Received: 5 May 2020 | Accepted: 19 February 2022
DOI: 10.1111/rssc.12558

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY | Series C Applied Statistics

Multifidelity computer model emulation with high-dimensional output: An application to storm surge

Pulong Ma[1] | **Georgios Karagiannis**[2] | **Bledar A. Konomi**[3] | **Taylor G. Asher**[4] | **Gabriel R. Toro**[5] | **Andrew T. Cox**[6]

[1]School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, USA
[2]Durham University, Durham, UK
[3]University of Cincinnati, Cincinnati, Ohio, USA
[4]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[5]Lettis Consultants International, Inc., Augusta, Georgia, USA
[6]Oceanweather, Inc., Stamford, Connecticut, USA

Correspondence
Pulong Ma, Assistant Professor in School of Mathematical and Statistical Sciences at Clemson University, 220 Parkway Drive, Clemson, SC 29634, USA.
Email: plma@clemson.edu

Abstract
Hurricane-driven storm surge is one of the most deadly and costly natural disasters, making precise quantification of the surge hazard of great importance. Surge hazard quantification is often performed through physics-based computer models of storm surges. Such computer models can be implemented with a wide range of fidelity levels, with computational burdens varying by several orders of magnitude due to the nature of the system. The threat posed by surge makes greater fidelity highly desirable, however, such models and their high-volume output tend to come at great computational cost, which can make detailed study of coastal flood hazards prohibitive. These needs make the development of an emulator combining high-dimensional output from multiple complex computer models with different fidelity levels important. We propose a parallel partial autoregressive cokriging model to predict highly accurate storm surges in a computationally efficient way over a large spatial domain. This emulator has the capability of predicting storm surges as

accurately as a high-fidelity computer model given any storm characteristics over a large spatial domain.

KEYWORDS

autoregressive cokriging, high-dimensional output, multifidelity computer model, storm surge, uncertainty quantification

1 | INTRODUCTION

Storm surge is one of the most severe natural disasters that can lead to significant flooding in coastal areas that brings multi-billion dollar damages and is responsible on average for half of lives lost from hurricanes (Rappaport, 2014). On average, inflation-normalized direct economic damages to the United States (1900–2005) are estimated at $10 billion per year and increasing as a result of storm surges (Pielke Jr et al., 2008). Since 2005, there have been 12 hurricanes whose total United States damages exceeded $10 billion (NOAA National Centers for Environmental Information, 2019). For instance, Hurricane Katrina (2005) caused over 1500 deaths and total estimated damages of $75 billion in the New Orleans area and along the Mississippi coast as a result of storm surge (FEMA, 2006). To mitigate these impacts, studies are carried out to evaluate the probabilistic hazard (e.g. Cialone et al., 2017; Niedoroda et al., 2010) and risk (e.g. Fischbach et al., 2016) from coastal flooding through a synthesis of computer modelling, statistical modelling and extreme-event probability computation. Here computer modelling is used to predict the storm surge hazard initialized by hurricanes, statistical modelling is used to determine the distribution of hurricane characteristics, and extreme-event probability is used to assess the flood hazard. These studies support development and application of flood insurance rates, building codes, land use planning/development, infrastructure design and construction and related goals by providing hazard levels at a range of frequencies (e.g. Aerts et al., 2014). Similarly, forecast simulations are used to support a wide array of operational needs, most notably disaster mitigation and evacuation planning/preparedness (e.g. Blanton et al., 2020; Georgas et al., 2016).

The ADCIRC ocean circulation model (Luettich & Westerink, 2004; Westerink et al., 2008) is the primary computer model in the United States to predict storm surges in coastal areas. It was certified by the Federal Emergency Management Agency (FEMA) for coastal flood hazard study and has been successfully used in a large number of applications, including FEMA flood hazard map updates (e.g. FEMA, 2008; Hesser et al., 2013; Jensen et al., 2012; Niedoroda et al., 2010) and in support of United States Army Corps of Engineers (USACE) projects (e.g. Cialone et al., 2017; Wamsley et al., 2013). These studies develop surge and wave hazard elevations corresponding to a range of frequencies, from the 50% annual exceedance level to the 0.01% annual exceedance level.

In the risk assessment of coastal flood hazard, ADCIRC needs to be run for a large number of storm characteristics. ADCIRC can be run at different levels of accuracy due to the sophistication of the physics incorporated in mathematical models, accuracy of numerical solvers and resolutions of meshes. Although ADCIRC can be run efficiently in parallel on large supercomputers (Tanaka et al., 2011), its computational cost scales with the cube of the spatial resolution, meaning that very high-fidelity models are several orders of magnitude more expensive than lower fidelity ones. By incorporating the physics of ocean waves, ADCIRC can generate storm surges with even greater fidelity, but this adds another order of magnitude to the run time as compared to the uncoupled ADCIRC model (Dietrich et al., 2012). For instance, a single high-resolution, coupled simulation in Southwestern Florida takes roughly 2000 core-hours on a high-performance supercomputer (Towns et al., 2014). As a result, research often uses coarser models without wave effects, even though the importance of more advanced, detailed models in estimating storm surge has been well demonstrated (e.g. Marsooli & Lin, 2018; Yin et al., 2016).

An important scientific demand is to develop an *emulator*—a fast probabilistic approximation of a simulator—that can produce highly accurate storm surges over a large spatial domain. The development of an emulator is needed in probabilistic flood studies dealing with climate change where a broad host of variables need to be considered. For instance, risk studies with relatively simple surge models estimate the current coastal flooding risk to New York City alone at over 100 million U.S. dollars per year (Aerts et al., 2014). But these estimates are highly sensitive to underlying assumptions that remain to be explored (e.g. Fischbach et al., 2016; de Moel & Aerts, 2011). Similarly, high-fidelity studies of coastal flooding changes associated with sea level rise and/or climate change have shown complex, nonlinear changes in flooding patterns that simpler studies cannot uncover (Liu et al., 2019).

The main scientific goal in this article is to develop an emulator that can not only predict highly accurate storm surges but also can be run very quickly over a large spatial domain. The development of an emulator for the high-fidelity storm surge model directly can be computationally prohibitive, since the high-fidelity storm surge model requires a tremendous amount of computing resources just to obtain a single run. An alternative is to develop an emulator that can combine a limited number of highly accurate simulations from a high fidelity but expensive surge model and a larger number of less accurate simulations from a low fidelity but cheaper surge model. Combining simulations from different fidelity levels relies on the idea that, after quantifying discrepancy between models of different fidelity levels, information from the low-fidelity surge model can facilitate prediction at high fidelity.

Several statistical works have been proposed to combine output from simulators at different fidelity levels based on a well-known geostatistical method called *cokriging* (see chapter 3 of Cressie, 1993). The idea to emulate multifidelity computer models is originated in Kennedy and O'Hagan (2000) using an autoregressive cokriging model—a cokriging model with an Markov assumption. The work in Kennedy and O'Hagan (2000) has been extended in several ways. For instance, Qian and Wu (2008) propose a Bayesian hierarchical formulation. Le Gratiet (2013) devises an efficient Bayesian approach to estimate model parameters. Konomi and Karagiannis (2021) introduce nonstationarity by partitioning the input space via a Bayesian tree structure. Ma (2020) develops objective Bayesian analysis for an autoregressive cokriging model. A typical assumption in univariate autoregressive cokriging models is the hierarchically nested design, which is not always preferred as discussed in Qian and Wu (2008). All these works focus on univariate computer model output without addressing the high-dimensionality challenge and possibly non-nested design in the storm surge application.

To address challenges due to high-dimensionality in the output space and non-nested design, we propose the parallel partial (PP) cokriging emulation methodology that can deal with high-dimensional output over a large spatial domain and non-nested design. In particular, the PP cokriging emulator is an extension of the PP kriging emulator (Gu & Berger, 2016) on different levels of fidelity code with massive output. To allow non-nested design, we develop a data-augmentation technique for parameter estimation via Monte Carlo expectation-maximization (MCEM) algorithm. For prediction, we develop a sequential prediction approach in which prediction at a higher fidelity level requires prediction to be made at a lower fidelity level. The proposed PP cokriging emulator explicitly introduces nonstationarity through spatially varying the mean parameters and variance parameters in Gaussian processes at each fidelity level and also allows fast computations in an empirical Bayesian framework.

The remainder of the article is organized as follows. Section 2 introduces the storm surge application with two storm surge simulators. In Section 3, we present the proposed methodology to handle high-dimensional output and non-nested design. In Section 4, an analysis of storm surge simulators is performed with the proposed methodology. Section 5 is concluded with discussions and possible extensions.

2 | APPLICATION OF INTEREST: STORM SURGE

This section describes the storm surge simulators with their discrepancy highlighted and the simulation design that is used for this study.

2.1 | Storm surge simulators

ADCIRC is a hydrodynamic circulation numerical model that solves the shallow-water equations for water levels and horizontal currents using the finite element method over an unstructured gridded domain representing bathymetric and topographic features (Luettich & Westerink, 2004). Information about ADCIRC can be accessed at https://adcirc.org. In what follows, we will refer to ADCIRC as the low-fidelity simulator, meanwhile we will refer to the coupled ADCIRC + SWAN model as the high-fidelity simulator. The latter incorporates the Simulating WAves Nearshore (SWAN) wave model (Booij et al., 1999; Zijlema, 2010) in order to enhance system physics and accuracy. This is achieved by tightly coupling the ADCIRC and SWAN models, simulating them on the same unstructured mesh (Dietrich et al., 2011, 2012). This coupling is important for accurate prediction of waves in the nearshore, which ride on top of storm surge and bring substantial destructive power to coastal structures and defences.

Figure 1 shows a basic diagram for the ADCIRC simulator and the ADCIRC + SWAN simulator. In this study, we focus on six input parameters to characterize the storm: ΔP, R_p, V_f, θ, B, ℓ, with their physical meaning given in Table 1. These parameters will be treated as inputs in the ADCIRC simulator. Although the behaviour of hurricanes is much more complex than this characterization, using this simplified storm parameterization is acceptable for the probabilistic characterization of future storms since no practical or robust model exists to represent these effects for surge-frequency calculations for future storms. In this application, the response variable of interest is the peak surge elevation (PSE) for each landfalling hurricane simulated from these surge models, where the peakness is taken across time over the course of one storm.

2.2 | Model validation

In this work, our goal is to develop an emulator for this coupled ADCIRC + SWAN model for storm surge prediction. The detailed validation of the coupled ADCIRC + SWAN model has been performed against a variety of data sources (tide harmonic constituent data, surge measurements, water level gages and wave buoy data) and storm events (Hurricane Charley 2004; Tropical storm Gabrielle 2001; Hurricane Donna 1960) in FEMA coastal flood hazard studies (e.g. FEMA, 2017).

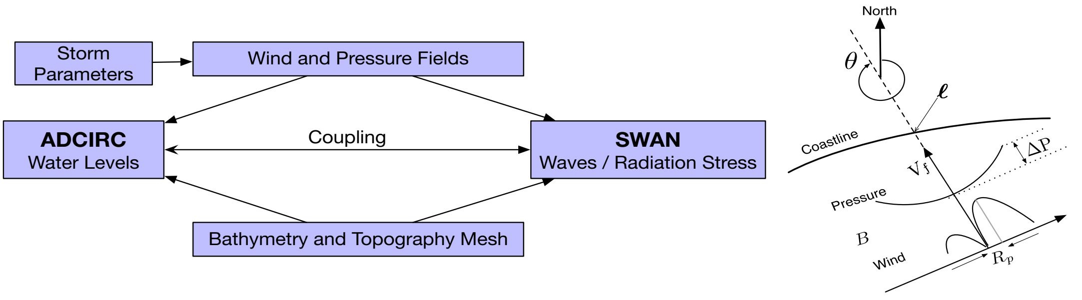


FIGURE 1 Diagram of storm surge models based on ADCIRC and storm parameters. The left panel shows ADCIRC and its coupling with SWAN. Then right panel shows storm parameters when a storm approaches the coastline (Toro et al., 2010). ADCIRC has bathymetry and topography mesh and wind and pressure fields as inputs. Storm parameters are used to derive the wind and pressure fields [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Storm characteristics parameters

Input variables	Physical meaning
ΔP	Central pressure deficit of the storm (mb)
R_p	Scale pressure radius in nautical miles
V_f	Storm's forward speed (m/s)
θ	Storm's heading in degrees clockwise from north
B	Holland's B parameter (unitless)
ℓ	Landfall location in latitude and longitude

The coupled ADCIRC + SWAN model has been validated to observed surges in several studies for historical storms, and shows good model performance with typical errors below 0.3 metres (Cialone et al., 2017; Dietrich et al., 2011, 2012; FEMA, 2017).

From a physics perspective, ADCIRC + SWAN explicitly incorporates ubiquitous wave effects on water levels and currents through the SWAN model. Before modern computing, wave setup effects were determined via approximate equations and added onto surge hazard estimates. Now, model coupling is standard practice for both researchers and practitioners to more correctly represent the physical processes, with clear benefits (Dietrich et al., 2011). ADCIRC + SWAN has been used in all regions of the U.S. Gulf and Atlantic coasts (including our study region) by both FEMA and USACE in their flood hazard studies. Flooding forecasting work still struggles to utilize coupled models because of the computational cost, and our work also has the potential to aid such efforts, partly thanks to the speed of emulator construction. There is a substantial need to develop an efficient emulator for the high-fidelity simulator ADCIRC + SWAN to aid flood hazard studies and forecasting work.

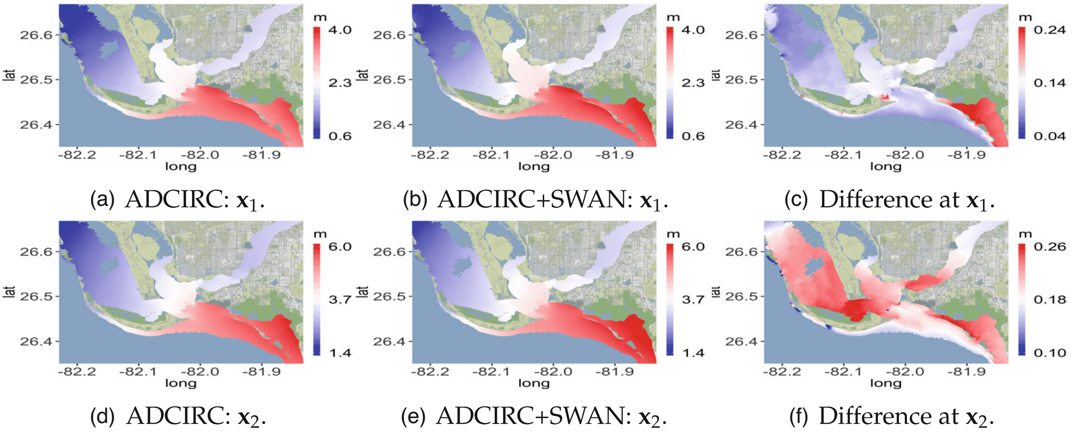
2.3 | Model simulation set-up

In FEMA coastal flood hazard studies (FEMA, 2017), storm surge hazard assessment is accomplished via the annual exceedance probability (AEP) for hurricane-prone areas. The quantification of AEP requires large-scale numerical simulations from ADCIRC. Statistical modelling is used to develop characteristics of synthetic storms based on historical tropical cyclones. The wind and pressure fields are used as inputs into hydrodynamic models such as ADCIRC to predict storm surges. In this application, the ADCIRC simulator and the ADCIRC + SWAN simulator are run on the same mesh with 148,055 nodes (spatial points). The mesh and simulation characteristics were constructed for a FEMA coastal flood hazard study in Southwest Florida (FEMA, 2017) and these simulations were carried out using the same standards and methods documented in that study. The peak surge hazard estimates produced by that study are considered to represent current conditions (i.e. sea level and climate in the years around when the study is done), and the joint probability distribution of tropical cyclone parameters is constructed from regional historical data.

We primarily focus on peak storm surges at $N = 9284$ spatial locations in the Cape Coral subregion of the study area, since Cape Coral is a study region in the FEMA Region IV's mission. Section S.1 of the Supplementary Material gives a brief description of coastal flood study and shows the full mesh of ADCIRC and the selected spatial locations. To design the experiment, we select 50 unique combinations of storm parameters $(\Delta P, R_p, V_f, \theta, B)$ based on the maximin Latin hypercube design (LHD) in the input domain $[30, 70] \times [16, 39] \times [3, 10] \times [15, 75] \times [0.9, 1.4]$. This parameter range corresponds to a core region of major surge hazards of interest in the FEMA coastal study (FEMA, 2017). For each combination of these five storm parameters, the landfall location ℓ is repeated with one R_p spacing along the coastline in Cape Coral. For each of the five storm parameters, the initial position of landfall location is randomly chosen, meaning that no two storms make landfall at the same location, and the number of landfalls for each of the 50 parameter combinations varies, with a higher number of smaller storms; this reflects the smaller spatial scale of storm surge for smaller storms, which necessitates higher sampling to capture the more localized response. The meteorological forcing in both the ADCIRC simulator and the ADCIRC + SWAN simulator is produced by a single group (Oceanweather, Inc.) using the work in Cardone and Cox (2009). The full simulation can be found from the Coastal Flood Modeling Database - Southwest Florida (Asher & Liu, 2022).

In total, we obtained 226 inputs, meaning that on average, each of the 50 parameter combinations is used to generate 4.5 storms. These 226 inputs will be referred to as $\mathcal{X}_0$. We randomly selected 60 inputs from the 226 inputs to run the ADCIRC + SWAN simulator, which will be referred to as $\mathcal{X}_2$. Here 60 model runs are selected because the folklore for Gaussian process emulation is to use $10d$ model runs. Then we randomly selected 150 inputs from the remaining 166 inputs to run the ADCIRC simulator, which will be referred to as $\mathcal{X}_1^1$. To characterize the difference between the ADCIRC simulator and the ADCIRC + SWAN simulator, we also randomly chose 50 inputs from the 60 inputs to run the ADCIRC simulator, which will be referred to as $\mathcal{X}_1^2$. Let $\mathcal{X}_1 := \mathcal{X}_1^1 \cup \mathcal{X}_1^2$ be the collection of 200 inputs from $\mathcal{X}_1^1$ and $\mathcal{X}_1^2$. Notice that only 50 inputs in the ADCIRC + SWAN simulator are nested within the 200 inputs in the ADCIRC simulator.

Figure 2 shows peak surge elevations over 9284 spatial locations from the ADCIRC simulator and the ADCIRC + SWAN simulator at two different input settings $\mathbf{x}_1 = (48.30, 20.48, 6.187, 62.28, 1.260, -82.08, 26.59)^\top$ and $\mathbf{x}_2 = (68.85, 34.34, 8.778, 55.57, 1.066, -82.15, 26.69)^\top$. The output surfaces also have very different variations in different regions at each fidelity level. This indicates that a spatially varying mean function or a spatially varying variance function may help capture the spatial variations in the output space. The third column shows that PSEs have their maximum difference less than 0.3 metres between the low- and high-fidelity simulators, and also shows that the discrepancy has different spatial structures. At some specific regions such as the Fort Myers Beach (on the top right



(a) ADCIRC: $\mathbf{x}_1$. (b) ADCIRC+SWAN: $\mathbf{x}_1$. (c) Difference at $\mathbf{x}_1$.

(d) ADCIRC: $\mathbf{x}_2$. (e) ADCIRC+SWAN: $\mathbf{x}_2$. (f) Difference at $\mathbf{x}_2$.

FIGURE 2 Comparison of model runs from the ADCIRC simulator and the ADCIRC + SWAN simulator at two input settings. The first and second columns show the model runs from the low-fidelity simulator and the high-fidelity simulator. The third column shows the difference between the high-fidelity run and the low-fidelity run [Colour figure can be viewed at wileyonlinelibrary.com]

panel), very sharp changes can be detected. Physically, the changes in the surge elevation arise from differences in the spatial and temporal structures of the surge-only versus the wave response to storm forcing. For instance, in the top panel, the addition of wave-driven water level set-up leads to greater overtopping (a situation when waves are higher than the dunes or structures they encounter) of coastal barrier islands, bringing greater water into the semi-protected bays in the southeastern portion of the figure. It can be difficult to detect these sorts of patterns without modelling wind-driven wave effects. In what follows, we develop a cokriging-based emulator to approximate the high-fidelity simulator by combining simulations from a limited number of high-fidelity runs and a larger number of low-fidelity simulation runs.

3 | MULTIFIDELITY COMPUTER MODEL EMULATION

Section 3.1 gives a brief introduction of the general autoregressive cokriging framework, and Section 3.2 presents the proposed methodology called *parallel partial cokriging*.

3.1 | Background on autoregressive cokriging modeling

Assume that the computer model can be run at s levels of sophistication corresponding to output functions $y_1(\cdot), \dots, y_s(\cdot)$. The computer model associated with $y_t(\cdot)$ is assumed to be more accurate than the one associated with $y_{t-1}(\cdot)$ for $t = 2, \dots, s$. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, which is assumed to be the input space of computer model. Further assume that the computer model $y_t(\cdot)$ is run at a set of input values denoted by $\mathcal{X}_t \subset \mathcal{X}$ for $t = 1, \dots, s$, where $\mathcal{X}_t$ contains n_t input values. The autoregressive cokriging model at any input $\mathbf{x} \in \mathcal{X}_t$ is

$$y_t(\mathbf{x}) = \gamma_{t-1}(\mathbf{x}) y_{t-1}(\mathbf{x}) + \delta_t(\mathbf{x}), \quad t = 2, \dots, s, \tag{1}$$

where $\delta_t(\cdot)$ is the unknown location discrepancy representing the local adjustment from level $t-1$ to level t, and $\gamma_{t-1}(\mathbf{x})$ is the scale discrepancy representing the scale change from level $t-1$ to level t at input $\mathbf{x}$. This well-interpreted model is induced from the so called Markov assumption: no further information is gained about $y_t(\mathbf{x})$ by observing $y_{t-1}(\mathbf{x}')$ for any other $\mathbf{x} \neq \mathbf{x}'$ (Kennedy & O'Hagan, 2000).

To account for uncertainties in the unknown functions $y_1(\cdot)$ and $\delta_t(\cdot)$, we assign Gaussian processes

$$\begin{aligned} y_1(\cdot)|\boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\phi}_1 &\sim \mathcal{GP}\left(\mathbf{h}_1^\top(\cdot)\boldsymbol{\beta}_1, \sigma_1^2 r(\cdot,\cdot|\boldsymbol{\phi}_1)\right), \\ \delta_t(\cdot) &\sim \mathcal{GP}\left(\mathbf{h}_t^\top(\cdot)\boldsymbol{\beta}_t, \sigma_t^2 r(\cdot,\cdot|\boldsymbol{\phi}_t)\right), \end{aligned} \tag{2}$$

for $t = 2, \dots, s$. $\mathbf{h}_t(\cdot)$ is a vector of basis functions and $\boldsymbol{\beta}_t$ is a vector of coefficients at fidelity level t. In practice, the basis functions $\mathbf{h}_t(\cdot)$ along with $\mathbf{w}_t(\cdot)$ should be determined by exploratory data analysis following standard Gaussian process modelling procedure. $\gamma_{t-1}(\mathbf{x})$ can be modelled as a basis-function representation, that is, $\gamma_{t-1}(\mathbf{x}) = \mathbf{w}_{t-1}(\mathbf{x})^\top \boldsymbol{\xi}_{t-1}$, where $\mathbf{w}_{t-1}(\mathbf{x})$ is a vector of known basis functions and $\boldsymbol{\xi}_{t-1}$ is a vector of unknown coefficients. Here, $r(\cdot,\cdot \mid \boldsymbol{\phi}_t)$ is a correlation function with correlation parameters $\boldsymbol{\phi}_t := (\phi_{t,1}, \dots, \phi_{t,d})^\top$ at fidelity level t, where d denotes the number of input parameters. Following the seminar work (Sacks et al., 1989), we use the product form of correlation structure to allow anisotropy in each input dimension, that is, $r(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\phi}_t) = \prod_{i=1}^d r(x_i, x_i' | \phi_{t,i})$, where each $r(x_i, x_i'|\phi_{t,i})$ can be chosen as the Matérn correlation function although other choice is available (see Ma & Bhadra, 2022). The Matérn correlation function is

$$r(u|\phi) = \frac{2^{1-\upsilon}}{\Gamma(\upsilon)}\left(\frac{\sqrt{2\upsilon}u}{\phi}\right)^{\upsilon} \mathcal{K}_\upsilon\left(\frac{\sqrt{2\upsilon}u}{\phi}\right),$$

where $u = |x_i - x_i'|$ is the Euclidean distance, $\mathcal{K}_\upsilon$ is the modified Bessel function of the second kind and υ is the smoothness parameter that controls the mean-square differentiability of random processes. Following the standard practice of computer model emulation, $\upsilon = 2.5$ is chosen because of its closed-form expression and twice mean-square differentiability of random processes (e.g. Ma, 2020).

One way to implement the autoregressive cokriging model defined in Equations (1) and (2) for the storm surge application is to treat spatial coordinates as an additional input parameter. This can avoid dealing with the high-dimensional output, but it requires dealing with large-scale computations in evaluating the likelihood function. In particular, we have 9284×200 model output values for the low-fidelity simulator ADCIRC and 9284×60 ones for the high-fidelity simulator ADCIRC + SWAN. Fitting the autoregressive cokriging model by treating spatial coordinates as an additional input would require about $1.4 \times 10^{19} = (9284 \times 260)^3$ flops to evaluate the likelihood and about 46,252 gigabytes memory to store the covariance matrix with double precision. In addition, the computation of the predictive distribution would be infeasible in standard computers as large number of unknown parameters could not be analytically integrated out; this is because the experimental design in our application is not necessarily fully nested across levels of code as required in existing autoregressive cokriging implementations (Kennedy & O'Hagan, 2000; Le Gratiet, 2013; Qian & Wu, 2008). Another challenge is to model nonstationarity in the output space, where the storm surges in Cape Coral show strong heterogeneous spatial dependence structures.

3.2 | The parallel partial cokriging emulator

We propose the parallel partial cokriging emulator that couples ideas from the parallel partial Gaussian process (Gu & Berger, 2016) with the cokriging model. This approach mitigates the aforementioned challenges in the storm surge application.

We consider that at each level the output functions and their additive discrepancy functions are modelled as multivariate GPs where each dimension corresponds to a spatial location. This induces a cokriging model with high-dimensional output due to the massive spatial locations available. Two popular ideas for modelling multivariate output in computer models are the nonseparable covariance model between the input space and output space via basis-function representations for outputs (Higdon et al., 2008) and the separable covariance model between input space and output space (Conti & O'Hagan, 2010).

However, these approaches have not been developed in the multifidelity setting. What follows is to discuss whether they can provide useful ideas to solve the storm surge application. Specifically, the basis-function representation approach (Higdon et al., 2008) requires the output to be represented in terms of a few principal components. However, exploratory analysis suggests that this is not possible in the storm surge application, due to the large changes in the storm output from different input simulations. For the separable model (Conti & O'Hagan, 2010), it is computationally infeasible due to massive output that the storm surge simulators generate for each simulation. To deal with the high-dimensional output in the separable model, Gu and Berger (2016) propose the PP Gaussian process emulator, which assumes conditionally independent Gaussian processes for each spatial coordinate with the same range parameter. This approach is able to borrow information across the data-poor input dimensions when the data are rich in the spatial dimension with linear computational cost in terms of the number of simulator outputs N. However, the adoption of the PP ideas in the multifidelity setting with possibly non-nested designs is not straightforward. What follows is to introduce the proposed emulator methodology.

3.2.1 | The parallel partial cokriging model

In the storm surge application, we have $s = 2$ fidelity levels for computer models: ADCIRC and ADCIRC + SWAN. Each simulation of these computer models generates output values at the same $N = 9284$ spatial locations. Let n_t denote the number of computer simulations at fidelity level t, where $n_1 = 200$ and $n_2 = 60$ in the application. Let $\mathbf{y}_{t,j}$ be a vector of output values over all inputs in $\mathcal{X}_t$ at coordinate j and fidelity level t. Let $\mathbf{y}^{t,\mathscr{D}} := [\mathbf{y}_{t,1}, \ldots, \mathbf{y}_{t,N}]$ be a $n_t \times N$ matrix of output values across all input values and all spatial locations at level t. Let $\mathbf{y}_j^{\mathscr{D}} := (\mathbf{y}_{1,j}^\top, \ldots, \mathbf{y}_{s,j}^\top)^\top$ be a vector of output values at coordinate j over all inputs in $\mathcal{X}$ and all fidelity levels.

For each coordinate $j = 1, \ldots, N$ and fidelity level $t = 2, \ldots, s$, we specify the cokriging model for any input $\mathbf{x} \in \mathcal{X}_t$ as

$$y_{t,j}(\mathbf{x}) = \gamma_{t-1,j}(\mathbf{x}) y_{t-1,j}(\mathbf{x}) + \delta_{t,j}(\mathbf{x}). \tag{3}$$

Here, for $y_{t-1,j}(\mathbf{x})$ and $\delta_{t,j}(\mathbf{x})$, we assign Gaussian processes priors:

$$\begin{aligned} y_{1,j}(\cdot) | \boldsymbol{\beta}_{1,j}, \sigma_{1,j}^2, \boldsymbol{\phi}_1 &\sim \mathcal{GP}\left(\mathbf{h}_1^\top(\cdot)\boldsymbol{\beta}_{1,j}, \sigma_{1,j}^2 r(\cdot,\cdot|\boldsymbol{\phi}_1)\right), \\ \delta_{t,j}(\cdot) &\sim \mathcal{GP}\left(\mathbf{h}_t^\top(\cdot)\boldsymbol{\beta}_{t,j}, \sigma_{t,j}^2 r(\cdot,\cdot|\boldsymbol{\phi}_t)\right), \end{aligned} \tag{4}$$

where $\mathbf{h}_t(\cdot)$ is a vector of common fixed basis functions across all N spatial locations. In the storm surge application, these basis functions are assumed to be constant functions based on exploratory analysis. For $\gamma_{t-1,j}(\mathbf{x})$, it is represented as a function of $\mathbf{x}$, but it will be assumed as unknown constants in the real application, that is, $\gamma_{t-1,j}(\mathbf{x}) := \gamma_{t-1,j}$, since the physics knowledge suggests that the coupled computer model ADCIRC + SWAN tends to generate higher storm surges than ADCIRC due to the wave effects. For each coordinate j, we assume different regression parameters $\boldsymbol{\beta}_j := \{\boldsymbol{\beta}_{1,j}, \ldots, \boldsymbol{\beta}_{s,j}\}$, different variance parameters $\boldsymbol{\sigma}_j^2 := \{\sigma_{1,j}^2, \cdots, \sigma_{s,j}^2\}$, and different scale discrepancy parameters $\boldsymbol{\gamma}_j := \{\gamma_{1,j}, \ldots, \gamma_{t-1,j}\}$. The correlation parameters $\boldsymbol{\phi}_t$ are assumed to be the same across different spatial coordinates to simplify computations following (Gu & Berger, 2016).

We call the proposed model defined in Equations (3) and (4) the *parallel partial cokriging* model. The 'parallel partial' reflects the fact that our model can be thought of as involving different autoregressive cokriging models at each coordinate which share common input correlation parameters. As mentioned in Section 3.1, likelihood-based inference requires that the collection of input runs at each level is nested in order to have closed-form inference, that is, $\mathcal{X}_t \subset \mathcal{X}_{t-1}$. In the next section, we present a data-augmentation technique to deal with possibly non-nested design so that statistical inference based on the PP cokriging model can be carried out.

3.2.2 | Data augmentation for non-nested design

The specification of convenient priors facilitating the tractability of the marginal likelihood and the analytic integration of the unknown parameters require the available experimental design to be fully hierarchical nested; that is, $\mathcal{X}_{t+1} \subset \mathcal{X}_t$. This restrictive requirement can be found by examining the likelihood that results from Equations (3) and (4) and it is inherited from the cokriging model. As this requirement is not satisfied in our application, to address this issue we propose a suitable data-augmentation remedy that imputes the data with missing values to create a fully nested design.

We replace the original input design $\mathcal{X}_t$ by $\tilde{\mathcal{X}}_t = \mathcal{X}_t \cup \mathring{\mathcal{X}}_t$ such that $\mathring{\mathcal{X}}_t := \mathcal{X}_{(t+1):s} \backslash \mathcal{X}_t$ represents a collection of missing inputs that have not been run by the simulator at fidelity level t in order to form a nested design, where $\mathcal{X}_{(t+1):s} := \cup_{k=t+1}^{s} \mathcal{X}_k$ represents the collection of observed inputs from fidelity level $t+1$ up to the highest fidelity level s. Let $\mathring{\mathbf{y}}_{t,j}$ be a vector of missing output values all inputs in $\mathring{\mathcal{X}}_t$ at coordinate j and fidelity level t. In what follows, we use $\tilde{n}_t$ to denote the number of input values in the augmented set $\tilde{\mathcal{X}}_t$. Let $\mathring{\mathbf{y}}_j^{\mathscr{D}} := (\mathring{\mathbf{y}}_{1,j}^\top, \ldots, \mathring{\mathbf{y}}_{s,j}^\top)^\top$ be a vector of missing output values at coordinate j over all fidelity levels, where $\mathring{\mathbf{y}}_{s,j}^\top$ is defined to be empty for notational convenience. Let $(\mathring{\mathbf{y}}_t)_{j=1}^N := (\mathring{\mathbf{y}}_{t,1}^\top, \ldots, \mathring{\mathbf{y}}_{t,N}^\top)^\top$ be a vector of missing output at fidelity level t over all N spatial coordinates. Let $\tilde{\mathbf{y}}_{t,j} := (\mathbf{y}_{t,j}^\top, \mathring{\mathbf{y}}_{t,j}^\top)^\top$ be a vector of augmented output over all inputs at fidelity level t at the jth spatial coordinate and $\tilde{\mathbf{y}}_j^{\mathscr{D}} := (\mathbf{y}_j^{\mathscr{D}\top}, \mathring{\mathbf{y}}_j^{\mathscr{D}\top})^\top$ be a vector of augmented output over all inputs at coordinate j. Then the augmented sampling distribution at coordinate j is

$$L\left(\tilde{\mathbf{y}}_j^{\mathscr{D}} | \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \boldsymbol{\sigma}_j^2, \boldsymbol{\phi}\right) \propto \pi\left(\tilde{\mathbf{y}}_{1,j} | \boldsymbol{\beta}_{1,j}, \sigma_{1,j}^2, \boldsymbol{\phi}_1\right) \prod_{t=2}^{s} \pi\left(\tilde{\mathbf{y}}_{t,j} | \tilde{\mathbf{y}}_{t-1,j}, \boldsymbol{\beta}_{t,j}, \gamma_{t,j}, \sigma_{t,j}^2, \boldsymbol{\phi}_t\right), \tag{5}$$

with

$$\pi\left(\tilde{\mathbf{y}}_{1,j} | \boldsymbol{\beta}_{1,j}, \sigma_{1,j}^2, \boldsymbol{\phi}_1\right) = \mathcal{N}\left(\tilde{\mathbf{H}}_1 \boldsymbol{\beta}_{1,j}, \sigma_{t,j}^2 \tilde{\mathbf{R}}_1\right),$$
$$\pi\left(\tilde{\mathbf{y}}_{t,j} | \tilde{\mathbf{y}}_{t-1,j}, \boldsymbol{\beta}_{t,j}, \gamma_{t,j}, \sigma_{t,j}^2, \boldsymbol{\phi}_t\right) = \mathcal{N}\left(\tilde{\mathbf{H}}_t \boldsymbol{\beta}_{t,j} + \tilde{W}_{t-1,j}\gamma_{t-1,j}, \sigma_{t,j}^2 \tilde{\mathbf{R}}_t\right),$$

where $\tilde{\mathbf{H}}_t := \mathbf{h}_t(\tilde{\mathcal{X}}_t)$, $\tilde{\mathbf{R}}_t := r(\tilde{\mathcal{X}}_t, \tilde{\mathcal{X}}_t \mid \boldsymbol{\phi}_t)$, and $\tilde{W}_{t-1,j} := \mathbf{y}_{t-1,j}(\tilde{\mathcal{X}}_t)$. The proposed augmentation allows the joint likelihood of the complete output values to be factorized into Gaussian likelihood kernels of smaller dimensionality, where we can specify conditional conjugate priors and break the training problem into smaller more tractable ones.

In Sections S.2 and S.3 of the Supplementary Material, artificial examples are used to illustrate the performance of autoregressive cokriging with a *non-nested design* with the proposed data-augmentation technique. These examples show that Incorporating information from a low-fidelity code can improve inferential results on the high-fidelity code as well as demonstrate the good performance of the PP cokriging emulator.

3.2.3 | Empirical Bayesian inference via an MCEM algorithm

Let $\tilde{\mathbf{y}}^{\mathscr{D}} := (\tilde{\mathbf{y}}_1^{\mathscr{D}}, \dots, \tilde{\mathbf{y}}_N^{\mathscr{D}})$ be a $(\sum_{t=1}^{s} n_t) \times N$ matrix of augmented outputs over all fidelity levels and all spatial locations. We introduce the following notation: $\boldsymbol{\beta} := (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_N^\top)^\top$, $\boldsymbol{\gamma} := (\boldsymbol{\gamma}_1^\top, \dots, \boldsymbol{\gamma}_N^\top)^\top$, and $\boldsymbol{\sigma}^2 := (\boldsymbol{\sigma}_1^{2\top}, \dots, \boldsymbol{\sigma}_N^{2\top})^\top$. The overall augmented sampling distribution across all N spatial locations is the product of each augmented sampling distribution

$$L(\tilde{\mathbf{y}}^{\mathscr{D}} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}) = \prod_{j=1}^{N} L(\tilde{\mathbf{y}}_j^{\mathscr{D}} \mid \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \boldsymbol{\sigma}_j^2, \boldsymbol{\phi}). \tag{6}$$

We specify the following a priori model for the unknown parameters

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi}) \prod_{j=1}^{N} \left\{ \pi(\boldsymbol{\beta}_{s,j}, \sigma_{s,j}^2) \prod_{t=1}^{s-1} \pi(\boldsymbol{\beta}_{t,j}, \gamma_{t,j}, \sigma_{t,j}^2) \right\}, \tag{7}$$

where independent Jeffreys priors can be assigned on $\boldsymbol{\beta}_{t,j}, \gamma_{t-1,j}, \sigma_{t,j}^2$: that is, $\pi(\boldsymbol{\beta}_{t,j}, \gamma_{t-1,j}, \sigma_{t,j}^2) \propto \frac{1}{\sigma_{t,j}^2}$ for $t = 2, \dots, s$, and $\pi(\boldsymbol{\beta}_{1,j}, \sigma_{1,j}^2) \propto \frac{1}{\sigma_{1,j}^2}$ at each coordinate j. For each level t, we assign a jointly robust prior (Gu, 2019) on $\boldsymbol{\phi}_t$ which is a proper prior and has desirable properties for Gaussian process emulation.

After integrating out model parameters $\{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2\}$, the conditional distribution of $\tilde{\mathbf{y}}^{\mathscr{D}}$ given $\boldsymbol{\phi}$ is

$$\pi(\tilde{\mathbf{y}}^{\mathscr{D}} \mid \boldsymbol{\phi}) = \prod_{j=1}^{N} \pi(\tilde{\mathbf{y}}_{1,j} \mid \boldsymbol{\phi}_1) \prod_{t=2}^{s} \pi(\tilde{\mathbf{y}}_{t,j} \mid \boldsymbol{\phi}_t, \tilde{\mathbf{y}}_{t-1,j}), \tag{8}$$

where each conditional distribution is given in Section S.4 of the Supplementary Material. The augmented posterior distribution $\pi(\boldsymbol{\phi} \mid \tilde{\mathbf{y}}^{\mathscr{D}})$ can be obtained via Bayes' theorem: $\pi(\boldsymbol{\phi} \mid \tilde{\mathbf{y}}^{\mathscr{D}}) \propto \pi(\tilde{\mathbf{y}}^{\mathscr{D}} \mid \boldsymbol{\phi}) \times \pi(\boldsymbol{\phi})$, where $\pi(\boldsymbol{\phi})$ is a proper prior.

To estimate $\boldsymbol{\phi}$, we adopt an empirical Bayesian approach which maximizes the integrated posterior $\pi(\boldsymbol{\phi} \mid \tilde{\mathbf{y}}^{\mathscr{D}})$, since empirical Bayes approaches provide faster computational results compared to the fully Bayesian ones which often require Markov chain Monte Carlo methods. As direct maximization of $\pi(\boldsymbol{\phi} \mid \tilde{\mathbf{y}}^{\mathscr{D}})$ is impossible due to the intractable form of $\pi(\boldsymbol{\phi} \mid \tilde{\mathbf{y}}^{\mathscr{D}})$, we introduce a Monte Carlo expectation-maximization (MCEM) algorithm (Dempster et al., 1977; Wei & Tanner, 1990) to tackle this challenge. The MCEM algorithm consists of two important steps: in E-step, the Q-function is defined as a function of certain conditional expectation that is approximated by Monte Carlo methods; in M-step, maximization of this function is performed with respect to the unknown parameters; for detailed development of this MCEM algorithm; see Algorithm 1 of the Supplementary Material.

3.2.4 | Prediction

For any new input $\mathbf{x}_0 \in \mathcal{X}$, the goal is to make prediction for $\{y_{s,j}(\mathbf{x}_0), j = 1, \ldots, N\}$ based on the data $\mathbf{y}^{\mathscr{D}}$. With the prior model (7), the predictive distribution of interest is $y_{s,j}(\mathbf{x}_0|\mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi})$ for $j = 1, \ldots, N$.

In what follows, we derive a new approach to predicting $y_{s,j}(\mathbf{x}_0)$ and termed it as a *sequential prediction* approach. The idea is to add the new input $\mathbf{x}_0$ to each collection of missing inputs $\mathring{\mathcal{X}}_t$ such that a hierarchically nested design can be obtained. To fix the notation, we define $\mathring{\mathcal{X}}_t^0 := \mathring{\mathcal{X}}_t \cup \{\mathbf{x}_0\}$ and $\tilde{\mathcal{X}}_t^0 := \mathcal{X}_t \cup \mathring{\mathcal{X}}_t^0$. Hence the collection of inputs $\{\tilde{\mathcal{X}}_t^0 : t = 1, \ldots, s\}$ also forms a nested design with $\tilde{\mathcal{X}}_t^0 \subset \tilde{\mathcal{X}}_{t-1}^0$. Let $\mathbf{y}(\mathbf{x}_0) := (\mathbf{y}_1(\mathbf{x}_0)^\top, \ldots, \mathbf{y}_s(\mathbf{x}_0)^\top)^\top$ with $\mathbf{y}_j(\mathbf{x}_0) := (y_{1,j}(\mathbf{x}_0), \ldots, y_{s,j}(\mathbf{x}_0))^\top$. The predictive distribution of $\mathbf{y}(\mathbf{x}_0)$ given $\tilde{\mathbf{y}}^{\mathscr{D}}$ and $\boldsymbol{\phi}$ is the product of N independent distributions with

$$\pi\left(\mathbf{y}(\mathbf{x}_0)|\tilde{\mathbf{y}}^{\mathscr{D}}, \boldsymbol{\phi}\right) = \prod_{j=1}^{N} \pi\left(\mathbf{y}_j(\mathbf{x}_0)|\tilde{\mathbf{y}}_j^{\mathscr{D}}, \boldsymbol{\phi}\right).$$

The following result gives the predictive distribution at each spatial coordinate. Its proof follows from standard kriging theory.

Proposition 1 (Sequential Prediction). *Given the PP cokriging model and the non-informative priors in Equation (7), the predictive distribution across all fidelity levels at spatial coordinate j for $j = 1, \ldots, N$ is*

$$\pi\left(\mathbf{y}_j(\mathbf{x}_0)|\tilde{\mathbf{y}}_j^{\mathscr{D}}, \boldsymbol{\phi}\right) = \pi(y_{1,j}(\mathbf{x}_0)|\tilde{\mathbf{y}}_{1,j}, \boldsymbol{\phi}_1)\prod_{t=2}^{s-1}\pi(y_{t,j}(\mathbf{x}_0)|\tilde{\mathbf{y}}_{t,j}, \tilde{\mathbf{y}}_{t-1,j}, y_{t-1,j}(\mathbf{x}_0), \boldsymbol{\phi}_t) \times \pi(y_{s,j}(\mathbf{x}_0)|y_{s-1,j}(\mathbf{x}_0), \mathbf{y}_{s,j}, \boldsymbol{\phi}_s). \tag{9}$$

The conditional distributions on the right-hand side of Equation (9) are Student-t distributions with degrees of freedom, location, and scale parameters given by

$$\begin{aligned}
\nu_{t,j} &:= n_t - q_t,\\
\boldsymbol{\mu}_{t,j} &:= \mathbf{T}(\mathbf{x}_0)\hat{\mathbf{b}}_{t,j} + \mathbf{r}_t^\top(\mathbf{x}_0)\tilde{\mathbf{R}}_t^{-1}(\tilde{\mathbf{y}}_{t,j} - \tilde{\mathbf{T}}_{t,j}\hat{\mathbf{b}}_{t,j}),\\
V_{t,j} &:= \frac{S^2(\boldsymbol{\phi}_t, \tilde{\mathbf{y}}_{t,j})}{n_t - q_t} c_{t,j}(\mathbf{x}_0),
\end{aligned}$$

with $\hat{\mathbf{b}}_{t,j} := \left(\tilde{\mathbf{T}}_{t,j}^\top\tilde{\mathbf{R}}_t^{-1}\tilde{\mathbf{T}}_{t,j}\right)^{-1}\tilde{\mathbf{T}}_{t,j}^\top\tilde{\mathbf{R}}_t^{-1}\tilde{\mathbf{y}}_{t,j}$, $S^2(\boldsymbol{\phi}_t, \tilde{\mathbf{y}}_{t,j}) := (\tilde{\mathbf{y}}_{t,j} - \tilde{\mathbf{T}}_{t,j}\hat{\mathbf{b}}_{t,j})^\top\tilde{\mathbf{R}}_t^{-1}(\tilde{\mathbf{y}}_{t,j} - \tilde{\mathbf{T}}_{t,j}\hat{\mathbf{b}}_{t,j})$,

$$\begin{aligned}
c_{t,j}(\mathbf{x}_0) &:= r(\mathbf{x}_0, \mathbf{x}_0|\boldsymbol{\phi}_t) - \mathbf{r}_t^\top(\mathbf{x}_0)\tilde{\mathbf{R}}_t^{-1}\mathbf{r}_t(\mathbf{x}_0)\\
&\quad + \left[\mathbf{T}_{t,j}(\mathbf{x}_0) - \tilde{\mathbf{T}}_{t,j}^\top\tilde{\mathbf{R}}_t^{-1}\mathbf{r}_t(\mathbf{x}_0)\right]^\top\left(\tilde{\mathbf{T}}_{t,j}^\top\tilde{\mathbf{R}}_t^{-1}\tilde{\mathbf{T}}_{t,j}\right)^{-1}\left[\mathbf{T}_{t,j}(\mathbf{x}_0) - \tilde{\mathbf{T}}_{t,j}^\top\tilde{\mathbf{R}}_t^{-1}\mathbf{r}_t(\mathbf{x}_0)\right],
\end{aligned}$$

where $\mathbf{r}_t(\mathbf{x}_0) := r(\tilde{\mathcal{X}}_t, \mathbf{x}_0|\boldsymbol{\phi}_t)$ *and* $\mathbf{T}_{t,j}(\mathbf{x}_0) = [\mathbf{h}_t(\mathbf{x}_0), y_{t-1,j}(\mathbf{x}_0)]$.

Proposition 1 shows that a random sample from the predictive distribution can be sequentially drawn from a collection of conditional distributions in an efficient manner, since the total computational cost required for such a simulation is $O(\sum_{t=1}^{s}\tilde{n}_t^3)$ at each spatial coordinate. As the correlation matrix is the same across all spatial locations at each fidelity level, the total computational cost to obtain one single random sample from the predictive distribution across all spatial locations is $O(\sum_{t=1}^{s}\tilde{n}_t^3 + N\sum_{t=1}^{s}\tilde{n}_t^2)$, which is linear in N when $\sum_{t=1}^{s}\tilde{n}_t^2 \ll N$. Notice that a sample from $\pi(\mathbf{y}_s(\mathbf{x}_0) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi})$ can be obtained via the composition sample technique based on $\pi(\mathbf{y}_s(\mathbf{x}_0) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi}) = \int \pi(\mathbf{y}_s(\mathbf{x}_0) \mid \tilde{\mathbf{y}}^{\mathscr{D}}, \boldsymbol{\phi})\pi(\mathring{\mathbf{y}}^{\mathscr{D}} \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi})d\mathring{\mathbf{y}}^{\mathscr{D}}$, with the missing data $\mathring{\mathbf{y}}^{\mathscr{D}}$ being generated from the distribution $\pi(\mathring{\mathbf{y}}^{\mathscr{D}} \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi})$ given in Section S.6 of the Supplementary Material. In practice, $\boldsymbol{\phi}$ needs to be replaced with its maximum a posteriori estimate obtained via the MCEM in Algorithm 1 of the Supplementary Material. Vanilla Monte Carlo approximation is used to compute the predictive mean and predictive variance of the predictive distribution $\pi(\mathbf{y}_s(\mathbf{x}_0) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi})$ when the design is not nested. For a nested design, closed-form expressions for the posterior predictive mean and posterior variance across all fidelity levels are given in Lemma 1 of the Supplementary Material.

In Section S.5 of the Supplementary Material, we derive the *one-step prediction* formula based on the idea in Kennedy and O'Hagan (2000) and Le Gratiet (2013) when the design is nested. The sequential prediction formula in Proposition 1 has several advantages over the one-step prediction formula in Section S.5 of the Supplementary Material. First, the high-dimentionality of simulator output makes the one-step prediction formula computationally infeasible in the storm surge application, since this sequential prediction formula has $O(N(\sum_{t=1}^{s}\tilde{n}_t^3))$ computational cost. Second, to obtain predictive distribution $\pi(y_{s,j}(\mathbf{x}_0) \mid \mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi})$, model parameters $\{\boldsymbol{\gamma}, \boldsymbol{\sigma}^2\}$ have to be numerically integrated out in the one-step prediction formula. Thus, Monte Carlo approximation is required to take account of uncertainty in both model parameters $\{\boldsymbol{\gamma}, \boldsymbol{\sigma}^2\}$ and missing data $\mathring{\mathbf{y}}^{\mathscr{D}}$. This will even hinder the practicality of the one-step prediction formula for large number of spatial locations. In contrast, the sequential prediction formula explicitly integrate model parameters $\{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2\}$ without relying on Monte Carlo approximations.

3.2.5 | Computational cost

The PP cokriging model can allow efficient computations in output space due to the following reasons. In parameter estimation, each iteration of the MCEM algorithm in Algorithm 1 of the Supplementary Material requires the computation of so-called Q-function in E-step of MCEM and its numerical optimization with respect to correlation parameters $\boldsymbol{\phi}_t$ at each level of code. The evaluation of Q-function requires matrix inversions and matrix multiplication of size $\tilde{n}_t \times \tilde{n}_t$. Such an evaluation requires $O(MN\tilde{n}_t^2 + \tilde{n}_t^3)$ computational cost across N spatial locations and M Monte Carlo samples. If the numerical optimization requires k evaluations of Q-function to find the optimal value, the overall computational cost in each iteration of the MCEM algorithm is $O(kMN\sum_{t=1}^{s}\tilde{n}_t^2 + k\sum_{t=1}^{s}\tilde{n}_t^3)$ without any parallelization. Notice that parallelization across t is possible according to Algorithm 1 of the Supplementary Material. This is a one-time computational cost. In the predictive distribution in Equation (9), each conditional distribution requires matrix inversions and matrix multiplication of size $\tilde{n}_t \times \tilde{n}_t$. This requires $O(\tilde{n}_t^3 + N\tilde{n}_t^2)$ computational cost. One random sample generated from the predictive distribution at one new input value requires $O(\sum_{t=1}^{s}\tilde{n}_t^3 + N\sum_{t=1}^{s}\tilde{n}_t^2)$ computational cost. As $\tilde{n}_t$ is typically small (a few hundreds at most) in many real applications, the computational cost in prediction is linear in the number of spatial locations, N. This indicates the scalability of the proposed method to handle high-dimensional output for multifidelity computer models.

3.3 | Near equivalence of PP cokriging and separable cokriging

The PP cokriging emulator turns out to have nearly the same marginal predictive distributions at each spatial location as a *separable autoregressive cokriging* model, where the separability refers to the fact that unknown spatial correlation matrices $\boldsymbol{\Sigma} := \{\boldsymbol{\Sigma}_t : t = 1, \ldots, s\}$ on the spatial domain are assumed for the Gaussian process that approximates the level 1 code and the Gaussian processes in the location-scale discrepancy function. This result is an analogy of Theorem 6.1 of Gu and Berger (2016) for autoregressive cokriging models. In what follows, we will assume a nested design. This leads to the following matrix-normal distribution for the separable autoregressive cokriging with s levels:

$$L\left(\mathbf{y}^{\mathscr{D}} | \mathbf{B}, \Gamma, \boldsymbol{\Sigma}, \boldsymbol{\phi}\right) = \mathcal{MN}_{n_1,N}(\mathbf{y}_1 | \mathbf{H}_1\mathbf{B}_1, \mathbf{R}_1, \boldsymbol{\Sigma}_1) \times \prod_{t=2}^{s} \mathcal{MN}_{n_t,N}(\mathbf{y}_t | \mathbf{H}_t\mathbf{B}_t + W_{t-1}\Gamma_{t-1}, \mathbf{R}_t, \boldsymbol{\Sigma}_t), \tag{10}$$

where $\mathcal{MN}_{n_t,N}(\cdot,\cdot,\cdot)$ is a $n_t \times N$ matrix normal distribution. $\mathbf{B} := \{\mathbf{B}_1, \ldots, \mathbf{B}_s\}$ with $\mathbf{B}_t := [\boldsymbol{\beta}_{t,1}, \ldots, \boldsymbol{\beta}_{t,N}]$ being a $q_t \times N$ matrix of unknown mean parameters. $W_{t-1} := [y_{t-1,1}(\mathcal{X}_t), \ldots, y_{t-1,N}(\mathcal{X}_t)]$ is an $n_t \times N$ matrix. $\Gamma_{t-1} := \text{diag}\{\gamma_{t-1,1}, \ldots, \gamma_{t-1,N}\}$ is an $N \times N$ diagonal matrix.

We can show that under the constant prior on the location parameters and any prior on spatial correlation matrices $\{\boldsymbol{\Sigma}_t : t = 1, \ldots, s\}$, the resulting predictive mean in the separable autoregressive cokriging model is simply the predictive mean in the PP cokriging model, and the resulting predictive variance in the separable autoregressive cokriging model is almost equal to the predictive variance in the PP cokriging model. Thus, if only the mean and marginal predictive variance are concerned, there is no need to introduce spatial correlation structures in the output space. This fact is established in the next theorem with its proof given in Section S.8 of the Supplementary Material.

Theorem 1 *For a separable autoregressive cokriging model with likelihood in Equation (10), given the objective prior*

$$\pi(\mathbf{B}_1, \ldots, \mathbf{B}_s, \Gamma_1, \ldots, \Gamma_{s-1} | \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_s, \boldsymbol{\phi}) \propto 1 \tag{11}$$

for the parameters of mean functions and scale discrepancy functions, the following hold:

1. *The posterior predictive mean at level t in the separable autoregressive cokriging emulator, at an unobserved input $\mathbf{x}_0$ and at spatial coordinate j, is identical to the PP cokriging emulator mean.*
2. *The posterior predictive variance at level t in the separable autoregressive cokriging emulator, at an unobserved input $\mathbf{x}_0$ and at spatial coordinate j, depends on $\boldsymbol{\Sigma}_t$ through the posterior mean of the* jth *diagonal term $E[\boldsymbol{\Sigma}_t^{jj} | \mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi}_t]$; it is identical to the PP cokriging emulator variance, if $E[\boldsymbol{\Sigma}_t^{jj} | \mathbf{y}^{\mathscr{D}}, \boldsymbol{\phi}_t] = (n_t - q_t)\hat{\sigma}_{t,j}^2/(n_t - q_t - 2)$, where $\hat{\sigma}_{t,j}^2$, defined in Lemma 1 of the Supplementary Material, is an estimator of $\sigma_{t,j}^2$ under nested design.*

Theorem 1 indicates that the PP cokriging emulator and the separable autoregressive cokriging emulator can have the same predictive mean and marginal predictive variance under nested design when $n_t - q_t$ is large as in practice, since the new posterior expectation of $\mathbf{\Sigma}_t^{jj} = \sigma_{t,j}^2$ will be approximately the same as $\hat{\sigma}_{t,j}^2$. Moreover, when the scale discrepancy function $\gamma_{t,j}$ is fixed at one. It is easy to check that the results in Theorem 1 still hold. When the design is not nested, it can be readily checked that PP cokriging possesses the properties as in the nested design with the same proof except for notational difference by introducing the missing data. That is, the fact that missing data are available only affects how computation is carried out and does not alter the properties of the PP cokriging emulator. Thus, in practice, the predictive mean and predictive variance in PP cokriging can be used in practice as long as posterior draws from the predictive distribution over the spatial domain is not concerned. In addition, risk assessment of storm surges in the FEMA report (FEMA, 2017) only requires computation of the AEP at each location, which is a function of predictive mean and predictive variance.

4 | ANALYSIS OF STORM SURGE SIMULATIONS

In this section, the PP cokriging emulator is used to analyse high-dimensional output from the ADCIRC simulator and the ADCIRC + SWAN simulator.

The analysis of emulation results and numerical comparison is presented to demonstrate the advantage of the parallel partial cokriging model with high-dimensional output. The proposed PP cokriging methodology is implemented in the R package `ARCokrig` (Ma, 2019). The PP cokriging model is trained on 200 inputs from the ADCIRC simulator and 60 inputs from the ADCIRC + SWAN simulator, where 50 inputs from the second fidelity level are nested within the first fidelity level. With such model runs, the proposed method can be applied readily. To measure predictive performance, we run the ADCIRC + SWAN simulator at 166 inputs from the original 226 inputs after excluding 60 inputs as described in Section 2.3.

Moreover, we train the PP kriging emulator via the R package `RobustGaSP` (Gu et al., 2019) with the same 60 high-fidelity runs used in the PP cokriging emulator. As the landfall location is along the coastline, we define a distance measure d_{ℓ} to replace the actual longitude and latitude coordinate of the landfall location. Specifically, we first choose a reference location ℓ_0 to be the landfall location that is in the most northwest direction along the coastline. Then for any landfall location ℓ, d_{ℓ} is defined as the spherical distance between ℓ and ℓ_0. As the coastline is unique, the landfall location determines the distance measure d_{ℓ} and vice versa. In the implementation of the PP kriging emulator and the PP cokriging emulator, the input variables are $\Delta P, R_p, V_f, \theta, B, d_{\ell}$. Evaluation of predictive performance is based on root-mean-squared-prediction errors (RMSPE), coverage probability of the 95% equal-tail credible interval (CVG(95%)), average length of the 95% equal-tail credible interval (ALCI(95%)), and continuous rank probability score (CRPS; Gneiting & Raftery, 2007). We also compute the Nash–Sutcliffe model efficiency coefficient (NSME):

$$\text{NSME} := 1 - \frac{\sum_{j=1}^{N}\sum_{\mathbf{x}\in A}\{m_j(\mathbf{x}) - y_{2,j}(\mathbf{x})\}^2}{\sum_{j=1}^{N}\sum_{\mathbf{x}\in A}\{m_j(\mathbf{x}) - \bar{y}_{2,j}\}^2},$$

where $m_j(\mathbf{x})$ is the value to predict the high-fidelity simulator $y_{2,j}(\cdot)$ at input $\mathbf{x}$ and j-th spatial coordinate and $\bar{y}_{2,j} := \sum_{\mathbf{x}\in\mathcal{X}_2} y_{2,j}(\mathbf{x})/n_2$ is the average of code $y_{2,j}(\cdot)$ at j-th spatial coordinate. The NSME computes the residual variance with the total variance, which has a similar meaning as the coefficient of determination. The closer NSME is to 1, the more accurate the model is. If the ADCIRC simulator is used to predict the ADCIRC + SWAN simulator at these 166 new inputs, NSME is −1.089, which indicates that the mean of the training data in the high-fidelity simulator is a much better predictor than the low-fidelity simulator at these inputs.

4.1 | Emulation accuracy

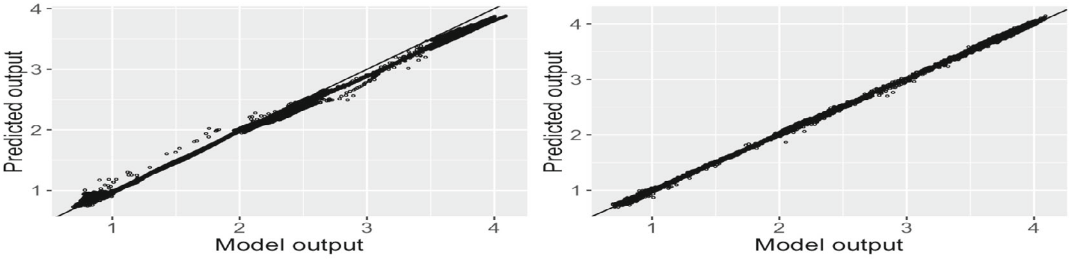
In the PP cokriging model, we include constant basis functions $\mathbf{h}_t(\cdot)$ according to exploratory data analysis. The scale discrepancy function $\gamma_{t-1,j}$ is assumed to be an unknown constant that does not depend on input. This assumption still allows the scale discrepancy function to vary across different fidelity levels and spatial coordinates. For parameter estimation, the MCEM algorithm is initialized with multiple starting values and took about 5 h to achieve convergence for a pre-specified tolerance on a 2-core Macbook Pro with 8 GB random access memory. The predictive mean and predictive variance is approximated by 30 random draws from the distribution in Equation (9). Negligible improvement is seen from increasing the number of draws. The estimated range parameters show that the peak surge elevation is highly dependent on the inputs: central pressure deficit (ΔP), Holland's B parameter (B), since these two inputs have relatively large range parameters compared to their input ranges in the training sets. The small impact of the landfall location (ℓ) is due to our focus on a small coastal region in Cape Coral.

A direct approach to building an emulator for the high-fidelity model run is to use the PP kriging emulator trained only with the high-fidelity simulations. As an additional comparison, we also include the prediction results based on the PP kriging emulator using the 200 low-fidelity runs only. The results in Table 2 show that the PP cokriging emulator gives best prediction results all the three emulators, since the PP cokriging model gives smallest RMSPE and CRPS and largest CVG and NSME. It is also interesting to see that PP cokriging is able to give largest CVG with a modest ALCI. As a baseline, the low-fidelity simulator is directly used to predict the high-fidelity simulator over these 166 new inputs, resulting in the RMSPE at 0.132, which is similar to the RMSPE obtained using the PP cokriging emulator trained on the low-fidelity runs only. The PP kriging emulator trained on high-fidelity runs only gives much better CVG and CRPS than it does when trained on the low-fidelity runs only, while PP kriging emulator trained on the high-fidelity runs only gives larger RMSPE than it does when trained on the low-fidelity runs only. This indicates that the folklore of using $10d$ runs for emulating the high-fidelity simulator does not perform satisfactorily and that the low-fidelity simulator is a biased surrogate model for the high-fidelity simulator in the storm surge application. Figure 3 indicates that the PP cokriging emulator performs much better than the PP kriging emulator at the input setting $\mathbf{x}_1$, since its predicted PSE are scattered around the 45-degree straight line. In contrast, the PP kriging emulator has a comparably worse performance. All these numerical and visual results confirm the advantage of using the PP cokriging emulator in the storm surge application by effectively borrowing information across different fidelity levels.

In the storm surge application, the high-fidelity simulator is about 10 times slower than the low-fidelity simulator. Increasing the number of model runs in the high-fidelity simulator is therefore computationally prohibitive. The computational cost of predicting a new high-fidelity model run via the PP cokriging emulator is negligible compared to that needed to get a single run from the actual ADCIRC + SWAN simulator. This implies that emulating the high-fidelity simulator by using our proposed PP cokriging emulator that combines only a small number of high-fidelity runs and a few hundred low-fidelity runs is preferable than using the low-fidelity simulator, in

TABLE 2 Predictive performance of emulators at $n^* = 166$ held-out inputs over all $N = 9{,}284$ spatial locations. PP kriging was trained based on low-fidelity runs and high-fidelity runs separately. PP cokriging was trained based on both the low-fidelity simulation and high-fidelity simulation. PP = parallel partial

	RMSPE	CVG(95%)	ALCI(95%)	CRPS	NSME
PP kriging with low-fidelity data	0.130	0.042	0.023	0.109	0.979
PP kriging with high-fidelity data	0.174	0.913	0.532	0.083	0.966
PP cokriging	0.040	0.992	0.257	0.024	0.998



(a) PP kriging trained on high-fidelity runs only

(b) PP cokriging

FIGURE 3 Scatter plot of predicted peak surge elevation against held-out PSE over $N = 9284$ spatial locations at the input setting $\mathbf{x}_1$

terms of both accuracy and computational cost. The capability to use the low fidelity simulator, without substantial loss of accuracy through use of the PP cokriging emulator, to explore more of the parameter space greatly enhances the feasibility of achieving high-precision modeling results without a massive computational budget.

4.2 | Uncertainty analysis

Cross-validation in the previous section showed that the PP cokriging emulator can provide very accurate predictions when compared to the true high-fidelity surge model runs in an overall sense. Figure 4 compares the predicted storm surges against held-out storm surge from the high-fidelity surge model across $N = 9284$ spatial locations at two storm inputs that are used in Figures 2 and 3. At these two inputs, the PP cokriging emulator seems to have large predictive uncertainties in the southeast region of Cape Coral and to have small predictive uncertainties in the Pine Island Sound Aquatic Preserve and the Caloosahatchee River. The largest predictive standard deviation in the PP cokriging emulator across all spatial locations is around 0.2, which is smaller than the difference between the high-fidelity simulator and the low-fidelity simulator.

Next, we explore the relationship between storm inputs and error structures in the PP cokriging emulator. We compute the prediction errors across all spatial locations at all held-out inputs. Figure 5 shows that the majority of emulation errors range from −0.5 to 0.5. This indicates that the PP cokriging emulator can capture the input–output relationship quite well. The residuals become larger as the central pressure deficit and the forward speed increase. The scale pressure radius seems to impact the emulation error in an opposite way as central pressure deficit. The residuals across different spatial locations are different as shown in Figure 5. This indicates that

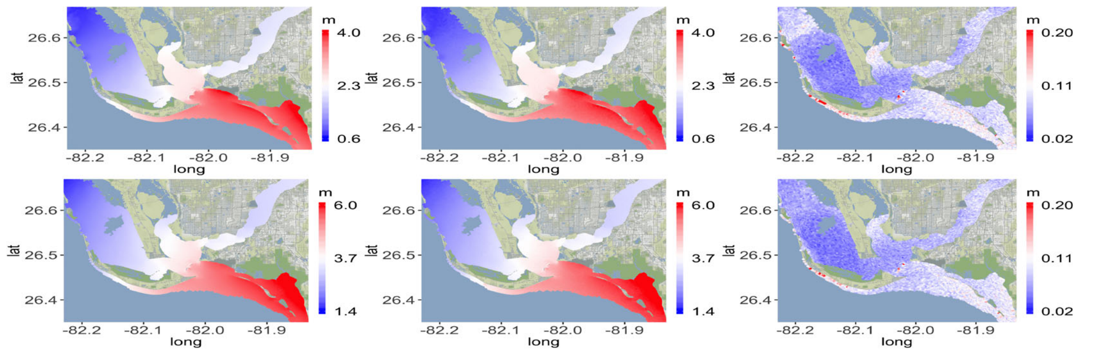


FIGURE 4 High-fidelity runs and predicted peak surge elevations with predictive standard errors at two input settings. The first column shows the high-fidelity runs at two different input settings. The second and third columns show the corresponding predicted peak surge elevation and associated predictive standard errors [Colour figure can be viewed at wileyonlinelibrary.com]

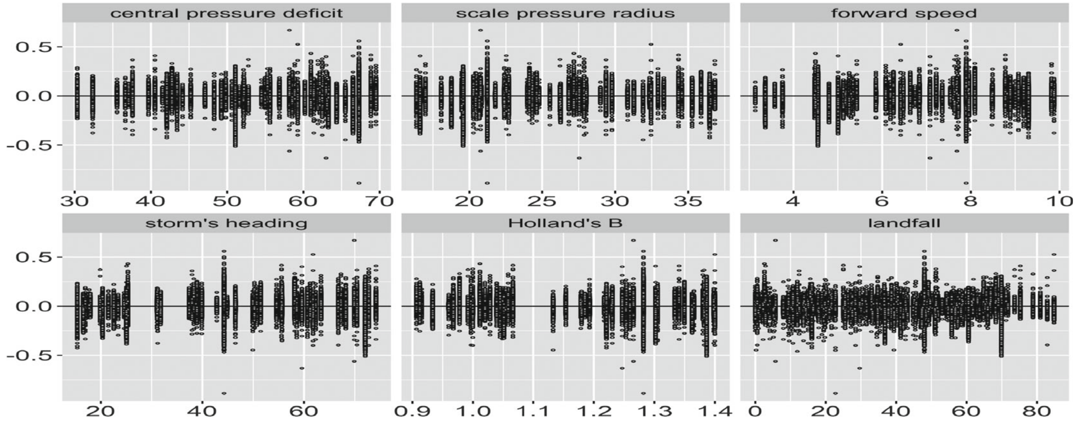


FIGURE 5 Prediction errors across all $N = 9284$ spatial locations against each storm parameter at all held-out inputs

the current PP cokriging emulator can partially capture the inhomogeneous structures in the output space, and some variations due to inputs are still left; see Section 5 for discussions on nonstationarity modelling in input space.

Finally, we show the parameter estimates for $\boldsymbol{\beta}_1$, $\boldsymbol{\sigma}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}_1$, and $\boldsymbol{\sigma}_2$ in Figure 6. As we can see, these estimated parameters show strong spatially varying structures at different regions. The estimated regression parameters $\hat{\boldsymbol{\beta}}_1$ and standard deviation $\hat{\boldsymbol{\sigma}}_1$ at the low-fidelity level seem to be smoother than those estimates at fidelity level 2. This is because more variations are captured by the Gaussian process at the low-fidelity level. The remaining variations captured by the discrepancy function $\delta_{2,j}(\cdot)$ are small. This implies that the Gaussian process at the low-fidelity level fits well with model runs from the ADCIRC simulator and the discrepancy between the low-fidelity simulator and the high-fidelity simulator is relatively small. The estimated scale discrepancy parameters $\hat{\boldsymbol{\gamma}}_1$ at all locations also show strongly heterogeneous spatial structures with values slightly greater than 1. This indicates that the high-fidelity simulator is more likely to generate higher values of storm surges than the low-fidelity simulator, but this trend is very

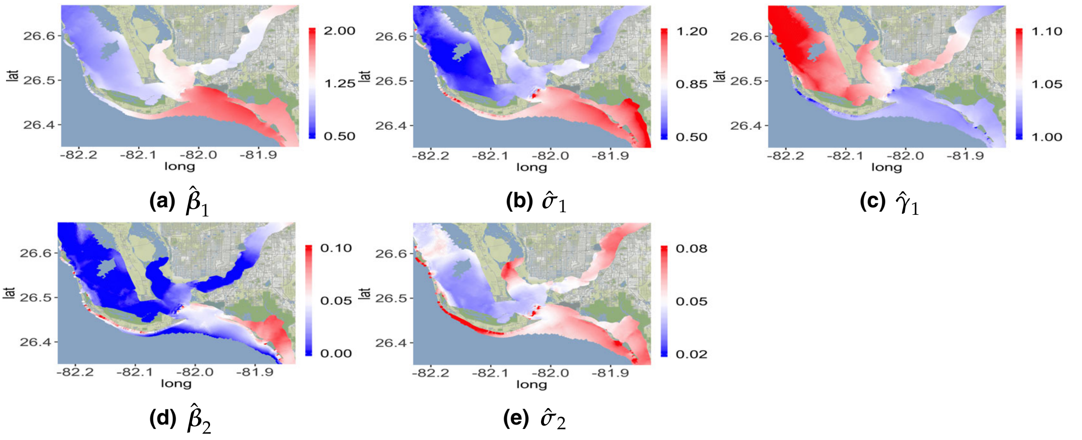


FIGURE 6 Estimated parameters across all spatial locations. The estimated parameters show strong heterogeneous spatial patterns [Colour figure can be viewed at wileyonlinelibrary.com]

small. The estimated standard deviations $\hat{\boldsymbol{\sigma}}_1$ and $\hat{\boldsymbol{\sigma}}_2$ seem to have more local structures than their corresponding regression parameters. This makes sense because we expect the regression trend in Gaussian processes to capture large-scale variations, and covariance structure to capture small-scale variations.

A central task in coastal engineering is to perform risk assessment of storm surges. The proposed emulator can produce predictive mean and predictive variance of storm surges over large number of spatial locations, which are practically useful and advantageous in risk assessment of storm surges in coastal flood hazard studies (e.g. Cialone et al., 2017; FEMA, 2017; Niedoroda et al., 2010), because computation of annual exceedance probabilities at one frequency level can require several thousands of storm surge simulations generated by a surge prediction model, which is prohibitively costly (e.g. FEMA, 2017). Practitioners in ocean engineering can use the proposed PP cokriging emulator to evaluate AEP using Monte Carlo approximations. This would be computationally efficient as running large number of storm surge simulations is not required once the emulator is readily available. The application of an emulator in risk assessment of storm surges is pursued elsewhere.

5 | DISCUSSION

Coastal flood hazard studies by FEMA and USACE use ADCIRC + SWAN to quantify the storm surge hazard, where simulation from this computer model is time-consuming and resource-intensive. We have built a parallel partial cokriging emulator to predict storm surges using simulations from both ADCIRC + SWAN and ADCIRC. The PP cokriging emulator effectively provides the marginal predictive distributions of storm surges for any storm characteristics at each spatial location. These marginal predictive distributions are also almost identical to the marginal predictive distributions obtained in situations where any spatial correlation matrix is used. These marginal predictive distributions can be directly used to compute annual exceedance probabilitiies in coastal flood hazard studies, which provide a convenient tool for practitioners in ocean engineering to perform risk assessment and surge forecasting research.

The PP cokriging emulator not only has similar prediction accuracy as the high-fidelity simulator ADCIRC + SWAN, but also has a linear computational cost in terms of output values and also induces nonstationarity in output space, which is crucial to capture non-smooth storm surge surface. Field measurements of historical storm data including both observed surges and characteristics of storms that making landfall in Southwest Florida (SWFL) are very limited with only one gauge station operating in Fort Myers. If there were sufficient storm data available, the proposed emulator could be further used to model the discrepancy between actual storm surges and ADCIRC + SWAN simulations, since accurate prediction of the actual storm surges can help perform more accurate coastal flood hazard studies. Although this paper focuses on a small region in SWFL, there is also a significant interest in extending the current method from a local region to the entire SWFL region so that the relationship between AEP and storm parameters can be quantified over a much larger coastal region.

The PP cokriging emulator assumes conditional independence across spatial locations that essentially leads to a separable covariance structure between input space and output space to simplify computations. This assumption can help capture nonstationary spatial patterns in the storm surge application. If interest lies in joint modelling across spatial locations, one can choose a spatial window to enable joint modelling. A related concern is the assumption of common correlation parameters at all spatial locations. If correlation parameters differ at each spatial location, the computational cost would not be linear in terms of spatial coordinates. This is a key advantage when the hazard from storm surge is assessed over a large spatial domain. One can potentially partition the domain into a set of subregions and allow different correlation parameters across these subregions.

In the storm surge application, we used a limited number of runs to train the emulator due to computational constraints. To aid coastal flood hazard studies and storm surge forecasting, one may use the proposed PP cokriging emulator to set-up the design in a statistical optimal way such as sequential design (Le Gratiet & Cannamela, 2015) or to use large number of model runs. The latter problem can be tackled via computationally efficient Gaussian process approximation approaches (e.g. Gramacy & Apley, 2015). The storm surge output shows quite rough structure across the spatial domain due to hurricane characteristics and heterogeneous topography. One can introduce nonstationarity in input space via treed Gaussian process (Gramacy & Lee, 2008; Konomi & Karagiannis, 2021). Another interesting exploration for the proposed methodology is related to non-nested design on how much gain is obtained by allowing the design is not hierarchically nested over the traditional nested design. There is also an interest in developing emulators when multifidelity computer codes have different spatial resolutions. These possible directions could be pursued in future.

ACKNOWLEDGEMENTS

This research began under the auspices and support of the Statistical and Applied Mathematical Sciences Institute (SAMSI) 2018-2019 research program on Model Uncertainty: Mathematical and Statistical (MUMS) in the Storm Surge Working Group that was supported by the U.S. National Science Foundation (NSF) under Grant DMS-1638521. This material is partly based on work supported by the U.S. Department of Homeland Security (DHS) under Grant Award Number 2015-ST-061-ND0001-01 and by the U.S. Coastal Research Program (USCRP) administered by the U.S. Army Corps of Engineering (USACE), Department of Defense (DoD). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, DHS and DoD, and no official endorsement should be inferred. Ma's work used the Extreme Science and Engineering Discovery Environment (XSEDE) Lonestar5 and Stampede2 at the Texas Advanced Computing Center (TACC) through allocation DMS180042. Ma gratefully acknowledges the support by the postdoctoral fellowship at SAMSI and by the USCRP program under Cooperative Agreement Number W912HZ-2020011 while part of this research was conducted at SAMSI and Duke University. Ma is grateful to Professor Jim Berger for his valuable discussion. Asher gratefully acknowledges the support by the USCRP program under Cooperative Agreement Number W912HZ-2020011 while part of this research was conducted. Toro gratefully acknowledges in-house research funding provided by Lettis Consultants International, Inc. for making possible his participation.

DATA AVAILABILITY STATEMENT

The full simulation can be found from the Coastal Flood Modeling Database - Southwest Florida (Asher and Liu 2022).The dataset in real application is available upon request.

SUPPLEMENT MATERIAL

The Supplement Material contains technical details and additional results. Code for the numerical examples can be found in https://github.com/pulongma/PPCokriging.

ORCID

Pulong Ma https://orcid.org/0000-0002-9847-1868
Taylor G. Asher https://orcid.org/0000-0003-0096-1680

REFERENCES

Aerts, J.C., Botzen, W.W., Emanuel, K., Lin, N., De Moel, H. & Michel-Kerjan, E.O. (2014) Evaluating flood resilience strategies for coastal megacities. *Science*, 344(6183), 473–475.

Asher, T.G. & Liu, Y. (2022) Coastal flood modeling database - Southwest Florida. PRJ-2433, DesignSafe-CI. Available at: https://www.designsafe-ci.org/data/browser/public/. In review.

Blanton, B., Dresback, K., Colle, B., Kolar, R., Vergara, H., Hong, Y. et al. (2020) An integrated scenario ensemble-based framework for hurricane evacuation modeling: Part 2-hazard modeling. *Risk Analysis*, 40(1), 117–133.

Booij, N., Ris, R.C. & Holthuijsen, L.H. (1999) A third-generation wave model for coastal regions: 1. Model description and validation. *Journal of Geophysical Research: Oceans*, 104(C4), 7649–7666.

Cardone, V.J. & Cox, A.T. (2009) Tropical cyclone wind field forcing for surge models: critical issues and sensitivities. *Natural Hazards*, 51(1), 29–47.

Cialone, M.A., Grzegorzewski, A.S., Mark, D.J., Bryant, M.A. & Massey, T.C. (2017) Coastal-storm model development and water-level validation for the North Atlantic coast comprehensive study. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 143(5), 4017031.

Conti, S. & O'Hagan, A. (2010) Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140(3), 640–651.

Cressie, N. (1993) *Statistics for spatial data*. New York: John Wiley & Sons. Revised edition.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1–38.

Dietrich, J., Zijlema, M., Westerink, J., Holthuijsen, L., Dawson, C., Luettich, R. et al. (2011) Modeling hurricane waves and storm surge using integrally-coupled, scalable computations. *Coastal Engineering*, 58(1), 45–65.

Dietrich, J.C., Tanaka, S., Westerink, J.J., Dawson, C.N., Luettich, R.A., Zijlema, M. et al. (2012) Performance of the unstructured-mesh, SWAN + ADCIRC model in computing hurricane waves and surge. *Journal of Scientific Computing*, 52(2), 468–497.

FEMA. (2006) *Hurricane Katrina in the Gulf coast*. Washington, D.C.: Federal Emergency Management Agency.

FEMA. (2008) Mississippi coastal analysis project. Project reports prepared by URS Group Inc., (Graithersburg MD and Tallahassee FL) under HMTAP Contract HSFEHQ-06-D-0162, Task Order 06-J-0018.

FEMA. (2017) Southwest Florida Storm Surge Study Intermediate Data Submittal 2. Report Prepared by the Risk Assessment, Mapping, and Planning Partners (RAMPP) under FEMA IDIQ Contract HSFEHQ-09-D-0369 and Task Order HSFE04-13-J-0097. FEMA Region IV.

Fischbach, J., Johnson, D. & Kuhn, K. (2016) Bias and efficiency tradeoffs in the selection of storm suites used to estimate flood risk. *Journal of Marine Science and Engineering*, 4(1), 10.

Georgas, N., Blumberg, A., Herrington, T., Wakeman, T., Saleh, F., Runnels, D. et al. (2016) The stevens flood advisory system: operational H3E flood forecasts for the greater New York/New Jersey metropolitan region. *International Journal of Safety and Security Engineering*, 6(3), 648–662.

Gneiting, T. & Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Gramacy, R.B. & Apley, D.W. (2015) Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2), 561–578.

Gramacy, R.B. & Lee, H.K.H. (2008) Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483), 1119–1130.

Gu, M. (2019) Jointly robust prior for Gaussian stochastic process in emulation, calibration and variable selection. *Bayesian Analysis*, 14(3), 877–905.

Gu, M. & Berger, J.O. (2016) Parallel partial Gaussian process emulation for computer models with massive output. *The Annals of Applied Statistics*, 10(3), 1317–1347.

Gu, M., Palomo, J. & Berger, J.O. (2019) RobustGaSP: Robust Gaussian stochastic process emulation in R. *The R Journal*, 11(1), 112–136.

Hesser, T.J., Cialone, M.A. & Anderson, M.E. (2013) Lake St. Clair: Storm wave and water level modeling.

Higdon, D., Gattiker, J. & Williams, B. (2008) Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482), 570–583.

Jensen, R.E., Cialone, M.A., Chapman, R.S., Ebersole, B.A., Anderson, M. & Thomas, L. (2012) Lake Michigan storm: Wave and water level modeling.

Kennedy, M. & O'Hagan, A. (2000) Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1), 1–13.

Konomi, B.A. & Karagiannis, G. (2021) Bayesian analysis of multifidelity computer models with local features and non-nested experimental designs: Application to the WRF model. *Technometrics*, 63(4), 510–522.

Le Gratiet, L. (2013) Bayesian analysis of hierarchical multifidelity codes. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1), 244–269.

Le Gratiet, L. & Cannamela, C. (2015) Cokriging-based sequential design strategies using fast cross-validation techniques for multi-fidelity computer codes. *Technometrics*, 57(3), 418–427.

Liu, Y., Asher, T.G. & Irish, J.L. (2019) Physical drivers of changes in probabilistic surge hazard under sea level rise. *Earth's Future*, 7(7), 819–832.

Luettich, R. & Westerink, J. (2004) Formulation and numerical implementation of the 2D/3D ADCIRC finite element model version 44.XX.

Ma, P. (2019) ARCokrig: Autoregressive cokriging models for multifidelity codes. R package version 0.1.2. Available at: https://CRAN.R-project.org/package=ARCokrig

Ma, P. (2020) Objective Bayesian Analysis of a Cokriging Model for Hierarchical Multifidelity Codes. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4), 1358–1382.

Ma, P. & Bhadra, A. (2022) Beyond Matérn: on a class of interpretable confluent hypergeometric covariance functions. *Journal of the American Statistical Association*. Available from: https://doi.org/10.1080/01621459.2022.2027775

Marsooli, R. & Lin, N. (2018) Numerical modeling of historical storm tides and waves and their interactions along the US east and Gulf coasts. *Journal of Geophysical Research: Oceans*, 123(5), 3844–3874.

de Moel, H. & Aerts, J. (2011) Effect of uncertainty in land use, damage models and inundation depth on flood damage estimates. *Natural Hazards*, 58(1), 407–425.

Niedoroda, A., Resio, D., Toro, G., Divoky, D., Das, H. & Reed, C. (2010) Analysis of the coastal Mississippi storm surge hazard. *Ocean Engineering*, 37(1), 82–90.

NOAA National Centers for Environmental Information. (2019) U.S. billion-dollar weather and climate disasters. Available at: https://www.ncdc.noaa.gov/billions/

Pielke Jr, R.A., Gratz, J., Landsea, C.W., Collins, D., Saunders, M.A. & Musulin, R. (2008) Normalized hurricane damage in the United States: 1900–2005. *Natural Hazards Review*, 9(1), 29–42.

Qian, P.Z.G. & Wu, C.F.J. (2008) Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50(2), 192–204.

Rappaport, E.N. (2014) Fatalities in the United States from Atlantic tropical cyclones: new data and interpretation. *Bulletin of the American Meteorological Society*, 95(3), 341–346.

Sacks, J., Welch, W.J., Mitchell, T.J. & Wynn, H.P. (1989) Design and analysis of computer experiments. *Statistical Science*, 4(4), 409–435.

Tanaka, S., Bunya, S., Westerink, J.J., Dawson, C. & Luettich, R.A. (2011) Scalability of an unstructured grid continuous Galerkin based hurricane storm surge model. *Journal of Scientific Computing*, 46(3), 329–358.

Toro, G., Niedoroda, A., Reed, C. & Divoky, D. (2010) Quadrature-based approach for the efficient evaluation of surge hazard. *Ocean Engineering*, 37(1), 114–124. A Forensic Analysis of Hurricane Katrina's Impact: Methods and Findings.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A. et al. (2014) XSEDE: accelerating scientific discovery. *Computing in Science & Engineering*, 16(5), 62–74.

Wamsley, T., Godsey, E., Bunch, B.W., Chapman, R.S., Gravens, M.B., Grzegorzewski, A.S. et al. (2013) Mississippi coastal improvement program, barrier island restoration numerical modeling.

Wei, G.C.G. & Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411), 699–704.

Westerink, J.J., Luettich, R.A., Feyen, J.C., Atkinson, J.H., Dawson, C., Roberts, H.J. et al. (2008) A basin-to channel-scale unstructured grid hurricane storm surge model applied to southern Louisiana. *Monthly Weather Review*, 136(3), 833–864.

Yin, J., Lin, N. & Yu, D. (2016) Coupled modeling of storm surge and coastal inundation: a case study in New York City during Hurricane Sandy. *Water Resources Research*, 52(11), 8685–8699.

Zijlema, M. (2010) Computation of wind-wave spectra in coastal waters with SWAN on unstructured grids. *Coastal Engineering*, 57(3), 267–277.

How to cite this article: Ma, P., Karagiannis, G., Konomi, B.A., Asher, T.G., Toro, G.R. & Cox, A.T. (2022) Multifidelity computer model emulation with high-dimensional output: An application to storm surge. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(4), 861–883. Available from: https://doi.org/10.1111/rssc.12558