the performance of the proposed framework. Synthetic water elevation data are generated by running ADCIRC with a reference Manning’s $n$ field, considered as the truth. SEIK is then implemented to estimate the Manning’s $n$ coefficients, both in the full nodally-defined and KL parameter space cases. We first study SEIK sensitivity to the ensemble size using the full parameter space and find that 100 ensemble members provide a reasonable trade-off between the filter performance and computational burden. Local analysis is also applied to alleviate the effect of spurious correlations between distant points. Increasing the number of observation stations from 15 stations to 24 stations further improves the filter performances.
Figure 3.13: Inferred Manning’s $n$ fields. 1$^{st}$ row: spatial plots of regular SEIK inference, 2$^{nd}$ row: spatial plots of ISEIK inference with 5 iterations, 3$^{rd}$ row: spatial plots of ISEIK-KL inference with 5 iterations.
SEIK with the full nodally-defined parameter vector proves to be successful at recovering the main patterns of the true Manning’s $n$ field in our idealized setting.

We then conduct the SEIK inference in the KL space. For small ensembles (e.g., 10 ensembles) and only 10 terms in the KL expansion, a significant improvement is observed compared to the results obtained using the regular SEIK filter. We also find that increasing the ensemble size requires increasing the number of KL terms in order for the KL-SEIK to outperform the regular SEIK. For the case with 100 ensemble members, 30 KL modes are required. However, the sensitivity of the filter performance to the number of KL modes and the ensemble size is nonlinear. In all cases, the KL-SEIK consistently outperforms the regular SEIK when the Manning’s $n$ field is represented using 30 KL terms, which preserves almost 100% of the total variance of the parameter space.

Finally, iterative SEIK (ISEIK) is implemented at almost no additional computational cost to enhance the SEIK performances. We apply ISEIK to both the nodally-defined pa-
rameter vector and KL cases. Even with a small number of iterations (e.g., 3 iterations), improvements are clearly observed, with the best results obtained using 5 iterations for both the nodally defined and KL cases were.

Overall, our results demonstrate the relevance of sequential ensemble data assimilation filtering schemes for estimating spatially varying parameters in the context of coastal ocean modeling. Future work will focus on exploring approaches to further improve our inference framework by developing efficient schemes to update and evolve in the KL basis of the parameter search space based on incoming data along the method proposed by [69].
Chapter 4

Bayesian Inference of Spatially-Varying Manning’s $n$ Coefficients in the Coastal Ocean Using a Generalized Karhunen-Loève Expansion and Polynomial Chaos

Building upon the results of the previous two chapters, the Bayesian inference of the Manning’s $n$ field will be further investigated. We are interested in improving the large dimensional parameter estimation by the mean of coordinate transformations of the KL expansion and PC-MCMC. The idea is to take into account the uncertainty of the hyper-parameters of the covariance model as part of the inference framework, which are traditionally assumed to be known \textit{a priori}. The generalized Karhunen-Loève expansion will be employed to construct the reduced parameter search space with uncertain hyper-parameters. PC with similar coordinate transformations will also be applied to accelerate the MCMC likelihood sampling. With this framework, Manning’s $n$ field and hyper-parameters posteriors will be inferred simultaneously, as opposed to conventional inference methods whose covariance hyper-parameters are assumed to be known. This development will contribute to a new approach for uncertainty reduction of spatially-varying parameters inference in the coastal ocean.

4.1 Introduction

Advanced coastal ocean models based on the shallow water equations (SWEs) have been widely utilized to simulate and predict shallow water circulations [6]. The SWEs are based on the simplified Navier-Stokes equations, which assume hydrostatic pressure and rela-
tively large horizontal length scales with respect to vertical length scales [12]. Even when the model is assumed to be perfect in describing the underlying physical processes, the model solutions are always subjected to uncertainties that emerge on account of errors originating from numerous sources [14], e.g., initial conditions, model parameters, forcing fields, and resolution of the implemented numerical methods. To provide robust and reliable forecasts of the ocean states, uncertainty quantification and reduction techniques have, in recent years, played a major role toward enhancing state-of-the-art coastal ocean simulation systems [5, 14, 15, 68, 111, 154, 155].

This work considers the uncertainty reduction of the model parameters in the coastal ocean, focusing on the inference of the Manning’s $n$ coefficient of roughness, introduced in the SWEs through the bottom stress components in the momentum equation [6, 16]. The Manning’s $n$ coefficients describe the bottom surface characteristics, which further define the resistance of water flow. [5] demonstrated that water elevation is highly sensitive to the change in this parameter, particularly in shallow coastal areas. The Manning’s $n$ coefficients are highly variable in space but cannot be measured directly [13], and as such often empirically derived. To infer such a parameter from the limited available observations, one may resort to Bayesian inversion [17, 18, 19, 20]. The Bayesian inference framework views the parameters estimation problem as the inference of a probability density function (pdf). A prior pdf is then updated into a posterior pdf by incorporating the likelihood of the observations [21].

Several well-established approaches to Bayesian inference are widely used in parameters inference problems, the most prominent of which is the Markov chain Monte Carlo (MCMC) method [23, 24, 26], which directly samples the posterior of the inferred parameters given the sample points of likelihood. A large number of model runs is required for MCMC to efficiently sample the posterior distribution [15, 28]. This renders the MCMC implementation computationally prohibitive for large-scale systems. To alleviate this hurdle, model reduction techniques have been exploited, both to build inexpensive surrogates
of the forward models and to reduce the dimension of the parameter estimation problems. Of particular interest is the spectral projection method used with the polynomial chaos (PC) basis proposed by [156]. The PC method expresses the dependence of the model predictions on the model parameters using orthogonal stochastic polynomials. This spectral expansion enables the construction of extremely cost-effective surrogate models that can be used as substitutes of the full model predictions to efficiently sample the likelihood [14, 81]. The PC method has been successfully implemented with a wide scope of large-scale problems, including mechanical systems (e.g., [75, 76, 77]), chemical systems (e.g., [78, 79, 80]), and more recently, large-scale oceanic systems (e.g., [15, 27, 68, 29]). Nevertheless, the PC method suffers from the ‘curse of dimensionality’ [68], in which only a limited number of stochastic parameters can be accommodated to build the PC surrogate before it becomes computationally prohibitive. To alleviate this issue, one may implement a parameter search space reduction technique, namely, Karhunen-Loève (KL) expansion [135, 136]. This technique expresses stochastic processes in terms of an orthonormal set of eigenfunctions, which can be used to parametrize highly spatially variant parameters with only a few dominant modes. Marzouk et al. [157] demonstrated the efficiency of combining PC and KL schemes to infer the spatially varying diffusivity of a medium in the context of a transient diffusion equation.

Sequential Bayesian inference method, such as the ensemble Kalman filter (EnKF) [34, 37, 38], has also been intensively utilized for parameter estimation of ocean models (e.g., [52, 51, 55, 70]). The EnKF has been found efficient, with advantages over the MCMC approach in accommodating large state-parameter vectors at reasonable computational cost [5, 27, 155]. Recently, the EnKF has been successfully used for the inference of a spatially varying Manning’s $n$ field, both in the full and reduced KL space [155]. The main merit in using an EnKF is its ability to accommodate larger member of parameters or KL modes. However, exploring the posteriors of these coordinates is limited to the directions defined by the subspace of the prior ensemble covariance model [48].
describing the Gaussian process needs to therefore be well known a priori.

In both MCMC- and EnKF-based inference, presented in [157] and [155], respectively, the underlying Gaussian process is associated with prior covariance models that are often poorly known. This is because a prior covariance model that describes the stochastic parameter field is subjected to the uncertainty in hyper-parameters, e.g., prior variance and length scales. An early attempt to include the hyper-parameter uncertainty in Bayesian inference problems was proposed by Tagade and Choi [158]. In this study, a KL expansion of a stochastic process is derived as a function of the hyper-parameters. The hyper-parameters are then cast as stochastic variables in the PC expansion, which can then be directly estimated alongside the KL coordinates via MCMC. However, this method can become computationally prohibitive due to the requirement in building the PC surrogate with increasing augmented stochastic dimensions.

In this work, we adopt the generalized KL expansion with the coordinate transformation approach introduced in [69] to account for the covariance hyper-parameters in a PC-based MCMC inference of spatially varying Manning’s $n$ coefficients of a coastal ocean model. This approach only requires building the PC surrogate based on a KL basis decomposed from a reference covariance matrix. A change of basis methodology is then utilized to transform the KL coordinates based on one basis to the a different basis. A similar coordinate transformation is finally applied to the KL expansion of the reference basis to obtain the PC expansion for any other basis required for the inference. Here, we avoid introducing the hyper-parameters as additional stochastic dimensions to the PC expansion. However, since the KL decomposition required to build the coordinate transformation matrix proposed in [69] for any set of hyper-parameters can be quite costly and we further introduce, in this study, the construction of the PC surrogate of the coordinate transformation matrix. This present contribution is essential for the application of a generalized KL expansion for any large-scale fluid system, in which the prior covariance matrices are very large.

The rest of this paper is organized as follows. Section 4.2 presents our Bayesian infer-
ence framework. Section 4.3 describes the change of coordinates methodology for incorporating hyper-parameters uncertainty to KL and PC expansions. The experimental setting for the evaluation of our proposed inference framework is outlined in Section 4.4. In Section 4.5, the results of the inference of the Manning’s $n$ coefficients coastal ocean model are presented and discussed. Concluding remarks are finally offered in Section 4.6.

4.2 Bayesian inference framework

Our objective is to estimate spatially varying parameters of a coastal ocean model given a set of (indirect) observations $d \in \mathbb{R}^{N_o}$, $N_o \geq 1$. The Manning’s $n$ roughness parameters, for example, can be viewed as a deterministic field $m(x)$, $x \in D$. We also assume that the observations can be predicted using the forward model (in our case the SWEs), which takes $m(x)$ as input and output, the model predictions of the observations. Bayes’ rule then updates the prior probability distribution function (pdf) of the unknown parameters $m$, conditioned on the discrepancy between the model prediction $u(m)$ and observations $d$ as \[82, 83\]:

$$p(m, \sigma_o^2 | d) \propto p(d | m, \sigma_o^2) p_m(m) p_o(\sigma_o^2),$$

(4.1)

where $p(m, \sigma_o^2 | d)$ is the posterior pdf, $p(d | m, \sigma_o^2)$ is the likelihood function of obtaining observations $d$ given the parameter field $m$ and the observations error model hyper-parameter $\sigma_o^2$, with priors $p_m(m)$ and $p_o(\sigma_o^2)$, respectively. In this work, we consider for simplicity an unbiased additive Gaussian error model, $\epsilon \sim d - u(m)$, $\epsilon \sim \mathcal{N}(0, \sigma_o^2 I_{N_o})$, where $\mathcal{N}(0, \sigma_o^2 I_{N_o})$ denotes the multivariate Gaussian distribution with zero mean and diagonal covariance $\sigma_o^2 I_{N_o}$.

We resort to the adaptive Metropolis-Hastings MCMC algorithm \[86, 159\] to efficiently sample the posterior pdf of the Manning’s $n$ parameters $m$. Metropolis-Hastings algorithm iteratively generates a Markov chain based on the accept/reject methodology, in
which the state of the chain at a given iteration depends only on the previous one. The high-dimensional nature of the Manning’s $n$ field discretized as the computational mesh, however, renders the node-wise parameter inference using MCMC computationally prohibitive. A model reduction technique, which statistically describes the Manning’s $n$ field as a stochastic process, parametrized using a few random variables, is thus required.

4.2.1 Karhunen-Loève (KL) expansion

The Manning’s $n$ roughness parameters physically characterizing the seabed surface are generally continuous in space, with correlated neighboring values. It is therefore effective to describe it as a stochastic process, $M(x, \omega)$, where $M$ is a function defined on the product space $\mathbb{D} \times \Omega, x \in \mathbb{D}$ is the index set of the location of the physical domain, and $\omega \in \Omega$ is the sample space of the probability space, $(\Omega, \mathcal{F}, P)$. Given a stochastic process $M(x, \omega)$, with mean $\mu(x)$ and symmetric positive semi-definite covariance function $C(x,x')$, the Mercer theorem [137] states that $C$ can be decomposed as:

$$C(x,x') = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(x'),$$  \hspace{1cm} (4.2)

where $\lambda_k > 0$ are the eigenvalues of $C$, and $\phi_k$ are the corresponding normalized eigenvectors. The eigenpairs satisfy the Freholm equation of the second kind:

$$\int_{\Omega} C(x,x') \phi_k(x') dx' = \lambda_k \phi_k(x), \hspace{1cm} k = 1, 2, \ldots$$ \hspace{1cm} (4.3)

Sequencing the eigenvalues in decreasing order, any realization of $M$ can then be written as

$$M(x, \omega) = \mu(x) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \phi_k(x) \eta_k(\omega).$$ \hspace{1cm} (4.4)
The coordinate coefficients $\eta_k(\omega)$ can be deterministically computed by evaluating the integrals

$$
\eta_k(\omega) = \int (M(x, \omega) - \mu(x))\phi_k(x)dx.
$$

(4.5)

In case $M$ is a Gaussian process (GP), $\eta_k(\omega)$ are Gaussian-independent identically distributed random variables with zero mean and unit variance, such that $\mathbb{E}[\eta_k \eta_k'] = \delta_{kk'}$. The parameter field $M$ is then fully characterized by a set of coordinates $\eta_k$’s given the eigen-basis is known. By truncating the series in Equation (4.4) to retain the first few $K$ terms, the KL expansion efficiently approximates the high-dimensional parameter field using a few dominant modes. The size of the retained terms $K$ depends on the desired energy percentage to be retained by the KL expansion, which is defined as $\sum_{k=1}^{K} \lambda_k / \sum_{k=1}^{\infty} \lambda_k$ [48]. In practice, it is preferable to choose $K$ such that the truncated KL expansion encapsulates as much information as possible with respect to its infinite counterpart. Larger $K$ are required for prior covariance functions with smaller correlation lengths for the KL expansion to capture a similar information percentage.

The truncated KL expansion is optimal, in the mean square sense, in approximating stochastic processes [76]. It is used as the model reduction technique for Bayesian inversion. The problem of inferring the continuous field $M$ from a few discretely collected observations is then reduced to the inference of the KL coordinates vector $\eta$, written as

$$
p(\eta, \sigma_o^2 \mid d) \propto p(d \mid \eta, \sigma_o^2)p_{\eta}(\eta)p_o(\sigma_o^2),
$$

(4.6)

where $p_{\eta}(\eta) = \exp(-\eta^T \eta / 2)/(2\pi^{K/2})$ is the Gaussian prior of the KL coordinates. The unbiased additive Gaussian observational error model considered in this work leads to the likelihood function of the form
\[ p(d \mid \eta, \sigma_o^2) = \prod_{i=1}^{N_o} p_\epsilon(d_i - u_i(\eta), \sigma_o^2), \quad p_\epsilon(d_i - u_i(\eta), \sigma_o^2) = \frac{1}{\sqrt{2\pi\sigma_o^2}} \exp \left( -\frac{(d_i - u_i(\eta))^2}{2\sigma_o^2} \right) \] (4.7)

4.2.2 Polynomial Chaos (PC) expansion

Even with a parameter reduction technique (e.g., KL expansion), MCMC sampling remains computationally prohibitive for realistic large-scale models. This is because the sampling of the posterior pdf requires a massive collection of likelihood evaluation is required, each one relying on a full model run. PC expansion is a suitable technique for building a model surrogate that helps accelerating this MCMC sampling process [77, 81].

The PC method expresses the dependency of model solution \( U \) on some uncertain model inputs, using a truncated spectral polynomial expansion [76, 77]. Generally, one has to approximate \( U(\xi) \), where \( \xi = (\xi_1, \ldots, \xi_s) \in \Omega^* \subseteq \mathbb{R}^s \) are independent second-order random variables, with given density \( p_\xi : \mathbb{R}^d \mapsto \mathbb{R}^+ \), and \( \Omega^* \) is the image of the sample space over \( \xi \). Let \( \{\Psi_k, k \geq 0\} \) be a set of \( s \)-variate orthogonal polynomials. The PC method expresses the model output \( U \) using a spectral decomposition of the form

\[ U(\xi) = \sum_{k=0}^{\infty} U_k \Psi_k(\xi) \cong \sum_{k=0}^{P} U_k \Psi_k(\xi), \] (4.8)

where \( U_k \) are called the stochastic modes of \( U \). Because we restrict our investigation to Gaussian processes, \( \xi \) will be a vector of standard Gaussian random variables, and \( \Psi_k \) are the normalized multi-variate Hermite polynomials [161]. The number of terms \( P \) in (4.8) is defined by the polynomial order \( o \) and the stochastic dimension \( s \), such that \( P + 1 = (o + s)!/(o!s!) \). The terms in the PC expansion therefore increases exponentially with respect to an increase in both the polynomial order and the number of the stochastic inputs, which results in the so-called ‘curse of dimensionality’ for PC expansions [68].
The orthogonal system formed by the polynomials implies

\[ \langle U, \Psi_k \rangle = U_k \langle \Psi_k, \Psi_k \rangle, \quad \text{with} \quad U_k = \frac{\langle U, \Psi_k \rangle}{\langle \Psi_k^2 \rangle}. \] (4.9)

Here, the moments \( \langle \Psi_k^2 \rangle \) can be evaluated analytically [77], while \( \langle U, \Psi_k \rangle \) is computed from the correlation

\[ \langle U, \Psi_k \rangle = \int_{\Omega^*} U(s) \Psi_k(s) d\mu(s), \quad k = 0, \ldots, P. \] (4.10)

The evaluation of \( U_k \), thus, involves computing the values of a set of \( P + 1 \) integrals over \( \Omega^* \).

Various projection methods can be used to determine the stochastic modes \( U_k \) of the PC expansion [98]. These methods can be categorized as the Galerkin and non-intrusive approaches. The Galerkin projection method [76] defines the PC coefficients through the reformulation of the system equations, using a weak form with respect to the stochastic dimension. This requires modifying the model code, which makes the method only suitable for small problems. The non-intrusive approaches, as the name implies, do not require modifying the model code by defining the stochastic modes based on the ensemble of the deterministic evaluation of \( U(\xi) \), using a suitable quadrature formula. In this study, we resort to the non-intrusive spectral projection (NISP) method to build the PC surrogate.

In NISP, the integral in Equation (4.10) is first defined as a finite sum, using an appropriate quadrature formula:

\[ \int_{\Omega^*} U(s) \Psi_k(s) d\mu(s) \simeq \sum_{j=1}^{N_q} w_j U(\xi_j) \Psi_k(\xi_j). \] (4.11)

Here, \( w_j \) denotes the weights associated to the node \( \xi_j \), and \( N_q \) is the number of quadrature...
nodes. The set of quadrature nodes is denoted by

\[ \mathcal{S} = \{ \xi_j \}_{j=1}^{N_q} \subset \Omega^* \]  

(4.12)

To evaluate the summation (4.11), one needs to evaluate \( U(\xi_q) \) for all \( \xi_q \in \mathcal{S} \). Let \( \Pi \in \mathbb{R}^{(P+1) \times N_q} \) denote the NISP projection matrix

\[ \Pi_{k,j} = \frac{w_j \psi_k(\xi_j)}{\langle \psi_k^2 \rangle}, \quad k = 0, \ldots, P, \quad j = 1, \ldots, N_q, \]

and \( \xi \) be the vector with coordinates \( \xi_j = U(\xi_j) \). Then, the vector of PC coefficients \( U \) can be expressed as \( \Pi \xi \), or in coordinate form,

\[ U_k = \sum_{j=1}^{N_q} \Pi_{k,j} \xi_j = \sum_{j=1}^{N_q} \Pi_{k,j} U(\xi_j), \quad k = 0, \ldots, P. \]  

(4.13)

A more comprehensive derivation of the NISP method can be found in [68]. The complexity in performing NISP scales with \( N_q \) and grows exponentially with stochastic dimension \( s \). Consequently, the implementation of the NISP method to construct the PC surrogate is usually restricted to a limited small number of stochastic dimensions.

### 4.3 Change of coordinates for uncertain covariance functions

The inference framework described in the previous section incorporates the parameter reduction technique (i.e., KL expansion) and PC surrogate to enable the use of MCMC sampling of large dimensional spatially varying parameters, assuming the underlying covariance function is known. However, in real-world applications, this is rarely the case. The covariance functions are often characterized by some (uncertain) hyper-parameters, such as the variances and correlation length-scales.

The unknown hyper-parameters were traditionally estimated prior to inference using a Gaussian process regression (GPR) [162], where the marginal likelihood function is max-
imized with respect to the hyper-parameters via stochastic interpolation, using a set of noisy observations of the parameter field and data collected at a few locations. One may further use the model outputs observation to further reduce the hyper-parameter uncertainty via Bayesian inference \cite{158}. This however requires building a more computationally demanding PC surrogate, compared to the case of fixed covariance model. Here, we resort to the ‘change of coordinates’ technique, proposed in \cite{69} to simultaneously infer the covariance function hyper-parameters and the KL coordinates vector $\eta$ without introducing the hyper-parameters as additional stochastic dimensions in the PC surrogate.

\subsection{4.3.1 Generalized Karhunen-Loève expansion}

Let $q$ be a random vector of covariance function hyper-parameters with joint density $p_q$ (e.g. $q = \{\sigma_f^2, l\}$, where $\sigma_f^2$ is the process variance, and $l$ is the correlation length scale). The centered stochastic prior process $M$ has a covariance function $C = C(x, x', q)$ parametrized by $q \subset \mathbb{R}^h$, where $h$ is the number of hyper-parameters. The KL expansion in \eqref{4.4} can be truncated and modified to reflect the dependency of the covariance function on $q$ as

$$M_K(x, \omega, q) = \sum_{k=1}^{K} \sqrt{\lambda_k(q)} \phi_k(x, q) \eta_k(\omega). \quad \text{(4.14)}$$

Hereafter, we omit the $x$ dependence to simplify the notations and introduce the scaled eigenfunction $\Phi_k(q) \equiv \sqrt{\lambda_k(q)} \phi_k(x, q)$, which gives \eqref{4.4} the compact form

$$M_K(\omega, q) = \sum_{k=1}^{K} \Phi_k(q) \eta_k(\omega). \quad \text{(4.15)}$$

To eliminate the $q$-dependence of the eigenfunctions, we introduce $\phi_k^r$, the ordered and normalized eigenvectors of reference covariance $C^r$, and the projection coefficients $b_{kk'}(q)$. Since a set of reference eigenfunctions constitutes an orthogonal basis, a change of basis
can be applied to express any scaled eigenfunction in the reference basis as

$$
\Phi_k(q) = \sum_{k=1}^{\infty} b_{kk'}(q) \phi_k^r, \quad b_{kk'}(q) = (\phi_k^r, \Phi_k(q))_D = \int_D \phi_k^r(x) \Phi_k(x, q) \, dx.
$$

(4.16)

In practice, the latter expansion is truncated to the first $K'$ terms. Here, shall assign $K' = K$ to allow for convergence analysis with respect to the first few dominant KL modes. This leads to the expression of the generalized KL expansion of $M$ as

$$
M_K(\omega, q) = \sum_{k=1}^{K} \Phi_k(q) \eta_k(\omega) \approx \sum_{k=1}^{K} \left( \sum_{k'=1}^{K} b_{kk'}(q) \phi_{k'}^r \right) \eta_k(\omega).
$$

(4.17)

To further simplify (4.17), we define:

$$
\tilde{M}_K(\omega, q) = \sum_{k=1}^{K} \phi_k^r \tilde{\eta}_k(\omega, q), \quad \tilde{\eta}_k(\omega, q) = \sum_{k=1}^{K} b_{kk'}(q) \eta_{k'}(\omega).
$$

(4.18)

Since $\tilde{\eta}_k(\omega, q)$ is a linear combination of i.i.d standard random variables, this term can be cast in matrix as

$$
\tilde{\eta}(\omega, q) = B(q) \eta(\omega),
$$

(4.19)

where $B(q)$ is called the coordinate transformation matrix.

Ultimately, Equation (4.19) is the result of recasting the $q$-dependence of $C$ into the expansion of $M_K$ in (4.15), using scaled eigenfunctions and the further approximation of $M_K$ using a $q$-dependent linear transformation $b_{kk'}(q)$. The above derivations require the continuity assumption of the scaled eigenfunctions $\Phi_k(q)$ with respect to $q$ (for more details, see [69] and [163]). To ensure such conditions, the orientation of the decomposed eigenfunctions needs to be consistent. Sraj et al. [69] achieved this by defining the orientation of the eigenfunctions, such that $(\Phi_k(q), \phi_k^r)_D$ has a consistent sign for all $q$. [164].
The KL decomposition of $C(x,x',q)$, required at the beginning of each MCMC iteration to compute the corresponding change of coordinates matrix $B(q)$; this computation becomes costly for large-scale systems, and motivates us to build a PC surrogate of the coordinate transformation $B(q)$ to bypass this expensive step. Hence, we introduce the PC expansion:

$$B(q) \approx \bar{B}(q) = \sum_{\alpha \in A} B_{\alpha} \Psi_{\alpha}(q), \quad \mathbb{E} \left[ \Psi_{\alpha}(q) \Psi_{\beta}(q) \right] = \delta_{\alpha,\beta},$$

(4.20)

where the $\Psi_{\alpha}$ are orthogonal random polynomials in $q$. The PC modes of $B(q)$ are defined by

$$\forall \alpha, B_{\alpha} = \mathbb{E} \left[ B(q) \Psi_{\alpha}(q) \right].$$

(4.21)

However, the convergence of the PC expansion requires sufficient smoothness of the function $q \mapsto B(q)$. In this study, the coastal domain is two-dimensional with several symmetries. As a result, eigenvalues with multiplicity $\geq 1$ and the crossing of eigenbranches (with $q$) emerge, complicating the numerical construction of a smooth function $q$, to obtain smooth dependencies of the matrix $B(q)$. To avoid the above complications, we limit ourselves to a small KL truncation order ($K \leq 4$) for experimental demonstrations in section 4.4.

### 4.3.2 PC surrogate with coordinate transformation

Given that the hyper-parameters $q$ are considered to be uncertain and inferred alongside the KL coordinates $\eta$, the model prediction is expressed as $U(\eta,q)$. Intuitively, this suggests building a PC surrogate that includes $q$ as stochastic inputs, that is, $U(\xi) = \sum_{k=0}^{p} U_k \Psi_k(\eta,q)$. However, this approach drastically increases the computational requirement for evaluating the PC coefficients due to the exponential increase in the terms of the PC expansion. This is particularly problematic for a large-scale ocean model, which
requires a large number of model integrations and a large polynomial order to obtain an accurate PC surrogate for such a system. Adopting a similar framework applied to KL expansion as presented in section 4.3.1 for the inference problem, we seek to replace the direct augmentation of $q$ to solve the PC problem with the change of coordinates method, based on a reference model problem dictated by a fixed covariance function.

To formulate the PC approximation of $U(\xi)$ with the change of coordinates, we first define the reference Gaussian process:

$$\tilde{M}_K^{PC}(\xi) = \sum_{k=1}^{K} \sqrt{\lambda^*_k} \phi^*_k \xi_k,$$  \hspace{1cm} (4.22)

where the $\xi_k$’s are independent standard Gaussian random variables. We now denote $\tilde{U}(\xi)$ the reference PC surrogate, which approximates the predictions of the reference model, that is,

$$U(\eta, q) \approx \tilde{U}(\xi(\eta, q)) = \sum_{k=0}^{P} U_k \Psi_k(\xi(\eta, q)).$$  \hspace{1cm} (4.23)

Similar to Equation (4.19), which expresses the $q$-dependent Gaussian process $\tilde{M}_K$ through the mapping $(\eta, q) \mapsto \tilde{\eta}$ using the transformation matrix $B(q)$, we can apply the change of basis principle to express the surrogate $\tilde{U}$ through the mapping $(\eta, q) \mapsto \xi(\eta, q)$ using another coordinate transformation matrix:

$$\xi(\eta, q) = \tilde{B}(q)\eta,$$  \hspace{1cm} (4.24)

where the $q$-dependent $\tilde{B}(q)$ denotes the change of coordinates, mapping the couples $(\eta, q)$ to $\xi$. Based on (4.22), it comes

$$\tilde{B}_{kl}(q) = \begin{cases} \frac{B_{kl}(q)}{\sqrt{\lambda_k^*}}, & \lambda^*_k / \lambda_k < k, \\ 0, & \text{otherwise}, \end{cases}$$  \hspace{1cm} (4.25)
where $\kappa$ is a small constant related to the level of numerical accuracy introduced to discard the diminishingly small $\lambda_k^r$ and avoid numerical overflows. Note that $\tilde{B}$ is derived from $B$, and therefore, the KL and PC expansions used for the Bayesian inference are said to have a similar coordinate transformation.

### 4.3.3 Bayesian inference with the change of coordinates (COC)

The Bayesian inference framework for uncertain hyper-parameters based on the change of coordinate (COC) methodology can be split into two main steps. The invariant quantities, which are used throughout the MCMC simulation, such as the reference basis and the PC coefficients, can be precomputed offline and stored for the later usage, while the calculation of $(\eta, q)$-dependent quantities, i.e., coordinate transformation matrix and approximated PC solution of $U$, are embedded inside each iteration of MCMC sampling, which is referred to as the online step.

**Offline step**

Given that the reference covariance function $C^r$ has been selected, the KL decomposition is applied to $C^r$ according to (4.2). The eigenpairs $(\lambda_k^r, \phi_k^r)_{k=1,K}$ obtained from KL decomposition are then stored in the memory to be used later in the coordinate transformation process embedded within the online (inference) step. Subsequently, according to (4.22), the same eigenpairs are utilized to build $\tilde{M}_K^{PC}(\xi)$, which represents the reference Gaussian process that characterizes the stochastic parameter field of the reference mathematical model $\tilde{\mathcal{L}}(\xi)$. Finally, the coordinates $\xi$, which specify the realization of $\tilde{\mathcal{L}}(\xi)$, are sampled using the suitable quadrature formula to construct the surrogate $\tilde{U}(\xi)$ of $U$. The PC coefficients $\{U_k\}_{k=0}^P$ computed in the latter step are also stored in the memory to be used later to evaluate the model solutions for likelihood sampling in the inference step. Additionally, in this contribution, we also construct the PC surrogate $\tilde{B}(q)$ of $B(q)$ to be used for fast coordinate transformation in the online step.
Online step

Each iteration of MCMC embedded in the online step involves several substeps. First, $q$ and $\eta$ are proposed. Then, the PC approximated coordinate transformation matrix $\tilde{B}(q)$ is used to compute $\xi = \tilde{B}(q)\eta$. Subsequently, $\xi$ enters the PC expansion alongside the precomputed PC coefficients to approximate $U(\eta,q)$ for a fraction of a second. This PC approximation of $U$ is then used for the likelihood computation, generating the MCMC chains of the posterior of the couple $(\eta,q)$, based on the accept/reject method imposed by the Metropolis-Hastings algorithm. The above processes are repeated until the MCMC simulation reaches its designated terminal iteration.

4.4 Experimental setup

The experimental setup for the problem of inferencing the Manning’s $n$ coefficients (and any other spatially varying parameters) of the coastal ocean is presented in this section. Albeit rather idealized, the setup we propose for the experimentation represents well the dynamics of the coastal ocean.

4.4.1 The advanced circulation model (ADCIRC)

We employ the advanced circulation (ADCIRC) model [99] to test the proposed inference framework. ADCIRC has been widely implemented for simulating coastal and estuarine systems, as well as the analysis of tides and currents [5, 12]. It has also been validated intensively in many storm surge hindcast studies, such as those of hurricanes Betsy (1965), Ivan (2004), Dennis (2004), Katrina (2005), Rita (2005) [2, 3, 103], Gustav (2008) [4], and Ike (2008) [16].

The machinery behind ADCIRC solves a modified version of the SWEs, in which the continuity equation is replaced by the second-order, hyperbolic generalized wave continuity equation (GWCE) to prevent spurious oscillations that often arise from the numerical
solution of the original form [100][101]. Together with the momentum equation, these governing equations are solved on unstructured, triangular elements using a first-order continuous Galerkin finite element scheme. The time derivatives are approximated using centered finite differences in the GWCE and forward differences in the momentum equations. ADCIRC supports a wide range of domain specifications and complex bathymetry structures, including the range of scales necessary to represent the deep ocean basins, continental shelves, and coastal inland areas [92].

4.4.2 The model discretization

We consider an idealized inlet with an ebb shoal domain as depicted in Fig 4.1. The domain comprises a lagoon that is connected to the open ocean on the west through an inlet with twin jetties. The domain is 4500 m wide and 3000 m long, and is discretized to have 1518 grid nodes and 2828 elements. The bathymetry of the domain is designed to resemble the realistic vertical profile of a coastal ocean with the deepest depth in the open ocean area. The depth of the ocean floor increases linearly from 3.8 m at the left-most boundary to 1 m at the mouth of the inlet on the west side of the domain, leveling off at the landlocked area. The diameter of the ebb shoal is 750 m. On the open ocean boundary, ADCIRC is forced with the principal lunar semi-diurnal $M_2$ tidal constituent, with an amplitude of 0.25 m relative to the geoid. This domain is a simplified version of a real ebb shoal, which is a common natural feature of coastal inlets.

4.4.3 Prior covariance function and the reference basis

We assume that the covariance function describing the Gaussian process underlying the uncertain Manning’s $n$ field is isotropic. Specifically, we also note here that the field $M$ is Gaussian, while the Manning’s $n$ coefficients have a finite range $[0.005, 0.2]$. Thus, we use the normal cumulative distribution function (cdf) to map $\mathcal{N}(0, 1)$ to $\mathcal{U}[0, 1]$ and our own customized rescale function that maps $\mathcal{U}[0, 1]$ to $\mathcal{U}[0.005, 0.2]$, to transform between the
realizations of a GP and the actual Manning’s $n$ fields.

A popular class of covariance kernels employed in the field of geostatistics is the Matérn function [165],

$$C(x,x') = \frac{2^{1-v}}{\Gamma(v)} \left( \frac{\sqrt{2}v|x-x'|}{l} \right)^v K_v \left( \frac{\sqrt{2}v|x-x'|}{l} \right).$$  (4.26)

Here, $v$ is a smoothness parameter, $\Gamma$ is the gamma function, and $K_v$ is a modified Bessel function of the second kind. The roughness of the stochastic process has the inverse relation with the value of $v$. As $v \to \infty$, this covariance kernel becomes infinitely differentiable, and the Matérn covariance function converges to the squared exponential covariance function, which can be written as

$$C(x,x',q) = \sigma_f^2 \exp \left( -\frac{(x-x')^2}{2l^2} \right).$$  (4.27)
where \( q = \{ \sigma_f^2, l \} \). We select this special case of the Matérn covariance to simulate the stochastic process; it is one of the most used covariance functions to simulate spatially varying parameter fields [166]. In our case, only the correlation length \( l \) affects the shape of the eigenfunctions, while the process variance \( \sigma_f^2 \) simply scales the eigenvalues. Therefore, we fix \( \sigma_f^2 = 1 \) for all experiments and only assume the uncertainty in \( l \), that is, \( q = \{ l \} \).

The choice of the reference basis is similar to that chosen in [69]. We compute the reference eigenbasis by solving (4.3), given the \( q \)-averaged covariance function defined as

\[
C^r = \mathcal{C} = \int C(q) p_q(q) dq. \tag{4.28}
\]

This choice of reference covariance function has been analyzed (see [69]) and contrasted with the choice \( C^r = C(l') \) for several values of \( l' \). It has been shown to yield the lowest representation error (averaged relative error over \( q \)) compared to selecting any reference covariance as \( C(l') \). This choice is optimal because it uses the eigenfunction spanning the optimal subspace to represent \( M(\omega, q) \) when \( q \) varies with probability law \( p_q(q) \). Additionally, the error in approximating \( U \) with the PC surrogate \( \hat{U} \) through coordinate transformation is consistent with the previous analysis, i.e., the error is minimized when \( \mathcal{C} \) is used for all truncation order \( K \).

In practice, the explicit expression of the averaged covariance function in (4.28) is rarely available. It therefore needs to be approximated using a sampling-based approach. In our case, the Gauss quadrature rule was used to appropriately sample a set of \( q = \{ l \} \) from \( U[l_{\text{min}}, l_{\text{max}}] \), where \( l_{\text{min}} \) and \( l_{\text{max}} \) are the minimum and maximum correlation length scales, respectively. In the present work, we take advantage of the moderate size of the mesh to create and store the 1518 × 1518 discrete covariance matrix whose entries are compute by quadrature. For each quadrature representing a certain \( l \), the value of each element \( x_{ij} \) in this matrix is simply \( C(x_i, x_j, l) \). We then perform the summation of this matrix over \( l \) using
the for loop and take the average over the number of quadrature points, to finally obtain the approximated $\mathcal{C}$ matrix. Note that this method only requires storing a single covariance matrix. For the problems with much larger spatial discretization in which computing a large covariance matrix is intractable, covariance approximation methods such as fixed rank kriging [167], covariance tapering [168], and the full-scale approximation [169] could be used.

### 4.4.4 Observation simulation system experiments (OSSEs)

To assess the performance of the proposed inference framework, we implement ‘observation simulation system experiments (OSSEs)’. In these experiments, the true Manning’s $n$ field is chosen (assumed to be known). Then, this true field is used in the ADCIRC model, to compute the reference solution (i.e., water elevation), denoted by $u(x,t)$, where $x$ is the location and $t$ is time. Water elevation data are then extracted from the reference solution at 15 measurement stations every hour for 4.5 days, after the 12 hour ramp-up period. This yields a total $N_o = 15 \times 108 = 1620$ observations. These are then perturbed with i.i.d. additive noise $\varepsilon_i \sim N(0, \sigma^2_e = 0.01)$ and are used to infer the true Manning’s $n$ field. We investigate two test cases:

- **Random profile**: $M^{ran}(x)$ drawn at random from a Gaussian process, $GP(0, \mathcal{C})$, where $\mathcal{C}$ is the squared exponential covariance. To do so, we sample the quadruplet, $\{\eta_1, \eta_2, \eta_3, l\}$, that characterizes the random field $M_{K=3}$. We limit ourselves to the inference only three modes ($K = 3$), on account of the limitation in the size of the stochastic dimension that can be accommodated in the PC construction scheme (i.e., NISP with full tensorization quadrature rules). This does not undermine the validity of our assessment, which is to identify the improvement of the inference using COC with respect to the inference without COC, using the same number of modes. Furthermore, one can alleviate this limitation of small stochastic dimensions due to the chosen PC scheme by using the sparse quadrature methods [68]. By constructing a
true field from a known prior covariance, we can assess our method by comparing the posterior pdf and maximum a posteriori probability (MAP) estimate with the true field, in terms of \( \eta \) and \( q \) directly (see in the next section). We show the image of the true field \( M^{true} \) on the left of Fig. 4.2. Here, the sampled quadruplet that generates this particular Manning’s \( n \) field realization is \( \{ \eta_1, \eta_2, \eta_3, l \} = \{ 1.73, 0.26, 0.04, 1255 \} \).

- **Zonal profile**: the true Manning’s \( n \) profile \( M^{zone}(x) \) is derived from the equation:

\[
M^{zone}(x) = \alpha M_1(x) + \beta M_2(x) + (\beta - \alpha) M_3(x).
\]  

(4.29)

Here, \( M_1(x) \), \( M_2(x) \) and \( M_3(x) \), are the parameterization coefficients, which take a value between 0 and 1. The Manning’s \( n \) coefficients in the open ocean area are represented by \( \alpha \), and \( \beta \) are the Manning’s \( n \) coefficients in the landlocked area. This parameterization is such that the east and the west sides of the inlet have constant Manning’s \( n \) values, and there is a linear increase (or decrease) in the Manning’s \( n \) coefficients inside the inlet, with the rate of change proportional to the difference between \( \alpha \) and \( \beta \). Therefore, we refer to this parameterization as the ‘zonal’ profile.

The zonal profile has also been used in [5] to represent the simplistic approximation of the realistic Manning’s \( n \) field in the coastal domain, where the seabed roughness varies depending on the distance from the coastline, and the monotonic increase in the Manning’s \( n \) parameter within the inlet is caused by the sediment transport between the open ocean and the lagoon. By specifying \( \alpha = 0.005 \) and \( \beta = 0.1 \), we generate the true field \( M^{zone} \), as shown in the right panel of Fig. 4.2.

For the inference, we consider the squared exponential prior covariance with \( q = \{ l \} \) for both cases. The prior for \( l \) is uniform \( \mathcal{U}[l_{\min} = 1000 \text{ m}, l_{\max} = 4000 \text{ m}] \). This range of the correlation length scale was chosen such that \( l_{\max} \) does not exceed the domain length in any direction, and \( l_{\min} \) is approximately 1/4 in diameter of the computational domain. This is to ensure that the prior covariance can be adequately represented by just a few eigen-
Figure 4.2: The true Manning’s $n$ field of the two test cases. Truths were used to generate the synthetic water elevation observations for the inference. Left: $M^{ran}$, generated from sampling the KL coordinates and $l$ of the squared exponential covariance model. Right: $M^{zone}$ generated by setting $\alpha = 0.005$ and $\beta = 0.1$ in Equation (4.29).

modes. In addition, we do not want to assign an exceedingly small $l_{\text{min}}$ to our inference problem because the sparse observation points in our domain cannot resolve the small-scale variability assumed by the prior covariance with small correlation length scale.

For the second test case, $M^{zone}$, we use a simplified 1D prior covariance in which we implicitly state that a priori, there is no dependence between two points in space on the transverse direction; the dependence between two points only exists in the horizontal (x) direction. To derive such a covariance function, we consider that the Gaussian kernel is separable, that is, the multidimensional kernel can be written as a product of one-dimensional kernels:

$$C(x, x') = \prod_{i=1}^{N_D} C_i(x_i, x'_i),$$  \hspace{1cm} (4.30)

where, $N_D$ is the number of the dimension of $D$. For 2D domain, we use

$$C_1(x_1, x'_1) = \exp \left( -\frac{(x_1 - x'_1)^2}{2l^2} \right), \quad C_2(x_2, x'_2) = 1,$$  \hspace{1cm} (4.31)
that is an infinite correlation in the second dimension. The variability range for \( l \) is the same as previously.

4.5 Results and Discussions

4.5.1 Offline step constructions results

We first construct the reference basis \((\lambda_k^r, \phi_k^r)_{k=1,K}\) derived from the reference covariance \( C^r = \overline{C} \) for both test cases. Fig 4.3 shows the first six eigenmodes (from left to right, top to bottom) and the spectrum (bottom) of the KL decomposition of \( \overline{C} \) of the first test case. We observe that, to retain about 90% of the energy, truncation order \( K = 7 \) is required. Similar plots for the second test case \( M^{zone} \) are shown in Fig 4.4. In this case, the reference covariance \( \overline{C} \) is simpler and \( K = 4 \) is sufficient to retain more than 90% of the energy.

To setup the MCMC inference with the change of coordinates method, we compute the coordinates transformation matrix \( B(q) \). For \( M^{ran}(x) \), the dimension of \( B(q) \) is \( 3 \times 3 \). For the second test case, \( M^{zone}(x) \), we increase \( K \) to 4; hence, \( B(q) \) is of the dimension \( 4 \times 4 \).

In a large-scale system, it is not practical to compute \( B(q) \), which requires performing KL decomposition each time \( q \) changes at every MCMC iteration. Therefore, we build the PC surrogate \( \hat{B}(q) \) of \( B(q) \) according to Equation (4.20) using an Pseudo-spectral projection (PSP) methods [170, 171], a variant of the NISP which uses the nodes of a sparse grid in the domain \( \Omega^* \). PC construction using 15 Gaussian quadrature points (15 different realizations of \( l \)) based on Legendre polynomials with PC order \( o = 5 \) was found to be sufficient for approximating \( \hat{B}(q) \) for both test cases. Fig. 4.5 shows the results of approximating the true field \( M^{ran} \) of the first test case, using the reference basis and the change of coordinates. The top-right plot is \( \hat{M}_K \), which is the approximation of \( M^{ran} \), using the reference basis and transformation \( B(q) \). Finally, the bottom plot is \( \hat{M}_K^{PC} \), the approximation of \( M^{ran} \) with additional PC approximation of the coordinate transformation itself. \( \hat{M}_K^{PC} \) exhibits a larger error in approximating the truth than \( \hat{M}_K \) because it accumulates both the errors from
Figure 4.3: Resulting basis from KL decomposition of $\overline{C}$ for the first test case, $M^\text{ran}$. First three panels: the first six eigenmodes (from left to right, top to bottom), bottom: the spectrum of $C$. 
Figure 4.4: Resulting basis from KL decomposition of $\overline{C}$ for the second test case, $M^{cone}$. First three panels: the first six eigenmodes (from left to right, top to bottom), bottom: the spectrum of $\overline{C}$.
the truncation in building the COC and the PC approximation of $B(q)$. However, for this simple test case, the difference in error between $\tilde{M}_K$ and $\tilde{M}_K^{PC}$ is almost negligible because a sufficient number of quadratures and polynomial order were used to construct the PC surrogate. Nevertheless, the discrepancy in estimating the truth in the reference basis is noticeable. The circular gradient pattern of $M^{ran}$ in the top-left plot is slightly compressed in the $x$-direction for both $\tilde{M}_K$ and $\tilde{M}_K^{PC}$ plots. Also, the range between the maximum and the minimum Manning’s $n$ coefficients is slightly underestimated. This is clearly due to the $K$-term truncation in Equation (4.17). However, overall, we have demonstrated that the COC method is efficient in approximating the Gaussian process $\tilde{M}_K$ on the $q$-averaged reference basis.

The second task in the offline step is to build the PC surrogate of the model solution $U(\eta, q)$ following (4.23), based on the reference Gaussian process $\tilde{M}_K^{PC}(\xi)$ in (4.22). Following the former setup, $\tilde{M}_K^{PC}(\xi)$ is a stochastic process with $q$-averaged covariance function $\mathcal{C}$, in which the KL decomposition is truncated to the first $K = 3$ dominant modes. Employing the NISP approach, we obtain the PC approximation $\tilde{U}(\xi(\eta, q))$ of the reference model solution. For the first test case, full tensorization of 1-D Gauss quadrature and Hermite polynomial order $o = 6$ were used to accurately emulate the response surface of the full solution $U(\eta, q)$. With stochastic dimension $s = 3$ and 5 Gauss nodes in each direction, 125 ADCIRC model runs are required. After the surrogate is successfully built, we can evaluate the PC approximation of the ADCIRC predictions, i.e., the water elevation at observation stations, denoted by $\tilde{u}(\xi)$. Given the location in space $x_i$ and time $t_i$, the components of $\tilde{u}(\xi)$ are

$$\tilde{u}_i(\xi) = \tilde{U}(x_i, t_i; \xi) = \sum_{k=0}^{P} U_k(x_i, t_i) \Psi_k(\xi), \quad i = 1, \ldots, 1620. \quad (4.32)$$

These model predictions then enter the likelihood computation in the MCMC inference. In
Figure 4.5: Approximation of $M^{ran}$, using the reference basis and change of coordinates. Top-left: $M_K$, the true profile, top-right: $\tilde{M}_K$, the approximation of $M_K$ in the reference basis, and bottom: $\tilde{M}_K^{\text{PC}}$, the approximation of $M_K$ in the reference basis with the PC approximation of the change of coordinates, $B(q)$.
Figure 4.6: Time-evolution of the water elevation as produced by the full model (blue) and the surrogate model (red). Each plot shows the water elevation at different observation stations and different quadrature points (i.e., different realization of $\xi$), which characterizes the Manning’s $n$ field.

Fig 4.6, we plot for the first test case, the time-evolution of the water elevation prediction, produced by the ADCIRC, contrasted with its surrogate counterpart at three observation stations, with each panel plotting a different realization of $\xi$. We notice that the water elevation predictions are approximated well by the surrogate and, thus, the surrogate can be used in place of the full ADCIRC in MCMC.

Similar offline construction procedures were carried out for the second test case $M^{zone}$. However, we increase the KL truncation to $K = 4$, for the Gaussian process to capture better the small-scale variability in the true Manning’s $n$ field we attempt to infer. This requires 625 ADCIRC runs to build the surrogate, using the full tensorization quadrature method.
4.5.2 1st case: inference of a random Manning’s $n$ field drawn from a Gaussian process

To perform the MCMC inversion, we start by specifying the priors. Since $\eta = \{\eta_1, \eta_2, \eta_3\}$ are the standard normal random variables, $p(\eta)$ is a centered Gaussian. As for the length scale $l$ and the noise hyper-parameter $\sigma_o^2$, we use the uninformative, improper, Jeffrey prior. However, instead of inferring the actual correlation length scale, we seek to infer the normalized correlation length scale $l_{\text{norm}}$, which can be computed through the mapping $[1000, 4000] \mapsto [0, 1]$. For example, the $l_{\text{norm}} = 0.085$ is equivalent to $l = 1255$ m.

Having identified all priors, we can use the data predicted by the surrogate $\tilde{u}(\xi(\eta, q))$ to compute the likelihood:

$$p(d | \eta, q, \sigma_o^2) \approx \tilde{p}(d | \eta, q, \sigma_o^2) = \prod_{i=1}^{N_o} p_e(d_i - u_i(\xi(\eta, q)), \sigma_o^2),$$

(4.33)

where $\xi(\eta, q)$ and $p_e$ is given in Equation (4.24) and Equation (4.7), respectively. To specify $\tilde{B}$ in Equation (4.25), we set $\kappa = 0$ because $\lambda_{k<3}^r/\lambda_k^1$ remains large enough for the current setting. With all components of the Bayes rule defined, we can proceed with generating the posterior $\tilde{p}(d | \eta, q, \sigma_o^2 | d)$ using an adaptive Metropolis-Hasting algorithm.

We first examine the inference results from using PC-MCMC with COC. A total of $3 \times 10^5$ MCMC iterations were found to be sufficient in obtaining the posterior with good resolution. We also observe well-mixed MCMC chains for all inferred parameters (not shown). The marginal posteriors are estimated using a standard kernel density estimation (KDE) method [172]. The posteriors of the inferred 3 first KL coordinates and the hyper-parameter $l$ for the inference of $M^{\text{ran}}$ are shown in Fig.4.7. We also infer $\sigma_o^2$, which we found a Maximum a Posteriori (MAP) close to the value used to generate the data, i.e., $\sigma_o^2 = 0.01$. Similar findings are also the case for all later experiments (not shown for brevity). Concerning the posterior pdf of the KL coordinates, the information gained from the observations is found to be significant only for $\eta_1$, which shows large divergence from
the prior and significant decrease in the posterior pdf spread. An interesting result in these plots is in the posterior pdf of the normalized length-scale $l_{\text{norm}}$ as depicted on the right of the second panel. Here, the MAP estimate of the posterior of $l_{\text{norm}}$ (presented with dashed blue) matches almost perfectly with the true normalized length-scale used to generate the field $M_{\text{ran}}$. This demonstrates the efficiency of PC-MCMC with COC to estimate the uncertain length-scale of the Gaussian process that best estimates the reference profile. However, Fig. 4.7 is not informative in term of validating the accuracy of the inferred KL coordinates with respect to the true field $M_{\text{ran}}$, since these plots only display the raw inference results, in which all parameters simultaneously change within the MCMC sampler. Each sample on the posterior is, in fact, defined on a different basis.

To assess the inference results against the truth in a fixed system of coordinates, one needs to express the sampled fields in the same basis. To this end, we express all sampled field in the reference basis using the relation $\xi(\eta, q) = \hat{B}(q)\eta$ as given in (4.25). Fig. 4.8 shows the posteriors of the KL coordinates $\xi$ compared to the prior and the KL coordinates of the true field $M_{\text{ran}}$. Similar to the conclusion drawn from Fig. 4.7, the clear information gained from the observations is only revealed by the improvement of the posterior of the 1st KL coordinate $\xi_1$. The MAP of the posterior of $\xi_1$ clearly converges to the value of the first KL coordinate of the truth, and the posterior spread is greatly reduced compared to the prior. This is not the case for $\xi_2$ or $\xi_3$, in which the MAP estimate accurately matches the truth, but shows no clear improvement in term of posterior spread. This is likely due to the fact that the mean of the priors was already closed to the true values and because of the observation noise. Nevertheless, Fig. 4.8 further emphasizes the strengths of the PC-MCMC with COC, which is efficient at inferring the hyper-parameters and concurrently improve the estimation of the KL coordinates.

After inferring $\xi$, we can reconstruct the inferred Manning’s $n$ field using the KL expansion with known reference eigenbasis, and compare this field to the true field $M_{\text{ran}}$. We present in Fig. 4.9 the comparison of the inferred Manning’s $n$ fields from different infer-
Figure 4.7: Comparison between the priors and marginal posteriors of the 3 KL coordinates $\eta_k$, the length-scale hyper-parameter $l_{\text{norm}}$, and the noise variance hyper-parameter $\sigma_0^2$ (only the posteriors are shown for $l_{\text{norm}}$ and $\sigma_0^2$) for the inference of $M_{\text{ran}}$ using COC method. The maximum a posteriori probability (MAP) estimate for each parameters are also presented using dashed blue line. The true $l_{\text{norm}}$ used to generate the true field is plotted in dashed green.
Figure 4.8: Comparison between the priors, marginal posteriors of the 3 KL coordinates $\eta_k$ and the true KL coordinates of $M^{\text{ran}}$ in the $q$-averaged reference subspace ($\xi$). In each plot, the prior is the black curve, the posterior is in blue, the MAP estimate in dashed blue, and the truth is in dashed green.
ence settings and methods: 1) the truth 2) the fixed \( l \) basis, 3) the reference basis without using the change of coordinates, and 4) PC-MCMC with COC. First, we observe the top right plot of Fig.4.9 where the correlation length-scale is fixed to \( l^{\text{norm}} = 0.189 \) or \( l = 1567 \) m. As expected, we obtain a poor estimate of the truth in this case, simply because the scheme is non-adaptive in \( q \). In addition, the size of fixed \( l^{\text{norm}} \), in this case, is larger than the \( l^{\text{norm}} \) so it cannot resolve small-scale oscillations. We also plot in the bottom-left of Fig.4.9 the inferred Manning’s \( n \) field using the basis derived from the \( q \)-averaged prior covariance \( \overline{C} \). In this case, the pattern of the inferred field shows more resemblance to the truth, which can be explained by the better fitness of the basis derived from \( \overline{C} \) compared to \( C^r \) with fixed \( q \). It is nevertheless, visually, a poor estimate of the true field. Finally, as expected, the PC-MCMC with COC case in the bottom-right of Fig.4.9 clearly suggests significant improvement in the spatial structure of the inferred field compared to the two former cases. It actually matches almost perfectly to the true field reconstruction based on the reference basis and COC present in Fig.4.5. This indicates that the COC approach is very robust, up to the limit of accuracy in the approximation of \( \overline{C} \) and \( B(\eta, q) \).

We further investigate the accuracy of the estimated fields by plotting the percentage error \( e_\% \) of each inferred field with respect to the truth as presented in Fig.4.10 calculated node-wised as

\[
e_\%(x_i) = \left| \frac{M^{\text{in}}(x_i) - M^{\text{true}}(x_i)}{M^{\text{true}}(x_i)} \right| \times 100, \quad i = 1, \ldots, 1518,
\]

where \( M^{\text{true}} = M^{\text{ran}} \) in the first test case, \( M^{\text{in}} \) represents the inferred field, and \( x_i \) represents the discretization node \( i \) in the domain. For the fixed \( l \) case (top-left), a large error is shown at the southeast corner of the land-locked area with the maximum error of about 150\%. This clearly highlights the mismatch between the truth and the inferred field when hyper-parameters uncertainty is not considered. Similar to the previous discussion, using the
Figure 4.9: Spatial maps of several inferred Manning’s $n$ fields compared to the truth: (top-left) the true Manning’s $n$ field, (top-right) the inferred Manning’s $n$ field using a fixed correlation length scale, (bottom-left) the inferred Manning’s $n$ field based on the basis derived from $\mathcal{C}$ without using COC, and (bottom-right) the inferred Manning’s $n$ field from using PC-MCMC with COC.
Figure 4.10: Spatial maps of the percent error with respect to the truth for several inferred Manning’s $n$ fields: (top-left) the inferred Manning’s $n$ field using a fixed correlation length-scale, (top-right) the inferred Manning’s $n$ field based on the basis derived from $\overline{C}$ without using COC, and (bottom) the inferred Manning’s $n$ field from using PC-MCMC with COC.

basis derived from $\overline{C}$ (top-right plot) reduces the error as can be observed by the blue color representing the low error in most area. However, the exceedingly large error is still shown in both lower corners of the domain. Finally, the PC-MCMC with COC yields relatively small $e\%$ throughout the domain, with the maximum error at about 60%.

In many scenarios, the ocean tides may not be very sensitive to the small change in the model parameters such as Manning’s $n$ coefficients. We are also interested in examining the resulting water elevation data produced by ADCIRC run subjected to the inferred Manning’s $n$ field we have obtained. Fig 4.11 show the time-evolution of the water elevation in the final 12 hours of the ADCIRC simulations. Each curve in each subplot is the water
elevation produced with the different MAP values of the Manning’s $n$ field. Here, we want to highlight the difference in the sensitivity of the water elevation to the inferred Manning’s $n$ coefficients with respect to the observed location. Therefore, the three subplots represent the different locations in space: the open ocean, the inlet, and the landlocked area, respectively. In the open ocean area, the model is less sensitive to the difference in the field of the Manning’s $n$ coefficient compared to the landlocked area. This is reasonable since the west side of the domain has deeper bathymetry, reducing the value of the drag law coefficient [5]. The errors are larger at stations where the model is more sensitive. Among the three inference results, the fixed hyper-parameter case clearly shows the largest error in water elevation in the landlocked area with respect to the truth (RMSE = 0.0023), consistent with the corresponding inaccurately inferred Manning’s $n$ field. For the landlocked area, using $q$-averaged basis with and without COC produce accurate recovery of the water elevation. However, the case without COC shows the best estimate, with the smallest RMSE among all three cases at 0.00015. This can also be expected since, given the sufficiently well-approximated structure of the true field, it is often the case that numerical models are not sensitive to its value at every point in the computational domain, but rather more sensitive to the range of the parameter in the domain itself. In our case, the range of Manning’s $n$ coefficients given by the $q$-averaged basis without COC is the closest to the range of the true Manning’s $n$ field, while the case with COC slightly underestimates the maximum Manning’s $n$ value. This may also suggest the need to better approximate $\mathbf{B}(\mathbf{q})$ by increasing the number of truncation terms $K$ in (4.17), at the cost of increasing the computational load for building the PC surrogate of $U$.

4.5.3 2nd case: Inference of a 2D parametrized zonal Manning’s $n$ field

In the 2nd test case, we proceed our assessment of the PC-MCMC with COC method in similar manners as presented in the 1st test case. However, the true Manning’s $n$ field $M_{cone}$ used in this test case is derived independently from the Gaussian process assumption.
Figure 4.11: Time-evolution of the water elevation in the final 12 hours of the ADCIRC simulations subjected to different Manning's $n$ field.
As a result, this test case better simulates the inference problem in real-world settings. According to the observed spectrum in Fig. 4.4, we set $K = 4$, which gives $\lambda_4/\lambda_1 = 0.101$. This ensures that the $q$-averaged reference eigenbasis encapsulates roughly 90% of the information contained in $\mathcal{C}$. Without loss of generality, we omit the inference results using fixed-$q$ basis on the experiments reported below, as we found in the previous section that the fixed-$q$ case is not very informative. We will instead focus on the comparison between the inference results using $q$-averaged basis with and without COC. For simplicity, we shall also refer to the inference using $q$-averaged basis without the change of coordinates as the baseline case, while the inference with the change of coordinates will be referred as the COC case.

Fig. 4.12 outlines the inference of $M^{zone}$ of the baseline case. The posteriors of the four coordinates are compared with their respective standard Gaussian priors. We notice the significant improvement of the posteriors distributions for the first two coordinates, while the other two only show the shift in the posterior with no improvement on the pdf spread. With COC, shown in Fig. 4.13, better information gain has clearly been discerned for all four coordinates. Namely, the posteriors of the first three coordinates clearly exhibit smaller spread compared to their respective posteriors in the baseline case, and the posterior of the fourth coordinate clearly show larger divergence from the prior than that of the baseline case, demonstrating the ability of the MCMC sampler to explore the posterior a more suitable subspace. These improvements are indeed the result of including the uncertainty in $l$ in the inference problem.

Next, we examine the fields of inferred Manning’s $n$ coefficients constructed using the MAP estimates of $\eta$ and $l$. The resulting Manning’s $n$ fields and their respective absolute errors with respect to the truth $M^{zone}$ are presented in Fig. 4.14. Looking at the second panel of Fig. 4.14, the inferred Manning’s $n$ fields from both the baseline and the COC case show resemblance in the overall spatial structure with respect to the truth. Both methods are observed to be equally successful in recovering the Manning’s $n$ value in the open
Figure 4.12: Comparison between the priors and marginal posteriors of the 4 KL coordinates for the inference of $M_{\text{zone}}$ using $q$-averaged covariance basis. The maximum a posteriori probability (MAP) estimate for each parameters are also presented using dashed blue line.
Figure 4.13: Comparison between the priors and marginal posteriors of the 4 KL coordinates and the length-scale hyper-parameter \( l \) (only the posterior is shown) for the inference of \( M^{zone} \) using COC method. The maximum a posteriori probability (MAP) estimate for each parameters are also presented using dashed blue line.
ocean area \((\alpha = 0.005)\). However, overall, the COC case provides more accurate results with the Manning’s \(n\) values being closer to the truth both in the inlet and the landlocked areas. The range of the Manning’s \(n\) value is overestimated in all cases but appeared to be more pronounced for the baseline case, as can be discerned by observing and comparing the maximum values in the color bars of the three first subplots. In the bottom panel of Fig[4.14] the percent error maps are presented. Overall, the percent error was found to be small over most of the area, except at the west side of the inlet, where the error reaches to almost 390\% for the baseline case and 360\% for the COC case along the edge of the two barriers. This behavior can be explained by the squared exponential prior covariance model considered to approximate the field, which is very smooth in space and therefore not well-suited for representing the discontinuous slopes of \(M^{z\mathrm{one}}\) at the two inlet boundaries.

To better discern the differences in the inferred field results between the baseline and the COC cases, we present in Fig[4.15] the cross-section plot of the parameter fields at \(y = 0\) with varying Manning’s \(n\) value in the \(x\)-direction. Clearly, the truth \(M^{z\mathrm{one}}\) is not a smooth function with sharp turns at the two entrances of the inlet, which is the result of the field parameterization in Equation (4.29). Similar to the conclusion we have drawn from the previous figure, both baseline and COC cases accurately estimate the parameter field in the open ocean with the exception of a small area around the western boundary. This is again the result of the attempt to fit the constant function in a basis of eigenmodes for squared exponential prior covariance. The consequence of this limitation is more pronounced when estimating large Manning’s \(n\) values in the even smaller landlocked area. Here, the inference results from both cases manifest significant discrepancy with respect to the truth. The shape of the inferred field subjected to a smooth Gaussian process with a few modes do not well-represent the constant profile in the landlocked area. That aside, the COC case demonstrates clear improvement compared to the baseline in recovering the \(M^{z\mathrm{one}}\) profile in this area, with the maximum Manning’s \(n\) value of 0.1088; less than 10\% larger than \(\beta = 0.1\). The baseline, on the other hand, produces the maximum Manning’s \(n\) value of
Figure 4.14: Spatial maps of several inferred Manning’s $n$ fields compared to the truth: (top) the true Manning’s $n$ field, (middle-left) the inferred Manning’s $n$ field based on the basis derived from $\mathcal{C}$ without using COC, and (middle-right) the inferred Manning’s $n$ field from using PC-MCMC with COC. Bottom panel, spatial maps of the percent error with respect to the truth: (bottom-left) the inferred Manning’s $n$ field based on the basis derived from $\mathcal{C}$ without using COC, and (bottom-right) the inferred Manning’s $n$ field from using PC-MCMC with COC.
Figure 4.15: Cross section plot at $y = 0$ (midway between north and south) of the inferred Manning’s $n$ fields compared to the true field $M^{true}$ (black). The inferred field based on the basis derived from $\mathcal{C}$ without using COC and with COC are plotted in red and blue, respectively.

0.1233; 20% larger than the true maximum. Furthermore, one can observe the clear benefit of including the uncertainty in $l$ in the COC case, that is, it also attempts to recover the maximum Manning’s $n$ coefficients of the eastern boundary of the landlocked area.

Finally, the sensitivity of the produced water elevation to the inferred field is investigated. The water elevation at three locations during the 12 final hours of ADCIRC simulation for different Manning’s $n$ fields are illustrated in Fig. 4.16. Similar to the 1st test case $M^{ran}$, the resulting water elevation in the open ocean of the baseline and the COC case is well estimated compared to the truth. This is both because the inferred Manning’s $n$ fields from both cases in this area are well recovered and the water height is not very sensitive to the differences in Manning’s $n$ values in the area with deep bathymetry. The distinction in the resulting water elevation between the three cases starts to be noticeable in the inlet and the landlocked region. The baseline produces the larger error compared to the COC case in both areas, mainly, due to the exaggeration in the range of the estimated Manning’s $n$ values with the maxima exceeding more than 20% of that of the true field. The improvement in the resulting water elevation produced by the COC case over the baseline case is clearly outlined in the landlocked area.
Figure 4.16: Time-evolution of the water elevation in the final 12 hours of the ADCIRC simulations subjected to different Manning’s $n$ field.
To summarize, this section has demonstrated the efficiency of the change of coordinates approach for inference in a more realistic coastal ocean scenario compared to the first test case, with significant improvements in the inference results.

4.6 Conclusions

We proposed a Bayesian inference framework to infer a 2D spatially varying parameter field from a prior defined by a Gaussian process in the context of coastal ocean modeling. A Karhunen-Loève (KL) expansion was then exploited to approximate the process using a few eigenmodes. To infer the KL coordinates, we used a polynomial chaos (PC) expansion to build an inexpensive surrogate, which is then used as the substitute of the full forward model to speed up the likelihood sampling in a MCMC algorithm. The uncertainty in the hyper-parameters \( q \) involved with the prior covariance kernel describing the prior Gaussian process has been included in our framework with the aim to further enhance the performance of the Bayesian inference in estimating the parameter fields compared to the case where the prior covariance function hyper-parameters are assumed to be known \textit{a priori}. To efficiently estimate both \( q \) and the KL coordinates without augmenting \( q \) as additional stochastic dimensions to the expensive PC surrogate construction, we adopted and implemented the change of coordinate (COC) method in the generalized KL expansion framework proposed in [69]. The derivation of the generalized KL expansion was motivated by the idea that the eigenfunctions of the KL expansion form a basis in a space dictated by \( q \). Consequently, the KL coordinates defined in one basis can be transformed into another using a coordinate transformation operator that also depends on \( q \).

The Bayesian inference with COC can be summarized as follows. In the offline step, we defined and decomposed the reference covariance function from which the reference eigenbasis are computed through KL decomposition. Then, we construct the PC surrogate of the forward model that accounts for the dependence of the model prediction on the KL coordinates defined on the reference basis. Next, in the online step, every MCMC itera-
tion samples $\eta$ and $q$, where $q$ is used together with the reference eigenbasis to compute the coordinate transformation matrix $B(q)$. The reference PC surrogate then use the KL coordinates $\xi$ in the reference basis, which is the result of transforming $\eta$ through the relation $\xi = B(q) \eta$, to compute the model prediction required in the likelihood evaluation to explore the posterior. The reference covariance was chosen as the $q$-averaged covariance which has been shown to be optimal in term of representative error both for building the coordinate transformation operator and to construct the PC surrogate of the reference Gaussian process. The computation of $B(q)$ at every MCMC iteration requires one KL decomposition, which is costly. Thus, we also add one key contribution to this paper by constructing the PC surrogate $\hat{B}(q)$ of $B(q)$ itself. This greatly accelerates the online step and allows the inference framework with COC to be applied to realistic ocean models.

We tested the efficiency the proposed inference framework in an idealized coastal ocean setting, by conducting OSSEs to infer the Manning’s $n$ field from synthetic water elevation observations generated by running ADCIRC model with the assumed known true field. Two test cases with different reference Manning’s $n$ profile were utilized for this analysis: 1) the true profile drawn randomly from a Gaussian process with a prescribed hyper-parameter value and 2) the zonal Manning’s $n$ profile parametrized using a linear equation. While the first test case is quite informative in term of assessing the proposed framework by comparing directly the inferred $\eta$ and $q$ to those used to generate the truth, the second test case represents a more realistic Bayesian inference scenario where the truth is not predetermined by a similar underlying stochastic process as the prior. In each test case, we compared the inference results using a fixed covariance function with those using the COC. We observed better information gain when the hyper-parameters were included in the inference. The COC case also accurately recovered the true correlation length-scale in the first test case, resulting in a more accurate recovery of the true Manning’s $n$ field. Similar results have been obtained in the second test case, that is, including the COC in the Bayesian inference helps producing a better estimate of the true parameter profile compared to the
baseline where the inference is performed without incorporating the change of coordinate technique.

Despite the efficiency of the Bayesian inference with COC, this approach calls for a few comments. Firstly, the COC method is very robust, but its accuracy in transforming the coordinates from one basis to another is limited by the approximation reference basis and the coordinate transformation matrix. Secondly, the COC method relies on the $q$-continuity of $B(q)$. In the large-scale problems where the domain has several spatial dimension, this assumption would not hold through the emergence of the multiplicity of eigenvalues and crossing of eigen-branches with $B(q)$. This limits the scope of using COC to the few first dominant eigenmodes before the crossing/multiplicity issues become intractable. Motivated by these limitations, we plan in our future work to develop the inference method accounting for the uncertainty in the hyper-parameters that works on quantities that directly present smoothness and unambiguous dependences with $q$, such as the prior covariance, thus alleviating the need to follow a mode (a branch) when $q$ varies.
Chapter 5

Combining Ensemble Kalman Filter and Multiresolution Analysis for Efficient Assimilation into Adaptive Mesh Models

In the preceding chapters, we focus on quantifying and reducing the coastal ocean parameters uncertainty through Bayesian inversion. This chapter offers a distinct contribution to the thesis by considering another aspect of the UQ, that is, the model discretization uncertainty. We introduce a new approach combining the EnKF with multiresolution analysis (MRA) to enable and simulate efficient data assimilation on the model with adaptive mesh refinement (AMR) capability. Five variants of MRA-EnKF are derived based on different projection strategies at the filter update step. We examine the performance of these MRA-EnKF schemes in term of solution estimation accuracy and computational complexity on the one-dimensional Burger’s equation and compare the results to those obtained from performing the traditional EnKF on the fine mesh.

5.1 Introduction

Uncertainty propagation/reduction methods based on Bayesian inference are now widely applied to enhance the predictive capability of the geophysical fluid dynamics and hydrological models e.g. [15, 27, 28, 155, 173]. Within the Bayesian framework, uncertainty in model input is represented using random variables with known probability laws. Once data becomes available, this prior knowledge is updated via Bayes rule, which yields a generally correlated posterior with reduced uncertainty. Two popular approaches in Bayesian inversion have been extensively applied by the modeling community. The first approach uses a
Markov Chain Monte Carlo (MCMC) method to sample the posterior distribution. MCMC techniques require a large number of model runs in order to generate a large enough sample representing the posterior distribution [24, 174, 175]; this often renders the direct application of MCMC to sample large-scale models a computationally prohibitive exercise. Another popular approach for tackling the Bayesian estimation problem is the filtering method, in which the distribution is updated sequentially as the data become available [21]. Among many Bayesian filters, the most popular filtering approach in the field of geophysical fluid dynamics is the so-called Ensemble Kalman Filter (EnKF) and its variants e.g. [34, 36, 37, 38, 39, 40]. Motivated by a Monte Carlo formulation, EnKFs represent the Bayesian filter statistics through an ensemble of system states. These ensemble members, which represent realizations of the random state, are integrated using the dynamical model to estimate the forecast. Once data become available, a linear Kalman update is applied in the so-called analysis step to update the forecast with the new information. Despite its Gaussian underpinning, one of the main advantages of EnKF methods over MCMC techniques is the ability of the former to accommodate large dimensional state vectors at reasonable computational requirements [49, 51, 52, 53, 55, 125, 176].

In fluid dynamics simulations, the loss of solution accuracy due to limited spatial resolution can constitute a major source of uncertainties. Traditionally, a spatial mesh is used that is a priori sufficiently fine to capture underlying large- and small-scale dynamics, at the cost of a heavy computational burden. This motivated the development of adaptive mesh refinement (AMR) methods that allow the mesh to adapt its resolution locally according to the features of the solution [56, 57, 58, 59, 177]. AMR enables the simulated mesh to adapt both in space and time, assigning high spatial resolution to the areas in which the solution varies rapidly, and coarsening the mesh in regions of weaker variability. This makes AMR particularly useful for capturing sharp fronts and shock formations [60, 61]. By limiting the fine resolution to regions where it is required, AMR-based simulations significantly reduce the computational complexity and accordingly enhance the computational perfor-
formance. AMR approaches can be separated into two main classes. The first AMR class splits computational cells into finer cells on the same grid [178], whereas the second one constructs the adaptive grid over a multi-level mesh structure [177].

Recently, the possibility of performing data assimilation using adaptively discretized models grasped the attention of the ocean and hydrodynamics modeling community [57, 62, 63, 64, 65]. However, the literature dealing with the problem of implementing AMR and data assimilation techniques remains scarce. Variational data assimilation methods such as 3D-VAR and 4D-VAR were applied to adaptive mesh ocean and meteorological models in [66] and [67], respectively. The first attempt to combine a sequential data assimilation method, e.g. the EnKF, with an adaptive ocean model was proposed in [62]. The recent contribution by [63] is the first attempt to investigate the implementation of EnKF with multi-level AMR for data assimilation into large-scale coastal ocean models. To address the difficulties in computing the mean and error covariance required in the analysis step of the EnKF, the updates in [63] were performed by projecting each ensemble member from their own mesh to a “supermesh”, which is the union of all the meshes associated with the ensemble members.

In this paper, new approaches combining an EnKF and adaptive mesh models are proposed and analyzed. We rely on a Multiresolution Analysis (MRA) [179] approach to decompose the model solution into Haar wavelets. A convenient tree structure representation is used for this purpose [180]. Depending on the criterion specified by the user, the leaf nodes of the tree are removed or added to adapt the mesh. We design our combined MRA-EnKF schemes in such a way that the adaptive mesh refinement is integrated with the forecast steps of the EnKF in a straightforward fashion, similar to the multi-level AMR proposed in [63]. By defining different projection spaces in the Kalman update step, several MRA-EnKF variants are derived. To analyze the performance of these alternative schemes, we design a simplified system, namely the one-dimensional Burger’s equation in a periodic domain. This setting enables us to efficiently analyze the behavior of all proposed schemes,
and to contrast their performances, both in terms accuracy and computational cost, and consequently assess their relative merit. The gained experiences would enable us to infer the significance of these new schemes in the context of more generalized multi-dimensional settings, as recently performed in [62, 63].

The study is organized as follows. The background of the EnKF and MRA are outlined in Section 5.2. Section 5.3 presents in detail the derivation of several MRA-EnKF schemes. The test problem and the numerical scheme of our experimental system are presented in Section 5.4. The results of the numerical experiments are presented and discussed in Section 5.5. The main conclusions are summarized in Section 5.6.

5.2 Background

5.2.1 Multiresolution analysis

We consider the following transient hyperbolic partial differential equation on the one-dimensional domain \( \Omega = (0, L) \),

\[
\partial_t u + \partial_x f(u) = 0, \tag{5.1}
\]

\[
u(x, 0) = u_0, \tag{5.2}
\]

\[
u(0, t) = u(L, t), \tag{5.3}
\]

where \( u \) is a scalar field, \( u_0 \) is the initial condition, and \( f : \mathbb{R} \to \mathbb{R} \) is the flux function. We consider a regular discretization of the domain into \( N \) cells of fixed size \( h \). The boundaries of the cells are denoted \( (x_i)_{i=0}^N \), where \( x_i = ih \). The equation is spatially discretized using a finite volume scheme, and advanced in time using an explicit time integration scheme with an adaptive time step, \( \lambda \), that is restricted by the Courant-Friedrichs-Lewy condition. The resulting discrete system of equations is denoted by

\[
u^{s+1} = \nu^s - \lambda \left( f^+ - f^- \right), \tag{5.4}
\]
where \( u^s \in \mathbb{R}^N \) is the solution vector, \( f^{+,s} \) and \( f^{-,s} \) are the numerical fluxes at time \( t = t_s \).

The solution is represented using the piecewise constant expansion

\[
\begin{align*}
  u(x, t_s) & \approx U(x, t_s) = \sum_{i=1}^{N} u_{is} \chi_i(x), \\
  \chi_i & \text{ is the characteristic function of the cell } (x_{i-1}, x_i). 
\end{align*}
\]

To simplify the notation, the variable \( t_s \) is omitted in the rest of the section.

In order to compress the approximation of \( u \), we introduce a multiresolution analysis method based on the Haar transform. To this end, we consider the mother wavelet function \( \psi^0(x) \) and the associated scaling function \( \phi^0(x) \) defined by

\[
\psi^0(x) = \begin{cases} 
1 & \text{for } 0 \leq x < \frac{1}{2}, \\
-1 & \text{for } \frac{1}{2} \leq x < 1, \\
0 & \text{otherwise},
\end{cases}
\]

and

\[
\phi^0(x) = \begin{cases} 
1 & \text{for } 0 \leq x < 1, \\
0 & \text{otherwise}.
\end{cases}
\]

We define a rescaled version of the Haar functions according to:

\[
\phi(x) = \frac{1}{L^2} \phi^0 \left( \frac{x}{L} \right) \quad \text{and} \quad \psi_{n,j}(x) = \frac{2^n}{L^2} \psi^0 \left( 2^n \frac{x}{L} - j \right),
\]

where the two integer subscripts \( n \) and \( j \) indicate the resolution level and the position index of the rescaled Haar functions, respectively. This above family constitutes an orthonormal system with the inner product defined by \( \langle v, w \rangle_{L^2} = \int_0^L v(x)w(x)dx \). An expansion of \( U \) can be computed by direct projection, i.e.

\[
U(x) = \langle U, \phi \rangle_{L^2} \phi(x) + \sum_{n} \sum_{j} \langle U, \psi_{n,j} \rangle_{L^2} \psi_{n,j}(x).
\]

In this work, we restrict ourself to the case \( N = 2^{n_0} \). Given the piecewise constant structure of \( u \) at each time step, the same wavelet transform of the discrete solution can be efficiently
computed using a fast Haar transform \[18\]. The exact resulting expansion is

\[ U(x) = v_0 \phi(x) + \sum_{n=0}^{n_0-1} \sum_{j=0}^{2^n-1} v_{n,j} \psi_{n,j}(x), \tag{5.5} \]

where \( v_0 \) is the so-called average coefficient and \( v_n \in \mathbb{R}^{2^n} \) is the vector of detail coefficients at level \( n \). It is convenient to index the details using a binary tree structure \( T_p \) such that:

- the root of the tree is the pair \((0,0)\),
- a node has two children or none,
- a node without a child is a leaf,
- for each non leaf node \((n,j)\), the left (resp. right) son is \((n+1,2j)\) (resp. \((n+1,2j+1)\)),
- the maximum height of the tree is \( n_0 \).

The subscript \( P \) indicates that we are considering the perfect binary tree satisfying the above constraints. We denote by \( \mathcal{I}(T_p) \) the set of interior nodes of the tree and \( \mathcal{L}(T_p) \) the set of leaves. (Note that the basis functions, \( \psi_{n,j} \), are indexed in the same fashion as their support). For a node \( \alpha = (n,j) \), the support of \( \psi_\alpha \) denoted by \( \Omega_\alpha = \left[ \frac{L}{2^{n}} j, \frac{L}{2^{n}} (j+1) \right) \), and the support of the function associated to its left and right sons denoted by \( \alpha_l \) and \( \alpha_r \), respectively, form a partition of the support of the parent, i.e.

\[ \Omega_{\alpha_l} \cup \Omega_{\alpha_r} = \Omega_\alpha \quad \text{and} \quad \Omega_{\alpha_l} \cap \Omega_{\alpha_r} = \emptyset. \]

The wavelet expansion of \( U \) is then written as

\[ U(x) = v_0 \phi(x) + \sum_{\alpha \in \mathcal{I}(T_p)} v_\alpha \psi_\alpha(x), \]

with the sum taken over the interior nodes of the tree.
We introduce a thresholding operator $\Pi_\varepsilon$ that satisfies

$$\|U - \Pi_\varepsilon(U)\|_{L^2} \leq \varepsilon.$$  

The operator is defined such that the basis function associated to a node $\alpha = (n, j)$ is discarded if the following three conditions are satisfied:

- the coefficient $v_\alpha$ satisfies
  $$|v_\alpha| \leq 2^{-\frac{n}{2}} n_0^{-\frac{1}{2}} \varepsilon, \quad (5.6)$$

- all the nodes of the left and right subtrees of $\alpha$ satisfy condition $(5.6)$,

- the subtree $T$ of $\mathcal{T}_p$ associated to $\Pi_\varepsilon(U)$ is graded.

The graded condition means that the ratio of the sizes between two neighboring leaves does not exceed two. A graded tree is such that, for two leaves $\alpha_1 = (n_1, j_1)$ and $\alpha_2 = (n_2, j_2)$ with adjacent associated support $\Omega_{\alpha_1}, \Omega_{\alpha_2}$, the difference of depth is smaller than or equal to one. Formally, if there exists $x \in \Omega$ such that $\overline{\Omega}_{\alpha_1} \cap \overline{\Omega}_{\alpha_2} = \{x\}$, then $|n_1 - n_2| \leq 1$, where $\overline{\Omega}_{\alpha}$ is the closure of $\Omega_{\alpha}$. This condition is naturally extended to the boundary (i.e. in the case $x = 0$ or $x = L$). As a consequence, the function $\Pi_\varepsilon(U)$ can be written as

$$\Pi_\varepsilon(U)(x) = v_0 \phi(x) + \sum_{\alpha \in \mathcal{I}(T)} v_\alpha \psi_\alpha(x). \quad (5.7)$$

We also introduce the refinement operator $\mathcal{R}$ such that, for any function $w \in \mathcal{U}$, $\mathcal{R}(w) = w$ and the leaf node $\alpha \in \mathcal{I}(T)$ is split once if the depth of $\alpha$ is strictly smaller than $n_0 - 1$, and the resulting tree remains graded. An adaptive version of the scheme in Eq $(5.4)$ can be defined as:

$$u^{s+1} = \Pi_\varepsilon \left( \mathcal{R}(u^s) - \tilde{\lambda}^s (\tilde{f}^{+,s} - \tilde{f}^{-,s}) \right),$$

where $\tilde{\lambda}^s, \tilde{f}^{+,s}$ and $\tilde{f}^{-,s}$ are extended versions of $\lambda^s, f^{+,s}$ and $f^{-,s}$ accounting for the
anisotropic discretization. In the following, we denote by $\mathcal{U}_T$ the subspace of $L^2(0,L)$ defined by

$$\mathcal{U}_T = \text{span}\{\phi\} + \text{span}\{\psi_\alpha; \alpha \in \mathcal{T}\}.$$ 

Hence, if $\mathcal{T}_0$ is a subtree of $\mathcal{T}_1$, we have $\mathcal{U}_T \subset \mathcal{U}_T$.

### 5.2.2 Ensemble Kalman filter

The Kalman filter is a recursive technique to sequentially estimate the state of a linear dynamical system using available measurements. It provides the best unbiased estimates under the assumption that the underlying distributions are Gaussian \[89\]. We consider here the ensemble Kalman filter that is designed for large scale nonlinear filtering problems and in which the moments of the Gaussian distributions are estimated using a discrete sample. Given a state-space models defined as

$$v^k = \mathcal{M}(v^{k-1}) + \eta^k,$$

$$y^k = H^k v^k + \epsilon^k,$$

where $\mathcal{M}$ is the dynamical operator describing the time evolution of the state from time step $k-1$ to $k$, and $H^k$ a linear observation operator.

In this work, $v^k$ is a $\mathbb{R}^{\# \mathcal{T}}$-valued random vector representing the state defined by the expansion coefficients $(v_\alpha)_{\alpha \in \mathcal{T}}$ as in Eq. (5.7), where $\# \mathcal{T}$ denotes the number of internal nodes in $\mathcal{T}$. $y^k$ is a $\mathbb{R}^m$-valued random observation vector, $\eta^k$ is the model error and $\epsilon^k$ is the observation error. We assume that $\eta^k$ and $\epsilon^k$ are Gaussian with zero mean and covariance matrices $Q^k$, and $R^k$, respectively, i.e., $\eta^k \sim \mathcal{N}(0, Q^k)$ and $\epsilon^k \sim \mathcal{N}(0, R^k)$. $v^0$, $\eta^k$ and $\epsilon^k$ are further assumed to be mutually independent. For the sake of simplification, we omit the superscript $k$ in the rest of the section.

Given an ensemble of $N_e$ forecasts defining a matrix $V_f = [v^q]_{q=1}^{N_e} \in \mathbb{R}^{\# \mathcal{T} \times N_e}$ and the realization of the data $y$, the (stochastic) EnKF \[39, 55, 123\] updates the forecasts according
to:
\[
\mathbf{v}^a_q = \mathbf{v}^f_q + \mathbf{CH}^T (\mathbf{HCH}^T + \mathbf{R})^{-1} (\mathbf{y}_q - \mathbf{Hy}^f_q),
\]  
(5.8)
where \( \mathbf{y}_q = \mathbf{y} + \mathbf{e}_q, \ q = 1, \ldots, N_e \) is a perturbed observation vector, and \( \mathbf{C} \) is the sample covariance
\[
\mathbf{C} = \frac{1}{N_e - 1} \sum_{q=1}^{N_e} (\mathbf{v}^f_q - \bar{\mathbf{v}}^f) (\mathbf{v}^f_q - \bar{\mathbf{v}}^f)^T, \text{ with } \bar{\mathbf{v}}^f = \frac{1}{N_e} \sum_{q=1}^{N_e} \mathbf{v}^f_q.
\]
We denote by \( \mathbf{h}_q, \bar{\mathbf{h}}, \mathbf{C}_H \) and \( \mathbf{M} \) the quantities defined by
\[
\mathbf{h}_q = \mathbf{Hv}^f_q, \quad \bar{\mathbf{h}} = \frac{1}{N_e} \sum_{q=1}^{N_e} \mathbf{h}_q, \text{ and } \mathbf{C}_H = \frac{1}{N_e - 1} \sum_{q=1}^{N_e} (\mathbf{h}_q - \bar{\mathbf{h}}) (\mathbf{h}_q - \bar{\mathbf{h}})^T.
\]  
(5.9)
Equation (5.8) can thus be recast as
\[
\mathbf{v}^a_q = \mathbf{v}^f_q + \frac{1}{N_e - 1} \sum_{r=1}^{N_e} (\mathbf{v}^f_r - \bar{\mathbf{v}}^f) (\mathbf{h}_r - \bar{\mathbf{h}}) (\mathbf{C}_H + \mathbf{R})^{-1} (\mathbf{y} + \mathbf{e}_q - \mathbf{h}_q).
\]
Let \( G_{qr}, \tilde{G}_q \) and \( F_{qr} \) be defined by
\[
G_{qr} = \frac{1}{N_e - 1} (\mathbf{h}_r - \bar{\mathbf{h}}) (\mathbf{C}_H + \mathbf{R})^{-1} (\mathbf{y} + \mathbf{n}_q - \mathbf{h}_q), \quad \tilde{G}_q = \frac{1}{N_e} \sum_{r=1}^{N_e} G_{qr}, \text{ and } F_{qr} = G_{qr} - \tilde{G}_q.
\]
One can then show that
\[
\mathbf{v}^a_q = \mathbf{v}^f_q + \sum_{r=1}^{N_e} F_{qr} \mathbf{v}^f_r,
\]  
(5.10)
where the matrix \( F \) only depends on the ensemble members through the predicted observations ensemble \( \{\mathbf{h}_q\}_{q=1}^{N_e} \). Moreover, given the linear nature of the update of the members, the functional representation of the members is also updated as
\[
U^a_q(x) = U^f_q(x) + \sum_{r=1}^{N_e} F_{qr} U^f_r(x).
\]  
(5.11)
Hereafter, we refer to the term \( \sum_{r=1}^{N_e} F_{qr} U^f_r(x) \) in Eq. (5.11) as the ‘correction’ term. The
analyzed ensemble members are then integrated with the model, $\mathcal{M}$, in the forecast step until the next observation become available, a new EnKF update-forecast cycle then begins.

### 5.3 MRA-EnKF methods

This section will present the derivation of the EnKF update step for ensemble members defined on different grids. Five MRA-EnKF schemes employing different projection strategies on the grid on which the EnKF update is performed are proposed, namely, union grid (MRAEnKF), mean grid (MSPEnKF), and members grid (CrPEnKF and CnSPEnKF).

#### 5.3.1 EnKF update in the MRA settings

For a given time $t$, we denote by $U_q(x)$ the $q^{th}$ ensemble member, $1 \leq q \leq N_e$, and by $\mathcal{T}_q$ the tree associated with $U_q$. Let $\bar{U}$ be the mean of the ensemble $(U_q)^{N_e}$, i.e. the EnKF estimate, and $\bar{\mathcal{T}}$ be the tree associated with $\bar{U}$. Then $\bar{U}$ can be written as

$$
\bar{U}(x) = v_0 f(x) + \sum_{\alpha \in \mathcal{F}(\mathcal{T})} v_\alpha \phi(x),
$$

where the function spaces of $U_q$, $\bar{U}$ and $U$ are related by

$$
\mathcal{U}_{\mathcal{T}} \subseteq \bigcup_{q=1}^{N_e} \mathcal{U}_{\mathcal{T}_q} \subseteq \mathcal{U}_{\mathcal{T}_p}.
$$

To compute the mean $\bar{U}$, all ensemble members are projected onto the union subspace $\bigcup_{q=1}^{N_e} \mathcal{U}_{\mathcal{T}_q}$, on which the average is calculated.

Based on Eq. (5.11), one can modify the EnKF update step to accommodate the change in the subspace of the ensemble caused by adaptive mesh refinement during the forecast. The general form of the EnKF update with AMR is then written

$$
U^a_q(x) = \mathcal{P}_{\mathcal{U}_{\mathcal{T}}} \left[ U^f_q(x) + \mathcal{P}_{\mathcal{U}_{\mathcal{T}_p}} \left( \sum_{r=1}^{N_e} F_{qr} U^f_r(x) \right) \right],
$$

(5.13)
where $\mathcal{P}_{\mathcal{U}_i}$, $i = 1, 2$ is the L2 projection of a function onto a subspace $\mathcal{U}_i$ defined as

$$\mathcal{P}_{\mathcal{U}_2} u = \arg \min_{v \in \mathcal{U}_2} \|u - v\|_{L^2}.$$ 

Since $u$ and $v$ can be expressed with Haar wavelets, which constitute an orthogonal basis in L2 space, if $u(x)$ is such that

$$u(x) = v_0 \phi(x) + \sum_{\alpha \in \mathcal{I}(\mathcal{T}_1)} v_{\alpha} \varphi_{\alpha}(x),$$

and $v(x)$ is such that

$$v(x) = v_0 \phi(x) + \sum_{\alpha \in \mathcal{I}(\mathcal{T}_2)} v_{\alpha} \varphi_{\alpha}(x),$$

then the projection of $u$ on $\mathcal{U}_1$ is

$$\mathcal{P}_{\mathcal{U}_1} u = v_0 \phi(x) + \sum_{\alpha \in \mathcal{I}(\mathcal{T}_1) \cap \mathcal{I}(\mathcal{T}_2)} v_{\alpha} \varphi_{\alpha}(x) + \sum_{\alpha \in \mathcal{I}(\mathcal{T}_2) \setminus \mathcal{I}(\mathcal{T}_1)} 0 \cdot \varphi_{\alpha}(x),$$

which is amount to assigning the correct detail coefficients to the correct nodes that are both in $\mathcal{U}_1$ and $\mathcal{U}_2$, and assigning 0 as the coefficients for the nodes that are in $\mathcal{U}_2$ but not in $\mathcal{U}_1$.

An EnKF update performed on the finest grid containing all ensemble members corresponds to using the subspaces $\mathcal{U}_1 = \mathcal{U}_2 = \mathcal{U}_p$ in Eq. (5.13), and is equivalent to Eq. (5.11). By introducing a thresholding operator $\Pi_c$ and assigning a different function spaces to the $\mathcal{U}_i$s, we introduce in the following subsections several variants of adaptive mesh refinement EnKF algorithms based on the definition of the ensemble mean in Eq. (5.12) and the EnKF update in Eq. (5.13).
5.3.2 MRAEnKF

In the standard (non-adaptive mesh) EnKF, the updates are performed on a fixed space because the ensemble members and the filter estimate are defined on the same mesh; the update is thus straightforward. This is not the case for adaptive mesh where each ensemble member independently adapts its own grid, which prevents the computation of the first- and second-order statistical moments required for the EnKF algorithm. It is then natural to seek a common space of all members to apply the EnKF updates.

In the MRAEnKF method, the ensemble members and the members correction terms of the EnKF are projected onto the subspace spanning the union of the grids of all ensemble members. The updates are then performed according to Eq. (5.11). This corresponds to $\mathcal{U}_{\mathcal{F}} = \mathcal{U}_{\mathcal{F}} = \bigcup_{q=1}^{N_e} \mathcal{U}_{\mathcal{F}}$ in Eq. (5.13). Accordingly, $U^a(x)$ is written as

$$U^a_q(x) = \mathcal{P}_{\bigcup_{q=1}^{N_e} \mathcal{U}_{\mathcal{F}}} \left[ U_f^q(x) + \mathcal{P}_{\bigcup_{q=1}^{N_e} \mathcal{U}_{\mathcal{F}}} \left( \sum_{r=1}^{N_e} F_{qr} U_f^r(x) \right) \right].$$

We will hereafter refer to the proposed combined multiresolution analysis with EnKF techniques as the “MRA-EnKFs”, namely to avoid any confusion with the specific MRAEnKF method described above.

5.3.3 Forecast Mean Space Projection (FMSP) EnKF

In the limit of large ensemble, the EnKF converges to the Kalman filter (KF) given the Gaussian noise assumptions on the forecast and observation models [182, 183]. The mean estimate of the EnKF also converges to a maximum a posteriori (MAP) estimate of the KF. Even for the case of nonlinear models, it is a common practice to take the ensemble mean as the filter state estimate. Therefore, it is intuitive to consider performing the ensemble projection on the space of the mean $\mathcal{U}_{\mathcal{F}}$. As a consequence, we modify the MRAEnKF by carrying out the update of all members in the mean space, using Haar wavelets function representation.
In FMSPEnKF, all the members and correction terms are projected on the subspace of the mean forecast before the analysis step. Let $\mathcal{U}^f$ be the mean forecast and $\mathcal{U}_{\mathcal{F}f}$ the subspace associated with $\mathcal{U}^f$. FMSPEnKF is initialized by coarsening the mean of the forecast,

$$
\mathcal{U}^f = \Pi_e \left( \frac{1}{N_e} \sum_{q=1}^{N_e} \mathcal{U}^f_q \right). 
$$

(5.14)

$\mathcal{U}^f_q$ and the correction term $\sum_{q=1}^{N_e} F_q \mathcal{U}^f_r(x)$ are projected onto $\mathcal{U}_{\mathcal{F}f}$ before applying the EnKF update. This corresponds to setting $\mathcal{U}_{\mathcal{F}f} = \mathcal{U}_{\mathcal{F}f}$ in Eq. (5.13), which leads to

$$
\mathcal{U}^a_q(x) = \mathcal{P}_{\mathcal{U}_{\mathcal{F}f}} \left[ \mathcal{U}^f_q(x) + \mathcal{P}_{\mathcal{U}_{\mathcal{F}f}} \left( \sum_{r=1}^{N_e} F_q \mathcal{U}^f_r(x) \right) \right].
$$

Note that the update in FMSPEnKF is performed in $\mathcal{U}_{\mathcal{F}f}$, which is a subset of $\mathcal{U}_{\mathcal{F}f}$ for certain tolerance, $\epsilon$, the FMSPEnKF solution is expected to be less accurate than that of MRAEnKF, as the latter performs the update in the union space of all members. However, less computational complexity is anticipated for FMSP as it involves a smaller number of grid points.

### 5.3.4 Analysis Mean Space Projection (AMSP) EnKF

The order in which one applies the projection and the analysis may affect the performance of the MRA-EnKFs. To explore this, we introduce the Analysis Space Projection (AMSP) EnKF, which consists in another variant of mean space projection EnKF derived by rearranging the order of implementing the projection and analysis.

In contrast with the FMSPEnKF, the AMSPEnKF projects all the members on the subspace of the mean after the analysis step. Let $(\bar{\mathcal{U}}^a_q(x))^{N_e}$ be the intermediate analysis en-
semble, the filter starts by computing $\tilde{U}_q^a(x)$ with MRAEnKF:

$$
\tilde{U}_q^a(x) = \mathcal{P}_{\bigcup_{q=1}^{N_e} \mathcal{W}_q} \left[ U_q^f(x) + \mathcal{P}_{\bigcup_{q=1}^{N_e} \mathcal{W}_q} \left( \sum_{r=1}^{N_r} F_{qr} U_r^f(x) \right) \right].
$$

Here $\tilde{U}_q^a(x)$ is defined on the union subspace $\bigcup_{q=1}^{N_e} \mathcal{W}_q$. Then the coarsened mean of $(\tilde{U}_q^a(x))^{N_e}$ is calculated as

$$
\overline{U}^a = \Pi_e \left( \frac{1}{N_e} \sum_{q=1}^{N_e} \tilde{U}_q^a \right).
$$

The final analysis $(U_q^a(x))^{N_e}$ is obtained by projecting $(\tilde{U}_q^a(x))^{N_e}$ onto $\mathcal{W}_q^i$, namely according to:

$$
U_q^a(x) = \mathcal{P}_{\mathcal{W}_q^i} \left[ \tilde{U}_q^a(x) \right],
$$

where $\mathcal{W}_q^i$ is the subspace associated with $\overline{U}^a$.

### 5.3.5 Correction Projection (CrP) EnKF

FMSPEnKF and AMSPEnKF updates result in all members belonging to the space of the mean. All members are then integrated with the model in the forecast step starting from the same mesh. This may result in each member losing some of its details as possessed prior to the EnKF update. Projecting each member onto its own space before forecasting with the model may enhance the filter performance. Therefore, we also consider projecting each member back to its original space after the EnKF update.

In Correction Projection (CrP) EnKF, all the member corrections are projected on the mean before the analysis step. Each member is then updated in the mean space $\mathcal{W}_q^f$ as computed in FMSPEnKF (i.e. Eq. (5.14)), and the members are finally projected back to each member’s original subspace $\mathcal{W}_q$. Each member is then propagated forward to the next
assimilation step. This corresponds to setting $\mathcal{U}_1 = \mathcal{U}_{q_1}$ and $\mathcal{U}_2 = \mathcal{U}_{q_2}$ in Eq. (5.13). Since the CrPEnKF update is performed on the coarse mean subspace, considerable reduction in computational cost can be achieved with respect to EnKF.

### 5.3.6 Constant Space Projection (CnSP) EnKF

In Constant Space Projection (CnSP) EnKF, the forecast and the update of each member are performed entirely on its own mesh. No coarsening is involved in the scheme. This requires projecting the correction term onto subspace $\mathcal{U}_{q_1}$, not $\mathcal{U}_f$ or $\mathcal{U}_a$. The update step of CnSP therefore corresponds to setting $\mathcal{U}_1 = \mathcal{U}_2 = \mathcal{U}_{q_1}$ in Eq. (5.13). Because each member is updated independently in its own space without gaining any extra mesh details from other members, we expect the CnSPEnKF to produce less accurate estimates compared to the other methods as time evolves.

Table 5.1 summarizes the update algorithms of the EnKF and its multiresolution variants. The differences between these schemes arise from the order of application of the update and the coarsening steps. They also depend on the spaces of ensemble members projection.

### 5.4 Computational test problem

#### 5.4.1 Model

We focus on the one-dimensional Burger’s equation with periodic boundary conditions, i.e. Eq. (5.1) with the flux defined as

$$f(u) = u^2 / 2.$$  

In the numerical experiments, the uncertain initial condition corresponds to the shifted Gaussian:

$$u(x, t_0) = b + a \cdot e^{-(x-\mu)^2/\sigma^2},$$  \hspace{1cm} (5.15)
<table>
<thead>
<tr>
<th>Method</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>EnKF</td>
<td>$\mathcal{U}_{\mathcal{F}<em>1} = \mathcal{U}</em>{\mathcal{F}<em>2} = \mathcal{U}</em>{\mathcal{P}}$.</td>
</tr>
<tr>
<td>MRAEnKF</td>
<td>$\mathcal{U}_{\mathcal{F}<em>1} = \mathcal{U}</em>{\mathcal{F}<em>2} = \bigcup</em>{q=1}^{N_e} \mathcal{U}_q$.</td>
</tr>
</tbody>
</table>
| FMSPEnKF   | 1. $\mathcal{U}^f = \Pi_{\varepsilon} \left( \frac{1}{N_e} \sum_{q=1}^{N_e} \mathcal{U}_q^f \right)$.  
            | 2. Obtain $\mathcal{U}_{\mathcal{F}_f}$.  
            | 3. $\mathcal{U}_{\mathcal{F}_1} = \mathcal{U}_{\mathcal{F}_f}$ and $\mathcal{U}_{\mathcal{F}_2} = \mathcal{U}_{\mathcal{F}_f}$. |
| AMSPEnKF   | 1. Compute $(\tilde{U}_q^a)^N_{\varepsilon}$ with MRAEnKF  
            | 2. $\tilde{U}_a = \Pi_{\varepsilon} \left( \frac{1}{N_e} \sum_{q=1}^{Q} \tilde{U}_q^a \right)$.  
            | 3. Obtain $\mathcal{U}_{\mathcal{F}_a}$.  
            | 4. $\mathcal{U}_{\mathcal{F}_1} = \mathcal{U}_{\mathcal{F}_a}$ and $\mathcal{U}_{\mathcal{F}_2} = \bigcup_{q=1}^{N_e} \mathcal{U}_q$. |
| CrPEnKF    | 1. Compute $\mathcal{U}_{\mathcal{F}_f}$ (See FMSP).  
            | 2. $\mathcal{U}_{\mathcal{F}_1} = \mathcal{U}_{\mathcal{F}_a}$ and $\mathcal{U}_{\mathcal{F}_2} = \mathcal{U}_{\mathcal{F}_f}$. |
| CnSPEnKF   | $\mathcal{U}_{\mathcal{F}_1} = \mathcal{U}_{\mathcal{F}_2} = \mathcal{U}_{\mathcal{F}_a}$. |

Table 5.1: Summary of EnKF and its multiresolution variants.
where \( b \sim \mathcal{U}(0.5, 1), a \sim \mathcal{U}(1, 3), \mu \sim \mathcal{U}(1, 4) \) and \( \rho \sim \mathcal{U}(0.1, 0.5) \). The size of the domain is equal to \( 4\pi \).

### 5.4.2 Roe flux calculation

According to Eq. (5.4), the solution at the \( i^{th} \) cell is advanced as:

\[
 u_i^{s+1} = u_i^s - \lambda^s \left( f_i^{+s} - f_i^{-s} \right),
\]

(5.16)

where \( f_i^{+s} \) and \( f_i^{-s} \) are the corresponding right and left numerical fluxes, respectively. In the case of a fine mesh, the domain is discretized into \( 2^{n_0} \) elements of equal size, where \( n_0 \) represents the maximum depth of the multiresolution representation. In the experiments, the maximum depth is 10; therefore the fine mesh has 1024 cells.

As mentioned earlier, the solution is advanced using an adaptive explicit time integration scheme. Specifically, the time step, \( \lambda^s \), is adjusted according to \( \lambda^s = \text{CFL} \times h/\Lambda \), where \( \Lambda \) is the maximum absolute value of \( u^s \) and CFL is a user-defined positive value \( \leq 1 \).

Let us denote \( \hat{u} \equiv (u_l + u_r)/2 \) the Roe state, and \( \triangle u \equiv u_r - u_l \) the state jump. The classical Roe flux is

\[
 F(u_l, u_r) = \frac{[f(u_l) + f(u_r)]}{2} - \frac{|\hat{u}|}{2} \triangle u.
\]

To prevent the Roe flux from violating the entropy condition, the entropy fix is applied, which consists in insuring sufficient diffusion in rarefaction problem with \( u_l < 0 < u_r \). Let

\[
 \delta = \max\{0, \hat{u} - u_l, u_r - \hat{u}\},
\]

and define

\[
 \hat{q} = \begin{cases} 
 \hat{u}, & |\hat{u}| \geq \delta \\
 \frac{\hat{u}^2}{2\delta} + \frac{\delta}{2}, & |\hat{u}| < \delta.
\end{cases}
\]
The fixed Roe flux is
\[ F(u_l, u_r) = \left[ \frac{f(u_l) + f(u_r)}{2} \right] - \frac{|q|}{2} \triangle u, \]
and the right and left fluxes in Eq. (5.16) are defined through \( f^+ = F(u^+_i, u^+_{i+1}) \) and \( f^- = F(u^-_{i-1}, u^-_i) \), respectively. Note that, here, \( u_l = u^+_i \) and \( u_r = u^+_{i+1} \) for \( f^+ \), and \( u_l = u^-_{i-1} \) and \( u_r = u^-_i \) for \( f^- \). Furthermore, because periodic boundary conditions are used, \( f^+_N = f^-_1 = F(u^+_N, u^+_1) \), where \( N \) is the number of cells in the domain.

### 5.4.3 Multiresolution Roe solver

Given a sequence of \( N \) values on the fine grid \( u = \{u_j\}_{j=1}^N \), [179] showed that there is a one-to-one transformation between \( u \) and its multiresolution representation, \( U \). The transformation can be expressed as:
\[ U = Mu, \]  
(5.17)
where \( M \) depends on the interpolation method implemented by the multiresolution analysis scheme. In this work, the central interpolation is used, and therefore \( M \) is a linear operator expressed by an \( N \times N \) matrix.

We can symbolically recast Eq. (5.4) in the form of cell-wise update on the finest grid according to
\[ u_{i+1}^s = u_i^s - \lambda^s \left( f^+ - f^- \right) \equiv (E \cdot u^s)_i, \quad 1 \leq i \leq N, \]  
(5.18)
where \( E \) is a nonlinear operator representing the action of the explicit update. The multiresolution scheme is obtained by applying \( M \) to Eq. (5.18), resulting in:
\[ U^{s+1} = U^s - \lambda^s M \left( f^+ - f^- \right) = ME \cdot (M^{-1} U^s) \equiv E_M \cdot U^s. \]  
(5.19)
Here \( E_M \) is a multiresolution version of the nonlinear operator \( E \). The thresholding operator can be introduced at the beginning of each time step to incorporate the compression of the
numerical solution, which leads to:

\[
\mathbf{M} \mathbf{E} \cdot (\mathbf{M}^{-1} \Pi_{\varepsilon} (\mathbf{U}^s)) \equiv \mathbf{E}_{\mathbf{M}}^s \cdot \mathbf{U}^s. \tag{5.20}
\]

It can be shown that [179], given an appropriate tolerance value and a monotone scheme \( \mathbf{E} \), the error between the compressed and the full model solution is of the same order of magnitude as the tolerance, which can be written as

\[
\| \mathbf{E}_M^s \cdot \mathbf{U}^s - \mathbf{E}_M \cdot \mathbf{U}^s \|_1 \leq \varepsilon, \tag{5.21}
\]

where \( \| \mathbf{x} \|_1 \) is the L_1 norm of the vector \( \mathbf{x} \).

Given the thresholding parameter \( \varepsilon \) and the resolution level \( n_0 \), the algorithm starts by constructing a tree \( \mathcal{T}^0 \) and the subspace \( \mathcal{U}^0 \), which are used to represent the cell average initial condition. After the discrete solution is initialized, it is evolved in time following an iterative process. This consists of five main steps: a refinement of the tree \( \mathcal{T} \) at the beginning of every time step, the evaluation of \( \mathcal{U} \) at the centers of the cells, the computation of the time step satisfying the CFL condition, a time advance of the solution, and finally projection and coarsening. We summarize the multiresolution-based Roe solver as follows.

**Steps I. and II.** The approximation space is refined in anticipation of additional steepening that may develop in the solution. For this purpose, the grid is refined to add one more level of resolution. Specifically, the tree leaves are refined according to the enrichment strategy presented in [180]. After the fine grid is obtained, the solution is evaluated at the centers of the cells.

**Step III.** The optimal time step for solution update is computed based on the prescribed CFL limit.
Step IV. In this step, the model integration from time $t_s$ to $t_{s+1}$ is performed. The solution in an individual cell is updated by integration of the Roe fluxes through the cell’s boundary as in Eq. (5.16).

Step V. After the time integration, the Haar transform is used to recover the multiresolution representation of the solution. Then the coarsening operator $\Pi_c$ described in Section 5.2.1 is applied. This step defines the details of the tree and provides a compressed solution for the next iteration.

5.5 Numerical experiments

In this Section, we assess the performance of the proposed MRA-EnKF schemes in term of estimation error and complexity with respect to both a reference solution and the EnKF solution.

5.5.1 Twin experiments

A twin experiment is designed to assess the performance of the proposed assimilation schemes against the EnKF, based on their efficiency to estimate the reference solution. The reference solution is computed by solving the Burgers equation with an initial condition generated by sampling the parameters in Eq. (5.15). From the reference solution, we extract the data at locations $x = 4, 5.5, 7, 8.5, 10$ and $11.5$, respectively. The data are collected at regular time intervals corresponding to the assimilation frequency of the EnKF schemes, i.e. every second. These data are perturbed with measurement noise $\varepsilon_k \sim \mathcal{N}(0, R = \sigma^2 I)$, with $\sigma^2 = 0.09$ and $I$ the identity matrix. The initial ensembles of the EnKF and MRA-EnKFs are generated by sampling the parameters in Eq. (5.15) from the same set of priors used to generate the initial condition of the reference solution. In each set of assimilation experiments, all filters are initialized from the same ensemble, and assimilate the same observations. We also perform the experiments with varying ensemble size and tolerance to
Figure 5.1: Reference solution vs EnKF solution.

holistically examine the behavior of the proposed MRA-EnKFs framework.

5.5.2 MRA-EnKFs vs. Reference

In the first experiment, the settings for the multiresolution Roe solver are $\varepsilon = 10^{-3}$, $N_0 = 1024$, and $CFL = 0.9$. We run the simulation for 1 unit of time with 40 assimilation steps (one assimilation every 0.025 unit of time). Figure 5.1 illustrates the result of a fine grid approximation of the Burger’s equation solution. Both the reference solution and the EnKF estimate with similar fine grid resolution are shown. The initial condition of the reference solution and of the ensemble members are illustrated in the first subplot of Figure 5.1. The following subplots depict the evolution of the system at different times. The black solid
line is the reference solution at a fixed time step and the red solid line is its EnKF estimate. A total of 48 ensemble members are used for the EnKF. The evolution of the ensemble members is plotted along with the 3 standard deviation bounds to represent the ensemble spread. The ensemble members are marked with the blue lines and the 3 standard deviation bounds around the mean are highlighted in transparent blue. As time evolves, the solution develops a shock between \( x = \frac{3}{2}\pi \) and \( x = 2\pi \) at \( t = 0.2 \) (8\(^{th}\) assimilation iteration). This discontinuity in the form of a shock is the result of the compressibility due to different wave speed at different points when the hyperbolic PDEs are solved [179]. This shock formation is well captured by the Roe scheme. It is also clear from the figures that, as the assimilation cycle advances, the ensembles spread around the mean decreases and all members move closer to the reference solution. The filter well maintains the spread around the shock formation.

Initially, the estimate is far from the reference solution. As time evolves, the members start to converge toward the truth. The shock pattern is also well recovered by the EnKF. However, at the 16\(^{th}\) assimilation step, we observe the phase shift of the shock: the shock pattern estimated by the EnKF has slightly preceded the reference solution. This is because, at the initial time step, some ensemble members are shifted to the front with respect to the reference initial function. EnKF hardly recovers the head and the tail of the shock due to large ensemble variance at these locations.

The time evolution of MRA-EnKFs estimates are plotted together with the EnKF and the reference solutions in Figure 5.2. With a sufficiently large maximum level of the tree, \( n_0 \), and ensemble size (48 members), all MRA-EnKFs solutions converge to the reference as the assimilation advances. Most MRA schemes appear to perform as good as EnKF in term of accuracy of the estimates, except CnPEnKF which slightly deviates away from the rest, depicted by the dashed red plot around the tail of the shock formation in the 8\(^{th}\) assimilation step of Figure 5.2. This is because whereas other MRA-EnKFs rely on the union space for the updates, CnPEnKF performs the model integration and the update
purely in the space of the members, which may result in loss of solution details.

5.5.3 Assessment of MRA-EnKFs performances

Two main criteria are used to evaluate the efficiency of a MRA-EnKF; estimation accuracy and computational complexity. The MRA-EnKFs results are evaluated against that of the EnKF, which is taken as the reference. This is justified because the EnKF is always performed on the fine mesh, and thus is considered to be the most accurate filter. An MRA-EnKF scheme is considered efficient if its estimation error is close to that of EnKF and its computational complexity is significantly less than that of the EnKF. A number of factors, such as the ensemble size and the value of thresholding parameter (the tolerance) play an important role in evaluating the overall performance of the MRA-EnKF schemes.
MRA-EnKFs convergence analysis

We first consider the time evolution of the normalized L2 error to examine the convergence of MRA-EnKFs estimates toward the EnKF estimate defined as

$$
L2\text{-error} = \frac{\mathbb{E}\left(\int_{\Omega} \left( u(x) - u^{\text{EnKF}}(x) \right)^2 \, dx \right)^{\frac{1}{2}}}{\| u \|_{L_2}},
$$

where $u^{\text{EnKF}}$ denotes the EnKF solution and $\| u \|_{L_2}$ denotes the L2 norm of $u$.

Figure 5.3 plots the time evolution of the normalized L2 error in log scale for each of the MRA-EnKF schemes. The x-axis is the assimilation time step and the y-axis is the average L2 error in log scale. The plotted errors are averages over five assimilation runs with different initial ensembles. The results are consistent with those of Figure 5.2 with CnSPEnKF producing the largest discrepancy from the EnKF. As time advances, the L2 error produced by CnSPEnKF is clearly larger than that of the other schemes. The overall decreasing L2 error trends over time indicate the relevance of the MRA-EnKFs for approximating the EnKF solution.

We also examine the differences between the MRA-EnKF estimates and the EnKF estimate as the root of the square of the differences of MRA-EnKF and EnKF solutions. Figure 5.4 shows the time evolution of the RMSE for each MRA-EnKF scheme. The x-axis is the assimilation step and the y-axis is the RMSE in log scale. Clearly, for this particular tolerance ($\epsilon = 10^{-3}$), MRAEnKF leads to the smallest error at every assimilation cycle. This is expected because the updates are performed in the members union space with high mesh resolution. The RMSE gradually increases with time for all MRA-EnKF schemes, except CnSPEnKF which shows a RMSE decrease after starting from a large error before it increases again toward the end of the assimilation window.

To investigate the sensitivity of the MRA-EnKFs solutions to a change in the tolerance
Figure 5.3: Normalized L2 error versus assimilation step.

Figure 5.4: RMSE versus assimilation step.
value, the twin experiment was repeated with different values of $\epsilon$. Figure 5.5 plots the time evolution of RMSEs with varying tolerance values ranging from $10^{-3}$ to $3 \times 10^{-1}$. For all schemes, larger tolerance results in larger errors, as expected, as these lead to coarser meshes. With the exception of CnSPEnKF, the same RMSE evolution trend is observed for all schemes, with the RMSE gradually increasing and leveling off at the later timesteps for all tolerances.

**Estimation error vs. Computational complexity**

The computational complexity of data assimilation with an adaptive mesh model is estimated based on the number of the Roe flux computations, which depends on the number of cells used to approximate the solution. Adapting the mesh of each member independently may significantly reduce the model integration cost in the forecast step. Previously, we demonstrated that, given a specific tolerance, the first four MRA-EnKF schemes show no significant differences in terms of estimation error. In this subsection, we simultaneously analyse the estimation error and computational complexity of the MRA-EnKF schemes.

Figure 5.6 plots the time-integrated RMSE and computational complexity of each MRA-EnKF scheme for varying tolerance and ensemble size. Each subplot representing individual scheme shows normalized complexity on the $x$-axis against time-integrated RMSE in log scale on the $y$-axis. We define time-integrated RMSE as the total sum of the RMSE over all assimilation steps. The normalized complexity is defined as a ratio between the number of computed fluxes of the MRA-EnKF scheme and the number of computed fluxes in the EnKF (which runs on the fine mesh). Each curve of the plots represents a given ensemble size but varying tolerance.

The first observation we make from Figure 5.6 is that, with the exception of CnSPEnKF, as we increase the computational complexity by decreasing the value of the tolerance, the error decreases for all schemes and for all ensemble sizes. This is because a smaller tolerance yields higher mesh resolution, and hence more accurate estimates. The second
Figure 5.5: RMSE versus assimilation step. For each scheme, curves are generated for different tolerances, as indicated.
observation is that for all schemes, increasing the ensemble size beyond $N_e = 16$ does not significantly improve the accuracy of the estimates. This suggests that for a sufficiently large ensemble, the filtering error is dominated by mesh resolution (coarsening). The third observation is that, except for CnSPEnKF, large drops in the error occur between the normalized complexity of 0.4 and 0.6. This indicates a range with in which the trade-off between error and complexity is optimal, which provides the best performance of MRA-EnKF schemes, in term of both estimation error and computational cost. Except for CnSPEnKF, all MRA-EnKF schemes successfully reduce the computational complexity up to 40%, with an almost negligible time-integrated RMSE of less than $\mathcal{O}(10^{-10})$ when a small tolerance ($\varepsilon = 10^{-10}$) is used.

The above results suggest that, for a specific scheme, increasing the ensemble size beyond some threshold does not improve the filter’s performance. However, for a specific ensemble size, the schemes may exhibit different behaviors. We compare the error and complexity of the MRA-EnKFs in Figure 5.7. The first four MRA-EnKF schemes produce approximately the same error-complexity plots with small ensembles ($N_e \leq 16$). This indicates that for small $N_e$, the error in the assimilation system is dominated by ensemble sampling errors, and not by grid coarsening. For larger ensembles ($N_e > 16$), the error-complexity plots become more distinct. First, for $N_e = 24$, with the normalized complexity of 0.57, FMSPEnKF produces almost two orders of magnitude larger error than the other schemes, except CnSPEnKF. This result suggests that for a specific ensemble size and tolerance, some schemes may perform better than the others in reducing the estimation error. Similar conclusions can be made for larger ensembles ($N_e > 24$); with the exception of CnSPEnKF, FMSPEnKF yields slightly larger error than the other schemes within the 0.4-0.6 normalized complexity interval. The MRAEnKF provides the lowest error for any ensemble size and tolerance, which is expected given that MRAEnKF performs the analysis on the union space that yields very high mesh resolution compared to the other schemes. As we increase the ensemble size to 96 members, the differences between the
Figure 5.6: time-integrated RMSE versus normalized complexity. For each scheme, curves are generated for different ensemble size, as indicated.
different filtering schemes in term of error become clear. MRAEnKF and CrPEnKF yield approximately the same smallest error for all tolerances. This suggests that by projecting each member back to its original subspace before proceeding with the forecast step, CrP greatly improves the filter performance compared to the other schemes. Nevertheless, the results in Figure 5.7 suggest that, given the same tolerance, the computational complexity of the first four schemes are quite similar regardless of the ensemble size.

We also directly investigated the effect of the change in the error to the change in the tolerance. In Figure 5.8, the errors are plotted against the tolerance used in the MRA-EnKF schemes. Each curve represents the RMSE for a single ensemble size and varying tolerances. Similar to our previous observations, these plots suggest that for all tolerances, increasing the ensemble beyond a certain size does not significantly improve the performances of these schemes in term of estimation error. Particularly for large tolerances, using large ensembles has a small impact on the filters results. With the exception of CnSPEnKF, the linear trend of the error versus tolerance plots is clear. For some tolerance interval ($10^{-10}$ to $10^{-2}$), the RMSE is approximately of the same order of magnitude as the tolerance. This is consistent with the inequality in Eq. (5.21), which shows that the error between the numerical solution and its coarsened multiresolution counterpart is bounded by the tolerance.

Overall, with the exception of CnSPEnKF, the results suggest that all schemes are comparatively efficient at approximating the EnKF solution computed on the fine gird. These effectively capture the details of the reference solution given appropriate tolerances, and are all competitive in term of reducing the computational complexity. These schemes demonstrate the ability to significantly reduce the computational complexity by half while providing a RMSE comparable to that of the EnKF.
Figure 5.7: time-integrated RMSE versus normalized complexity. For each ensemble size, curves are generated for different MRA-EnKF scheme, as indicated.
Figure 5.8: time-integrated RMSE versus tolerance. For each scheme, curves are generated for different ensemble size, as indicated.
5.6 Conclusions

We proposed a new framework combining a sequential data assimilation technique, namely the Ensemble Kalman Filter (EnKF), with adaptive mesh models. Specifically, a new class of adaptive mesh EnKF schemes has been developed, relying on a Multiresolution Analysis (MRA) approach to decompose the solution into Haar wavelets. A binary tree structure is used to index wavelet supports, which facilitates the implementation of coarsening and refinement operations. Based on user defined tolerances, the latter enable effective mesh adaptation of individual solutions represented by the ensemble members in the EnKF. Projections onto the wavelet spaces are incorporated into the EnKF update equation, which leads to the formulation of different MRA-EnKF schemes. By following different strategies for projecting and coarsening the ensemble members, and considering different ordering of projection and update operations, five different MRA-EnKF schemes were constructed. The first scheme, called “MRAEnKF” projects all members to the members union space for the EnKF update. The second scheme, FMSPEnKF, updates the members on the coarsened mean forecast space. The AMSPEnKF scheme (the counterpart of FMSPEnKF) updates the members on the coarsened mean analysis space (instead of the mean forecast space). The fourth proposed scheme, CrPEnKF, updates the members using FMSPE first, then projects each member back onto the mesh of the previous forecast step. In the fifth scheme, CnSPEnKF, the analysis of each member is performed on the corresponding mesh, solely relying on the coarsening and refinement algorithms of the MRA.

These variants were evaluated against a fine-grid EnKF solution in a twin experiments setting involving the 1D Burger’s equation. The analysis revealed that with an appropriate choice of tolerance, all MRA-EnKF methods accurately recovered the reference free-run solution except for CnSPEnKF, which poorly performed compared to the other schemes. This was not surprising because CnSPEnKF performs all operations in the space defined by each member. The numerical results also revealed that the other four methods accurately approximated the EnKF solution and lead to significant computational savings. Specifi-
cally, only marginal differences with the EnKF were obtained while computational complexity reductions of up to 50% were achieved relative to the reference fine grid solution.

These preliminary results also provide an insight into potential extensions of the proposed framework. Particularly, because with the same tolerance all methods led approximately to the same computational saving, one may select the preferred approach based on ease of implementation. In this regard, we found that for the same value of tolerance, the performance of MRAEnKF and CrPEnKF are similar for small ensembles. However, CrPEnKF requires several more projection steps to implement the algorithm. One may thus consider MRAEnKF as being a more convenient scheme.

The present work focused on an adaptive mesh formulation based on MRA, and restricted our application to a one-dimensional setting. Extension of the present constructions to multi-dimensional problems is in principle straightforward, particularly if one exploits generalizations of the presently used binary tree constructions, and coarsening/refinement operators (e.g. [180]). Though this would naturally involve more elaborate adaptive mesh algorithms, the computational savings achieved in multiple dimensions are generally anticipated to be much more substantial than in one space dimension. We finally note that our selection of the present mesh adaptation formalism was in large part based on the analytical capabilities of the used MRA approach. However, one should note that the developed adaptive mesh EnKF methodologies can be readily implemented with different discretization approaches and adaptation strategies. Such generalizations will be the subject of future studies.
Chapter 6

Summary and Concluding Remarks

6.1 Contribution of the Thesis

This thesis presented an in-depth study of Bayesian inference and data assimilation in the context of hydrological and coastal ocean modeling. Any coastal marine simulations are always subject to significant degrees of uncertainty, which can originate from several sources of errors. The goal of this thesis was, therefore, to identify, quantify and reduce these uncertainties, toward establishing a UQ framework for efficient coastal ocean forecasting. Great efforts were invested to improve the accuracy of the empirically derived SWEs bottom stress parameter so-called ‘Manning’s $n$ coefficients’, which are considered the major source of errors for storm surge and tidal forecasts. Both sampling-based approach (i.e. MCMC) and data assimilation approach (i.e. EnKFs) were considered for the parameter estimation/inversion. We used observation simulation system experiments (OSSEs) to assess the performances and compare these inference approaches, with the aim to study the merits, disadvantages, and their inter-relations. To achieve these goals, we also exploited various model reduction techniques from the field of UQ such as KL expansion and PC expansion, etc., to enhance the performance of the developed inference schemes, both in term of estimation accuracy and computational cost saving. In the final chapter, we expended the scope of the thesis to include model discretization uncertainty. We have done so by introducing a new data assimilation approach that enables the EnKF implementation on adaptive mesh model, based on the multiresolution analysis. This particular contribution is important for the ocean modeling community where the use of adaptive mesh refine-
ment in the simulation has recently become more common. More specifically, the main contributions of the thesis can be summed up in four points:

1. The results presented in the first chapter is among the first few studies to assess the performance of the EnKF for parameter estimation in coastal ocean settings against the benchmark MCMC, both in terms of the estimates and their corresponding uncertainty. This comparison between EnKF and MCMC based on a large-scale coastal ocean system was made possible through the use of polynomial chaos surrogate as the substitute of the full large-scale ADCIRC model to efficiently compute the likelihood in the MCMC sampler for the sampling of the posterior pdf. With a fine-tuned EnKF using appropriate ensemble size and inflation, the posteriors obtained from the EnKF were found to be good approximations, at least in terms of the first two moments, of those obtained from PC-based MCMC. The EnKF successful inference results were achieved on the grounds that the dynamics of the tidal simulations and the relation between the water elevations and Manning’s $n$ are weakly nonlinear. The Gaussian noise assumptions imposed on the EnKF still hold and well represents the uncertainty of the parameters.

2. Having validated the relevance of the EnKF for parameter estimation in the first chapter, we then use SEIK filter, a deterministic variant of EnKF, to infer spatially-varying field Manning’s $n$ coefficients in the second chapter. We studied the effect of increasing the ensemble size and observation points on the estimation result. We also introduced the iterative SEIK which repeatedly perform the analysis using the same set of incoming observations until a stopping criteria is met. This iterative scheme is equivalent to minimizing the cost function of the nonlinear least squares problem with respect to the model parameters, and is therefore expected to further enhance the estimation accuracy. Motivated by the work in the subsurface simulations community, we used a Karhunen-Loève (KL) decomposition to enhance the parameter search space. In the KL subspace, the Manning’s $n$ field can be efficiently represented
using a few dominant modes. Consequently, the EnKF correction was applied in the subspace of the dominant directions of the parameter search space. This enhancement is achieved on the assumption that the parameter field is a stochastic process following a certain correlation model and the stochastic field is constrained by a set of collected data. For a small KL subspace, only small ensembles are required for the inference result to outperform those inferred from the full space. As we increase the ensemble size, increasing the number of KL modes is also required to obtain comparable inference performances with respect to those of the full space cases. Overall, the iterated-based inference in the enhanced reduced KL space significantly boosts the SEIK estimation accuracy compared to the node-wised parameter update in the full space.

3. The EnKF inference of Manning’s $n$ coefficients in the KL space presented in the second chapter is limited to the subspace defined by a given prior covariance matrices. In the third chapter of the thesis, we considered the uncertainty in the prior covariance model, which is expressed by viewing the prior covariance hyper-parameters (denoted by $q$) as random variables. We then inferred these hyper-parameters alongside the KL coordinates using a polynomial-based MCMC method. Instead of including the hyper-parameters as the additional stochastic dimensions in the expensive PC surrogate construction, we resorted to the so-called 'generalized KL expansion', which map the KL coordinates from any $q$-dependent basis to the KL coordinates in the reference basis using the $q$-dependent coordinate transformation function. This coordinate transformation was used together with the PC surrogate of the ADCIRC model, built based on the fixed $q$-averaged covariance, in the MCMC sampler to evaluate the posterior of the KL coordinates and $q$. As an added contribution, we built the surrogate of the coordinate transformation operator offline, which can indefinitely be used in the inference step. The evaluation of the change of basis using the PC surrogate instead of the KL decomposition is fast. Hence, it allows the implementation
of the change of coordinates (COC) for inference in the large-scale coastal ocean systems. By incorporating the prior hyper-parameters uncertainty in the OSSEs, we found that the MCMC recovers much better the spatially-varying Manning’s $n$ fields, comparing to the cases where $q$ are assumed given.

4. The scope of the thesis has been finally expanded to includes adaptive grid uncertainty to the inference problem. The final chapter of the thesis introduced a new and innovative approach to accommodate the EnKF inference on adaptive mesh models. A multiresolution analysis (MRA) based on Haar wavelets was used to build a fast and robust Adaptive Mesh Refinement (AMR) algorithm for 1D fluids wave simulation. Orthogonality of Haar wavelets allows for the straightforward projection of the mesh from one space to another. Hence, the ensemble members which originally reside on the different meshes can all be projected into the union mesh for the calculation of the sample ensemble mean and covariance matrix required in the EnKF analysis step. Five variants of the combined MRA and EnKF algorithm based on different projection strategies were proposed. This variety offers future users the flexibility to resort to their preferred MRA-EnKF schemes based on one’s convenience. The MRA-EnKFs were evaluated against a fine-grid EnKF solution in an OSSEs involving the 1D Burgers equation. All introduced MRA-EnKF schemes, except one, were found to yield estimates with similar accuracy to the benchmark EnKF solution. However, they require only 50% overhead with respect to the EnKF when an appropriate tolerance parameter for mesh coarsening and refinement was selected. This computational saving is expected to be even more substantial in the system with larger space dimension.
6.2 Future Research Work

This thesis serves as a benchmark to formulate a comprehensive UQ framework for coastal ocean forecasting systems. Several aspects of this framework remain open for exploration. Currently, we envisage our future work to involve the broader and deeper investigation in the following directions.

- This thesis focuses on parameter inversion/estimation in relatively weakly nonlinear dynamical systems. Future work will consider the state-parameter estimation in a stronger nonlinear setting (e.g. the coupling between tides and storm surges). In this scenario, the presence of non-Gaussian and non-linear features may not be properly treated by the EnKF. This is expected to be more pronounced when the state is a strongly nonlinear function of the parameter. To this end, the combined Ensemble Kalman and Particle Filters schemes such as the Gaussian-mixture filters [176, 49, 50], the Localized Ensemble Kalman Particle Filter (LEnKPF) [184] and EnKF-PF [185] etc., can be exploited to tackle the aforementioned issues.

- Tides and storm surges are related to estuarine flooding events. Wetting and drying (WD) algorithms in numerical models based on the SWEs are essential for simulating coastal inundation [186, 187]. Data assimilation on the coastal ocean model (e.g., ADCIRC) with wet and dry capability is yet to be considered. Although variational data assimilation has already been tested on this type of problem, the EnKF literature for state-parameter estimation with the presence of wetting and drying remains scarce, due to several challenges that need to be addressed in deriving an EnKF that accounts for the way the numerical model define wet and dry nodes. For instance, with the removal of dry nodes in the simulation, each ensemble member may be different in size. This raises some challenges in calculating ensemble mean and covariances required for the filter update. Furthermore, the observation operator used in the computation of the innovation needs to be modified for the presence of ‘no ob-
servations’ when drying occurs. Presumably, the development of the wetting/drying-consistent EnKF will be based on the EnKF with local analysis, where the EnKF analysis is performed nodally using only a subset of observations that are locally specified. This allows the modification of the observation operator and offers more flexibility in the scheme to estimate the model solutions.

- The curse of dimensionality is a typical limitation for PC expansion. In the chapter 2 and 4 of this thesis, computational overhead required to perform PC spectral projection exploded rapidly with respect to increasing stochastic dimensions, especially because the full tensorization method of 1D quadrature was used. Therefore, the future work must include the use of sparse quadrature methods, such as adaptive Smolyak Pseudo-Spectral projections [188] to enable the inclusion of several times more random variables in the PC construction than the present work. This will be particularly fulfilling for the work in the chapter 4 as more KL modes will be needed for the inference of the highly complex parameter field.

- In chapter 4, the difficulties in constructing a smooth mapping of the hyper-parameters \(q\) to the \(q\)-dependent coordinate transformation operator \(B(q)\) lead us to reconsider the problem. Evidently, the difficulty arises from the dependences of the eigenmodes on \(q\), and our capacity to follow a mode (a branch) when \(q\) varies. This observation suggests reconsidering our construction and work on quantities that directly present smooth and unambiguous dependences on \(q\). Clearly, such unambiguous quantity is the covariance, as \(q \mapsto C(x, x'|q)\) is well defined and certainly smooth for the parametrized families we may be interested (for instance the Mattern’s class). Reconsidering our approach and trying to design a new one, we envision that changing \(q\) should affect the prior of the proposed field, not the field values themselves. Therefore, the change of coordinates that adapts the realization of the process, when \(q\) changes, to maintain its prior probability constant is somehow not very natural.
When changing $\mathbf{q}$, it seems more natural to keep the same realization of the process and modify the prior probabilities of its coordinates in the reference basis. The mathematical derivation of this alternative approach for Bayesian inference with hyperparameters uncertainty has been formulated and the validation of the method using numerical experiments is currently a work in progress.
REFERENCES


burgers equation in the presence of uncertainty,” *Journal of Computational 

dimensional circulation model for shelves, coasts, and estuaries. report 1. theory
and methodology of adcirc-2ddi and adcirc-3dl.” COASTAL ENGINEERING RE-


tion and inference of mannings friction coefficients using dart buoy data during the 

ike forerunner surge,” *Geophysical Research Letters*, vol. 38, no. 8, 2011.

[17] Y. Ho and R. Lee, “A bayesian approach to problems in stochastic estimation and 
control,” *IEEE Transactions on Automatic Control*, vol. 9, no. 4, pp. 333–339, Oct 
1964.


[20] I. Sraj, M. Iskandarani, A. Srinivasan, W. C. Thacker, J. Winokur, A. Alexanderian, 

siam, 2005.

[22] I. Hoteit, “A reduced-order simulated annealing approach for four-dimensional vari-
atational data assimilation in meteorology and oceanography,” *International journal 


[106] G. J. Arcement and V. R. Schneider, “Guide for selecting manning’s roughness coefficients for natural channels and flood plains,” USGPO; For sale by the Books and


[117] I. Navon, “Practical and theoretical aspects of adjoint parameter estimation and


APPENDICES

A  Papers Published and Submitted


