The Annals of Applied Statistics
2021, Vol. 15, No. 1, 460–480
https://doi.org/10.1214/20-AOAS1398
© Institute of Mathematical Statistics, 2021

HIGH-FIDELITY HURRICANE SURGE FORECASTING USING EMULATION AND SEQUENTIAL EXPERIMENTS

BY MATTHEW PLUMLEE[1], TAYLOR G. ASHER[2], WON CHANG[3] AND MATTHEW V. BILSKIE[4]

[1]*Industrial Engineering and Management Sciences, Northwestern University,* mplumlee@northwestern.edu

[2]*Department of Marine Sciences, University of North Carolina,* tgasher@live.unc.edu

[3]*Division of Statistics and Data Science, University of Cincinnati,* won.chang@uc.edu

[4]*School of Environmental, Civil, Agricultural, and Mechanical Engineering, University of Georgia,* mbilskie@uga.edu

Probabilistic hurricane storm surge forecasting using a high-fidelity model has been considered impractical due to the overwhelming computational expense to run thousands of simulations. This article demonstrates that modern statistical tools enable good forecasting performance using a small number of carefully chosen simulations. This article offers algorithms that quickly handle the massive output of a surge model while addressing the missing data at unsubmerged locations. Also included is a new optimal design criterion for selecting simulations that accounts for the log transform required to statistically model surge data. Hurricane Michael (2018) is used as a testbed for this investigation and provides evidence for the approach's efficacy in comparison to the existing probabilistic surge forecast method.

1. Introduction.

Major hurricanes such as 2005's Katrina, 2012's Sandy and 2017's Irma produced tens of billions of dollars in damages. A majority of the damage comes from coastal flooding as high wind speeds from tropical cyclones push water from the ocean and coastal estuaries up onto land (Resio and Westerink (2008)). This phenomenon, called storm surge, is the rise of water associated with coastal flooding caused by storms such as tropical cyclones. It can reach elevations of several meters above sea level and can extend over several tens of kilometers. Storm surge is driven by tropical cyclone characteristics like location, heading, speed, intensity and size.

This article will focus on forecasting the surge from an incoming storm. Predicting coastal storm surge as a storm nears landfall is critical to improve evacuation management and can be used to evaluate damage for recovery (Walker et al. (2018)). The interface between coastal geometry and a storm is the key factor for predicting coastal flooding. In shallow areas the seafloor acts as a ramp for the wind-driven ocean currents to flow to higher elevations. Funneling land features can further amplify surges. Elevated overland areas like barrier islands can partially block lesser storm surges but may be overtopped by larger ones, substantially changing flooding patterns (Bilskie et al. (2015)). A storm landfalling slightly to the east would have a radically different surge profile compared to a storm landfalling slightly to the west because the winds on either side of a storm's center are directed opposite of each other. Today, computer models are considered the best way to understand and predict the complex interactions between coastal geometry and a storm.

1.1. *Computer models for forecasting storm surge.*

Storm surge is simulated by solving a set of partial differential equations known as the shallow water equations to yield water elevation and velocity in space and time (Bode and Hardy (1997), Resio and Westerink (2008)).

Received July 2019; revised September 2020.

Key words and phrases. Computer experiments, Gaussian process, sequential experiments, surrogate modeling.

A mesh of nodes, which are points in geographic space, is constructed to capture the shape of the seafloor and overland topography. The partial differential equations are then solved on the mesh and integrated forward in time over several days for a single storm simulation.

When using surge simulations for forecasting, the substantial uncertainty in meteorological forecast of the storm itself (Cangialosi (2019)) necessitates simulating multiple possible storm variations. Current surge forecasting in the U.S. employs a numerical model that can be run cheaply to permit thousands of simulations: The U.S. National Hurricane Center (NHC) publishes a probabilistic surge forecast product called P-Surge (Taylor and Glahn (2008)) that is based on the SLOSH storm surge model (Jelesnianski (1966), Jelesnianski, Chen and Shaffer (1992)). SLOSH is used partly because of its low computational expense; a single simulation can be run on a laptop in about one hour. SLOSH's speed and numerics were chosen given the computational capabilities of the time and have been maintained to provide a reliable and efficient model.

SLOSH's computational speed comes from a lower mesh resolution and simpler physics, such as neglecting nonlinear advection terms in the momentum equations. Our study uses the ADCIRC model (Westerink et al. (2008)) which, including differences in mesh resolution, can be over 100 times more computationally expensive than SLOSH. In addition to its more realistic physics, ADCIRC's numerics allow more flexible meshes that can resolve small features, such as coastal inlets and elevated highways, that act as crucial pathways and barriers to flooding (Bilskie et al. (2015)) while also covering entire ocean basins to reproduce the large-scale ocean response to storm forcing (Morey et al. (2006)). High-fidelity verification and calibration studies on ADCIRC indicate model errors are typically below half a meter (Bilskie et al. (2016), Dietrich et al. (2011)), though errors in forecasting are expected to be larger due to forecast uncertainty (Cyriac et al. (2018)). To manage the added computational cost, ADCIRC scales efficiently in parallel from one to thousands of processors on supercomputing systems (Tanaka et al. (2011)). Despite the advantages in accuracy of ADCIRC over SLOSH, operational forecasting using a computationally demanding model such as ADCIRC remains challenging. A generic (and exhaustive) thousand-member ensemble using the ADCIRC model for our case study would require on the order of 100,000 dedicated computer cores in a high performance computing configuration to forecast a single geographic region.

1.2. *Setting and approach.* This article explains an approach to forecasting peak storm surge using the high-fidelity ADCIRC model with a small number of carefully chosen runs per forecast period. Using only a small number of high-fidelity model runs is crucial to provide time-sensitive information. Our proposed approach uses only 10–30 runs every 12 hours. It does this by using statistical tools to predict water levels for unsimulated storms, borrow information across forecast periods and intelligently choose storms to be run. Our goal was to construct an approach with less than one hour of overhead computation using a standard desktop computer which is time aside from the ADCIRC simulation that cannot be easily parallelized.

The input to our process is the NHC forecasts, which are given every six hours and consist of information at future intervals on location of the storm, the maximum sustained wind speed and directional isotach radii from storm center. These forecasts typically predict storm states for 12 hr, 24 hr, 36 hr, 48 hr, 72 hr, 96 hr and 120 hr from the time of forecast if the tropical cyclone exists through that time period. A raw NHC storm forecast thus consists of over a hundred values representing storm properties. Instead of examining all of these values, we focused on the storm properties when the center of the storm first makes landfall because these determine the bulk of the peak surge. We use a six-dimensional characterization of the storms at landfall: latitude (LAT), longitude (LONG), heading (H), forward speed (FS), maximum wind speed (MWS) and radius of 34 kt isotach (R34). This characterization is unique

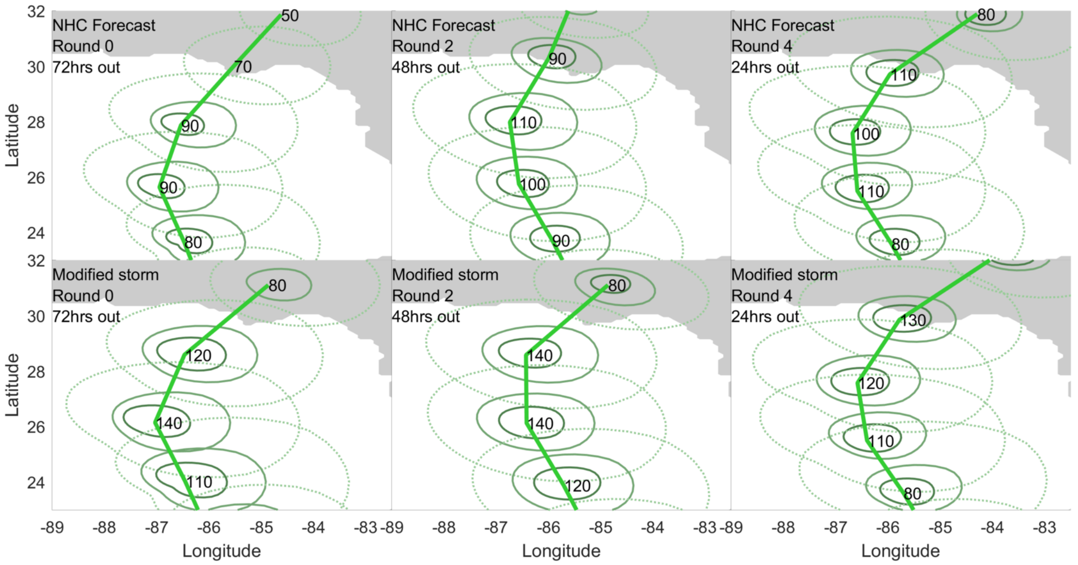

FIG. 1. *Hurricane forecast examples for our case study described in Section* 1.4. *The top panels are the NHC forecasts*, *and the bottom panels are hypothetical modified storms generated by our procedure described in Supplement A when the landfall characteristics are set to the* (*unknown at the time*) *actual landfall characteristics. The thick line represents the track of the center of the hurricane. The thin lines in circular shapes are the* 12 *hr increments of the* 64 *kt*, 50 *kt and* 34 *kt isotach lines*, *respectively* (*inward to outward*). *The numbers represent the maximum wind speed.*

to this paper, but it was anticipated to provide a good summary of storm properties relevant to storm surge. It also shares similarities with existing studies (Toro et al. (2010)). Once the landfall characteristics are specified or changed, the remaining aspects of the forecast are then filled based on current NHC forecast. Figure 1 illustrates the results of this operation for our case study. Using the historical error metrics published by the NHC, we also construct a probabilistic forecast of the storm's landfall characteristics. The derivation and procedure for drawing from the forecasted distribution of the storm's landfall characteristics is not the focus of this article, but details on the new procedure are given in Supplement A (Plumlee et al. (2021)).

Our approach involves existing and new computer model emulation and designed computer experiment tools (Currin et al. (1991), Sacks et al. (1989), Santner, Williams and Notz (2003)). The structure of our algorithm is to select computer experiments at each forecast period and then run those storms through the ADCIRC simulator. A fast emulator of the ADCIRC simulator is then built using this data set which provides an ability to integrate over thousands of storms without running them through the expensive ADCIRC simulator. The emulator is constructed using Gaussian process inference, also widely used in geostatistics (e.g., Cressie (1993)). The Gaussian process is used here as a statistical model of the surge response to predict at uninvestigated storms, not in a geographic sense to interpolate between nodes.

1.3. *Contributions summary.* The overarching goal of this project is demonstrating the effectiveness of statistical computer experiment inference for forecasting storm surge of an approaching tropical cyclone. Emulation for water level modelling has been considered previously (Parker et al. (2019), Rohmer and Idier (2012)) for forecasting surge with a fixed dataset (Jia and Taflanidis (2013), Jia et al. (2016), Taflanidis et al. (2012)). However, there appears to be no existing surge forecasting tool that both employs emulation techniques and adaptively selects new storms for sequential forecasting. We find that the integrative methodology described in this article results in significantly smaller predictive errors.

Several authors have applied data assimilation-based approaches to forecast storm surge (Altaf et al. (2014), Peng and Xie (2006)), though such efforts are functionally different from what we do here. Assimilation approaches leverage the covariance structure between observed and modeled water levels to produce a probabilistic prediction. They lack a mechanism to account for forecast uncertainties known to exist in the NHC storm forecast. Further, it is unclear whether such assimilation-based methods can provide useful information early enough to be of practical use (Asher et al. (2019)). This is because observations of surge, when the storm is far from landfall, are nearly uncorrelated with surge when the storm makes landfall, thus there is little information gained during assimilation in our forecast time period.

The novel contributions of this research can be broken into three categories. The first is the production of speedy and accurate statistical emulator of a deterministic computer model during forecasting. This required an agglomeration of tools but centered around parallel modeling and imputation. There are roughly half a million nodes of interest where peak surge needs to be simultaneously predicted. This can cause our computational overhead to swell if not carefully addressed (Chang et al. (2014), Gu and Berger (2016), Higdon et al. (2008)). We find in this setting that the partial parallel approach (Gu and Berger (2016)) is superior to a dimension reduction approach, such as that found in Chang et al. (2014). Some new justifications for a partial parallel approach from a frequentist perspective are in Supplement B (Plumlee et al. (2021)). The peak surges at nodes which were never below water on a given run are undefined physically and marked as missing in the software. This issue can consume our entire computational overhead if not consciously handled. A surface imputation approach that maintains interpolative properties proves effective.

Our second contribution is to enable the reuse of ADCIRC output from prior forecast periods in the current forecast. The simulation model effectively changes as new forecast information is given, even when the input vector of landfall properties remains the same. The bottom panels of Figure 1 illustrate this fact, where the same storm landfall characteristics produced slightly different storms because of alterations to other portions of the forecast outside of the landfall characteristics. This is because our approach used the NHC forecast and the current storm track to fill in the additional storm aspects aside from the landfall characteristics. One example of this effect is present in Figure 1, where at 72 and 48 hours out the NHC storm forecast diminished into landfall but at 24 hours out the NHC storm forecast reversed that trend and intensified into landfall. The bottom panels contain our modified NHC storm forecasts, where, even though all modified storms have the same landfall characteristics, the progression borrowed the diminishing or intensifying trend from the current NHC forecast. This differs from the emulation problem of computer models of dynamics (Conti et al. (2009)) and from computer models where the output given is time dependent (Liu and West (2009), Mak et al. (2018)). To account for this, a unique parameterization is used where the input vector includes the forecast time. Then, during prediction, the future landfall time is drawn and used in the input vector, correctly accounting for the future uncertainty induced by changes to the NHC forecast aside from landfall characteristics.

Lastly, our findings are that log-transformed surge data can be fit to a Gaussian process which has previously been used in conjunction with Gaussian process inference. Cressie (1990) dates this transform to the origin of kriging. However, this transform implies that the classical designed experiment criterion no longer functions, and a new criterion was needed. Designed experimentation on computer models in a sequential setting has been investigated in other contexts (Bect et al. (2012), Gramacy and Lee (2009), Ranjan, Bingham and Michailidis (2008)). In the practical context of storm surge prediction, we quantify the relative value of designed experimentation and find that the adaptive method significantly outperforms random draws of the storm characteristics from the forecast distribution.

TABLE 1

ADCIRC data collected and core hours used the case study described in Section 1.4. The marking $*$ on round 0 core-hours indicates the core hours are inflated due to cold starts of the ADCIRC model

			Experimental data		Verification data	
Round #	NHC advisory	Approx. hours until landfall	Runs	Core hours	Runs	Core hours
0	07	72	30	23,870*	0	0
1	09	60	30	12,569	0	0
2	11	48	20	6141	0	0
3	13	36	20	4976	20	5161
4	15	24	10	2077	10	1818
5	17	12	10	1219	10	1195
Total:			120	50,854	40	8174

1.4. *Michael case study.* A case study was needed to generate an efficacy evaluation of our approach. The computational demands of the ADCIRC model limited us to a single storm for this evaluation. We chose to study 2018's Hurricane Michael. Michael was a powerful but short-lived hurricane that caused 16 direct fatalities and $25 billion in damages in the U.S. as it made landfall in the Florida Panhandle (Beven, Berg and Hagen (2019)). Michael's forecast error was unusual. Forecast location errors were less than half the historical average, but forecast intensity errors were more than triple the historical average with forecasted intensity persistently below actual. This can be seen in Figure 1: the storm forecasts intensify closer to landfall.

In our study we reproduced the knowledge of Michael at select forecast intervals, which we call rounds, and let our algorithm evolve to meet the modeling demands observed at each interval. The computational cost breakdown of our case study using the configuration of ADCIRC from Bilskie, Hagen and Medeiros (2019) is provided in Table 1. Our attempt required a limited number of runs at prediction periods spaced 12 hours apart, doubling the NHC's six hour typical window to simplify this pilot study. We gave ourselves much more than 12 hours to build and test the statistical algorithm but kept the total run time of the algorithm within our one hour constraint.

Table 1 lists our case study's forecast periods, sample size and simulation computational cost. The hours until landfall approximate to within half an hour of actual landfall (Beven, Berg and Hagen (2019)). The number of runs per forecast period decreased in size as our accuracy goals appeared achievable with fewer runs. The computational cost per run is also decreasing as the ADCIRC simulation time decreases. The last three rounds of simulation included random simulations as verification data that we used to measure the accuracy of our emulator and evaluate designed experimentation in this setting compared to random experimentation.

1.5. *Structure of this article.* The remainder of the article is structured as follows. Section 2 describes the notation, and Section 3 describes our general statistical model, fitting details and our prediction algorithm for surge forecasting. Section 4 describes the new optimal experimental design approach to select each round's storms. Sections 5 and 6 give numerical analysis to evaluate the overall performance of the approach. The article offers some concluding remarks in Section 7.

2. Setting and data notation. In general terminology, our goal is to produce a probabilistic forecast of water level using a limited number runs from an expensive computer model of storm surge that changes based on the time at which it is evaluated. Toward this goal, we designed an emulator of the computer model as well as a sequential, optimal designed experiment tool that selects inputs to the computer model while accounting for the current forecast distribution. A more exhaustive description of the overall approach is located in Supplement A.

At some point in time t, our computer model, given landfall characteristics x, is a length M vector of peak surge labeled $f(x,t)$ with numerical values at wetted nodes and missing values for unwetted nodes. As stated in the Introduction, x, a landfall characteristic, is a six-dimensional vector of storm latitude, longitude, heading, forward speed, maximum wind speed and 34 kt isotach radius at landfall. We want to emphasize that the t in $f(x,t)$ refers to the time at which the model is evaluated because, depending on when the model is evaluated, the response will reflect some underlying properties of the current NHC forecast outside of x, as detailed in Section 1. After some time we will have generated surge data from n storms at $(x_1,t_1),\dots,(x_n,t_n)$ through ADCIRC simulations, where x_i is a vector of landfall characteristics for run i and t_i is the time of run i. Simulation outputs are stored and labeled as $f(x_i,t_i)$, a column vector of length M for all nodes of interest. Let $f_j(x_i,t_i)$ be peak water level for the jth node of the ith run. The geospatial locations of the M nodes are labeled $s_1,\dots,s_M$. We chose M as roughly half a million nodes—the computer model domain contains roughly two million nodes—covering overland areas and coastal waters with ground elevation at least -4 meters (i.e., four meters below sea level) in the study region.

Hurricane forecasts have uncertainty associated with prediction errors that decrease as the landfall draws near. The storm forecast distribution for the landfall characteristics reflecting such uncertainty at time t is labeled as $\Pi(t)$. This distribution is considered an input to our approach, specified by the NHC with some distributional assumptions pulled from historical accuracy measures. Moreover, the landfall time of the hurricane is unknown but can be drawn. The details of this process is in Supplement A. A random vector of landfall characteristics and landfall time drawn from this distribution is labeled $(\tilde{x},\tilde{\tau})$, and the surge response, given this random vector, is $f(\tilde{x},\tilde{\tau})$, where τ is used as the landfall time to distinguish it from t, the current time point.

Our statistical algorithm will take in the ADCIRC simulation data and the forecast distribution and return two items at an arbitrary time point t. The first item is a statistical emulator for ADCIRC simulation at time t. Due to the limited number of available runs, $f(x,\tau)$ is assigned a predictive distribution through statistical inference for any (x,τ) pair. To match current reporting practices at the NHC, we are only interested in the marginal distribution of each element in the vector $f(x,\tau)$. The second item we return is sequential experiments which are new storms that can be investigated at time t. That is, if we have budget for q storms, we want to choose $(x_{n+1},t_{n+1}),\dots,(x_{n+q},t_{n+q})$. We note that $t_{n+1}=\cdots=t_{n+q}=t$ by construction; thus, our decision is the choice of $x_{n+1},\dots,x_{n+q}$ using the past data and the current storm forecast distribution $\Pi(t)$. We do this by optimizing a design criterion that considers both the forecast distribution and the accuracy of the statistical estimation of $f(\tilde{x},\tilde{\tau})$.

We presume to have some initial data collected at $(x_1,0),\dots,(x_n,0)$. For our case study these landfall characteristics drawn were a Latin hypercube design of size 30. The landfall characteristic ranges were chosen to roughly cover the southeast U.S. Gulf Coast with LONG between -88.5 and -83.5 degrees, corresponding to the extent of the numerical model domain. The remaining landfall characteristics' ranges were roughly based on the observed ranges from historical hurricanes in the area and the current storm forecast: FS:[3 kts, 15 kts], H:[-10 deg, 40 deg], MWS:[60 kts, 160 kts] and R34:[75 nm, 200 nm].

3. Statistical modeling and forecasting.

This section explains our overall statistical modeling approach using collected data. This includes the dry node imputation, statistical inference and probabilistic forecasting.

3.1. *Imputation for spatial surge patterns.* We require fast simultaneous prediction at M ($\sim$ half a million) nodes. This eliminates a computation strategy where a prediction algorithm is used separately at each node. One simple approach that yields fast computation is to treat each node as independent with some shared parameters to allow for shared computation (Gu and Berger (2016)). However, surge data poses an important obstacle to the use of independent predictions: peak surge does not exist at some nodes for some runs. The data are missing not because of randomness nor computation error but due to the underlying physical system of interest. If a node is dry throughout the simulation, the peak surge is reported as a missing value. For an inland node, storm surge occurs only when the surge height is larger than the node's ground elevation. Moreover, higher ground near the ocean can block storm surge and create dry nodes behind it. Two approaches are to set these missing values to zero or to the ground elevation at the corresponding location. The former can lead to a large gap between simulated surges and the imposed zero values. The latter introduces an undesirable dependence on the ground elevation. Both of these methods lead to discarding the information from nearby nodes that can be used to impute reasonable surge values on dry nodes.

Our research found that within-surface imputation was an adequate strategy to surrogate the missing values in the sense that the overall forecasts are satisfactory and the prediction time is within the one hour time limit. The more sophisticated existing alternatives would not function well in our setting. For example, missing values in remote sensing problems are often imputed based on observations at both the same and different time steps (see, e.g., Cressie, Shi and Kang (2010)). Hung, Joseph and Melkote (2015) used a similar philosophy on high-dimensional computer model output with irregular grids. Transferring these ideas to the surge setting implies a dry node should be imputed based on runs where the same node is wet. One concern in using these type of algorithms in our setting is the intensive computational burden they would impose. Aside from the computational concern, storm surge data also involves complex missingness patterns that differ from these cases. For many inland locations at high elevation, the water level is missing unless the surge is extremely high. Storms where a node is wetted do not always provide useful information for storms where the water level is lower.

Our imputation approach can be interpreted as compositing the original surge function with the imputation scheme. Thus the imputation algorithm can be chosen without regard for the statistical approach we leverage for emulation. The imputation should not change the wet nodes' peak surges. Also, the imputation approach should be continuous, or nearly so, in the following sense: if a node is almost wetted, the imputed numeric value is very close to when the node is nearly dry. Let $y_i = f(x_i, t_i)$ be the vector of imputed data for storm generated with (x_i, t_i) and $y_{ij} = f_j(x_i, t_i)$ when the jth node is wetted. Our chosen within-surface imputation approach is inverse distance weighting interpolation (Shepard (1968)). Let z_{ij} denote the imputed value. Thus, if y_{ij} is wet, then $z_{ij} = y_{ij}$, and if y_{ij} is dry,

$$z_{ik} = \frac{\sum_{j \text{ where } f_j(x_i,t_i) \text{ is wet}} y_{ij} \|s_j - s_k\|^{-1}}{\sum_{j \text{ where } f_j(x_i,t_i) \text{ is wet}} \|s_j - s_k\|^{-1}},$$

where $\|s - s'\|$ is the distance between the geographic locations s and s'. The only modification was that the surge level was capped at the elevation of the node where a node is dry. This modification ensures that the presence of surge from the ADCIRC model matches the event when the imputed water level is larger than the elevation. That is, when a node is dry, the water level is less than or equal to ground elevation. When a node is wet, the water level is greater than ground elevation. The rationale for inverse distance weighting is that it is computationally fast and can impute an entire mesh's missing values in under a minute for a single surface. More discussions on this point are in the last section of this article.

3.2. *A fast, log-transformed Gaussian process model.* After imputation filled in the missing values, we opted for an approach termed as "partially parallel" by Gu and Berger (2016). This resulted in significant computational savings. Initially in our study, a Gaussian process emulator failed in providing accurate prediction for simulated surges from ADCIRC. The solution to the problem was to simply transform the data using a monotonic transform. While the full Box–Cox family was available (Sakia (1992)), our approach was to simply check if the two most popular transforms, square-root and log, would work, similar to the strategy of Johnson et al. (2018). We found that a log-transform was superior. While this fix alone is not particularly novel (Cressie (1990)), it is worth acknowledging the degree to which it resolved the problems in our case. The log-transformed Gaussian process model resulted in reasonable predictive models for the marginal distribution at each node.

To simplify notation in this section, let $g(\xi)$ be the imputed value at $\xi = (x, t)$, meaning that the input to the model can be represented by $\xi_1, \dots, \xi_n$. Our statistical model is

$$\log\bigl(g_j(\cdot)\bigr) \overset{\text{iid}}{\sim} \text{Gaussian process}\bigl(\mu_j, \sigma_j^2 r(\cdot, \cdot; \phi)\bigr) \quad \text{for } j = 1, \dots, M, \tag{3.1}$$

where μ_j and σ_j^2 are constants associated with the jth node and r is a correlation function depending on correlation parameters ϕ. In terms of the ADCIRC model, the coastal geography implies that nodes can differ in terms of the average surge and the variation in the surge; thus, it is important that each node be endowed with its own mean and variance parameters, as in (3.1). However, for the correlation function it makes sense to choose the same structure and parameters for all nodes to ease the computational burden. If we did not do this, we would have to invert hundreds of thousands of correlation matrices which by round 5 grow to size 120 × 120. If we did this inversion 100 times during optimization, this would expend approximately four times our entire overhead computational budget on matrix inversion alone. By using the same correlation structure and parameter ϕ, the speedup is massive and the fitting process takes less than *five* minutes on any round. We opted for a custom nonseparable power exponential correlation function; the separable variant was used by Welch et al. (1992). Supplement B describes the specific correlation structure used for our case study alongside a brief justification.

Let $Z = \log((z_1, \dots, z_n)^{\mathsf{T}})$ be an $n \times M$ matrix computed elementwise. The symbol T represents transpose. Following known derivations, our maximum likelihood vectors are

$$\hat{\mu}(\phi) = \begin{pmatrix} \hat{\mu}_1(\phi) \\ \vdots \\ \hat{\mu}_M(\phi) \end{pmatrix} = \frac{Z^{\mathsf{T}} R(\phi)^{-1} e}{e^{\mathsf{T}} R(\phi)^{-1} e} \quad \text{and}$$

$$\hat{\sigma}^2(\phi) = \begin{pmatrix} \hat{\sigma}_1^2(\phi) \\ \vdots \\ \hat{\sigma}_M^2(\phi) \end{pmatrix} = \frac{1}{n} \operatorname{diag}\bigl(\bigl(Z - e\hat{\mu}(\phi)^{\mathsf{T}}\bigr)^{\mathsf{T}} R(\phi)^{-1} \bigl(Z - e\hat{\mu}(\phi)^{\mathsf{T}}\bigr)\bigr),$$

where e is a length n vector of 1s, $R(\phi)$ is the n by n matrix

$$R(\phi) = \begin{pmatrix} r(\xi_1, \xi_1; \phi) & \cdots & r(\xi_1, \xi_n; \phi) \\ \vdots & \dots & \vdots \\ r(\xi_n, \xi_1; \phi) & \cdots & r(\xi_n, \xi_n; \phi) \end{pmatrix}$$

and the function diag returns a column vector of the diagonal elements of the matrix. The maximum likelihood estimate for the correlation parameters is then found by

$$\hat{\phi} = \operatorname*{argmin}_{\phi} \log\bigl(\text{determinant}\bigl(R(\phi)\bigr)\bigr) + \frac{n}{M} \sum_{j=1}^{M} \log \hat{\sigma}_j^2(\phi).$$

Supplement B gives additional computational details on how we found $\hat{\phi}$ in our case study.

We use a predictive distribution for the log of the surge at a new set of storm characteristics ξ, $\log(g(\xi))$, as a vector of normal distributions with vector of means

$$\hat{m}(\xi) = \hat{\mu}(\hat{\phi}) + \left(Z - e\hat{\mu}(\phi)^{\mathsf{T}}\right)^{\mathsf{T}} R(\hat{\phi})^{-1} \begin{pmatrix} r(\xi, \xi_1; \hat{\phi}) \\ \vdots \\ r(\xi, \xi_n; \hat{\phi}) \end{pmatrix}$$

and a vector of variances

$$\hat{v}(\xi) = \left(r(\xi, \xi; \hat{\phi}) - \left(r(\xi, \xi_1; \hat{\phi}) \quad \cdots \quad r(\xi, \xi_n; \hat{\phi}) \right) R(\hat{\phi})^{-1} \begin{pmatrix} r(\xi, \xi_1; \hat{\phi}) \\ \vdots \\ r(\xi, \xi_n; \hat{\phi}) \end{pmatrix} \right) \hat{\sigma}^2(\hat{\phi}).$$

These two functions, $\hat{m}$ and $\hat{v}$, are the conditional mean and variance of $\log(g(\xi))$ based on all available data and the maximum likelihood estimates for the Gaussian process parameters. This predictive distribution ignores the estimation uncertainty in the parameters, but this is a fast approximation compared to other popular, but slower, methods like Markov chain Monte Carlo. Given a landfall storm characteristic vector ξ, the mean of the predictive distribution of $g(\xi)$, often directly referred to as the emulator, is given as follows by a property of the log-normal distribution:

$$\hat{g}(\xi) = \exp\bigl(\hat{m}(\xi) + \hat{v}(\xi)/2\bigr).$$

The function $\exp(\hat{m}(\xi))$ represents our predictive median. Interestingly, $\hat{g}(\xi)$ is inflated beyond the median at uncertain storm characteristics because of the log-transform used in our statistical model.

3.3. *Forecasting technique.* Our forecast distribution for the imputed maximum water level at node j at time t is then

$$\text{Log-Normal}\bigl(\hat{m}_j(\tilde{x}, \tilde{\tau}), \hat{v}_j(\tilde{x}, \tilde{\tau})\bigr), (\tilde{x}, \tilde{\tau}) \sim \Pi(t), \tag{3.2}$$

where $\Pi(t)$ is the forecast distribution at time t. This implies the forecast for $f(x, t)$, which should have dry values indicated, should be a truncated log-normal distribution, where the value is marked dry when the prediction is below that node's elevation. This can be quickly simulated without rerunning the ARCIRC model. The current NHC surge forecast provides the 90% quantile, so we used $N = 3500$ samples from the forecast distribution at time t to ensure an adequate estimation of this quantile via a large enough sample size. We label our draws of landfall characteristics as $(\tilde{x}_1, \tilde{\tau}_1), \dots, (\tilde{x}_N, \tilde{\tau}_N)$. It is important to note that these are not the same as $(x_1, t_1), \dots, (x_n, t_n)$ which are run through the ADCIRC simulator. While $(\tilde{x}_1, \tilde{\tau}_1), \dots, (\tilde{x}_N, \tilde{\tau}_N)$ are chosen via random sampling, $(x_1, t_1), \dots, (x_n, t_n)$ are more carefully chosen, as described in the next section. Moreover, $t_1, \dots, t_n$ are forecast times, whereas $\tilde{\tau}_1, \dots, \tilde{\tau}_N$ are storm landfall times.

The usage of forecasted and actual landfall time is a unique and important component in prediction that departs from existing studies of time and computer emulation (Conti et al. (2009), Mak et al. (2018)). Forecast time, not landfall time, is the major driver of discrepancy in surge model output. If ignored, this could break our Gaussian process inference because a storm with the same landfall characteristics could return two different outputs. While we could treat each response as having random error, sometimes refereed to as a nugget (Gramacy and Lee (2012)), we opted for a procedure that reflects the actual generation of the discrepancy. The previous data include which forecast time the surge measurement is based, $t_1, \dots, t_n$. However, the predictive distribution is based on the random landfall occurrence time $\tilde{\tau}$. When we are far from landfall, thus $\tilde{\tau}$ is far from t_n, the model output is less correlated with predictions, and the predictive distribution gets more diffuse. As forecasted landfall $\tilde{\tau}$ draws near to t_n, the model outputs and predictions become more correlated. This mirrors the intuitive effect of nonlandfall information that is borrowed from NHC forecast: far from landfall, many things can aggregate for large effects, and we should have a broader predictive distribution; closer to landfall, these extraneous forecast properties are closely aligned with the actual storm and should have a relatively minor effect. This is automatically tuned during maximum likelihood estimation by estimating the correlation parameter corresponding to t.

4. Sequential sampling using optimal designs. In our case study the emulator using our initial data from 30 runs was not sufficient to end the statistical learning process. This section describes a new sequential sampling strategy for choosing storm landfall characteristics using optimal designed experiments. We will select a batch of experiments of size q to be run through the ADCIRC simulator in parallel. This section discusses how the log-transform used in our statistical model impacts the typical optimal design criterion. Aside from this transformation issue, our problem differs from a static problem where the goal is to improve the emulator in a fixed environment, as done in Loeppky, Moore and Williams (2010) and Vernon, Goldstein and Bower (2014). Two important changes happen each round as 12 hours of wall-clock time passes and the storm marches toward landfall. First, the distribution of the landfall storm characteristics becomes more concentrated around the unknown true storm's landfall characteristics, reducing the need to improve the emulator globally. Second, the simulation model changes as a function of the landfall characteristics each round, as detailed in Section 1. These two changes affect the resulting designs and overall performance but not the optimal design criterion we introduce.

We presume our statistical model is fitted using the previous data. Without loss of generality, we refer to these as $(x_1, t_1) \ldots, (x_n, t_n)$, acknowledging that n increases after each round of data collection. We then invert the fitted model by optimizing a criterion to find the best possible selection of q new landfall characteristics, between 10 and 30, to investigate. One widely used model-based criterion for computer experiments comes from Sacks et al. (1989). This takes a Gaussian process model, with no transform applied, and evaluates designs based on the integrated mean squared error (IMSE), given by

$$\text{IMSE} = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} \hat{v}_j^{\text{prop}}(\tilde{x}_i, \tilde{\tau}_i). \tag{4.1}$$

The word "integrated" is used here in place of the more obvious "average" to match historical notation. The variance $\hat{v}_j^{\text{prop}}$ is the variance of $\log(f_j(\tilde{x}, \tilde{\tau}))$ conditional on the data up to this point, the parameter estimates at this time as well as on both the choice of the new q points and the previously used storms. For us, it is given by

$$\hat{v}_j^{\text{prop}}(x, \tau) = \hat{\sigma}_j^2(\hat{\phi}) \left(r\big((x, \tau), (x, \tau); \hat{\phi}\big) - h(x, \tau)^{\mathsf{T}} \begin{pmatrix} R & H \\ H^{\mathsf{T}} & G \end{pmatrix}^{-1} h(x, \tau) \right),$$

where

$$h(x, \tau) = \begin{pmatrix} r\big((x, \tau), (x_1, t_1); \hat{\phi}\big) \\ \vdots \\ r\big((x, \tau), (x_n, t_n); \hat{\phi}\big) \\ r\big((x, \tau), (x_{n+1}, t); \hat{\phi}\big) \\ \vdots \\ r\big((x, \tau), (x_{n+q}, t); \hat{\phi}\big) \end{pmatrix},$$

$$H = \begin{pmatrix} r\big((x_1,t_1),(x_{n+1},t);\hat{\phi}\big) & \cdots & r\big((x_1,t_1),(x_{n+q},t);\hat{\phi}\big) \\ & \vdots & \\ r\big((x_n,t_n),(x_{n+1},t);\hat{\phi}\big) & \cdots & r\big((x_n,t_n),(x_{n+q},t);\hat{\phi}\big) \end{pmatrix}, \quad \text{and}$$

$$G = \begin{pmatrix} r\big((x_{n+1},t),(x_{n+1},t);\hat{\phi}\big) & \cdots & r\big((x_{n+1},t),(x_{n+q},t);\hat{\phi}\big) \\ & \vdots & \\ r\big((x_{n+q},t),(x_{n+1},t);\hat{\phi}\big) & \cdots & r\big((x_{n+q},t)(x_{n+q},t);\hat{\phi}\big) \end{pmatrix}.$$

These formulae leverage $t_{n+1} = \cdots = t_{n+q} = t$ since we are selecting q storms at time t. Minimizing the IMSE criterion typically avoids previously used storms, seeking to fill gaps where needed in the input space. Minimizing the IMSE criterion also considers the forecast distribution $\Pi(t)$ through the draws $(\tilde{x}_1, \tilde{\tau}_1), \dots, (\tilde{x}_N, \tilde{\tau}_N)$. Minimizing the criterion will place experimental landfall characteristics near the draws from the forecast distribution to reduce the variance in those locations. Thus, minimizing IMSE will automatically both fill gaps in the existing data while adhering to the current forecast distribution.

For us, the traditional IMSE in (4.1) is not representative of the predictive accuracy of our emulator. IMSE is measuring the error in the log-transformed space, $\hat{m}(\tilde{x}, \tilde{\tau}) - \log(g(\tilde{x}, \tilde{\tau}))$, but our concern is the actual error in terms of storm surge, $\hat{g}(\tilde{x}, \tilde{\tau}) - g(\tilde{x}, \tilde{\tau})$. A more nuanced criterion for our case with a log-transformed response would incorporate the results, such as those presented in Section 3.2.2 of Cressie (1993). This new criterion is termed an exponential integrated mean square error (E-IMSE) criterion and is defined as

$$\text{(4.2)} \qquad \text{E-IMSE} = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} w_j(\tilde{x}_i, \tilde{\tau}_i)\big(1 - \exp(-\hat{v}_j^{\text{prop}}(\tilde{x}_i, \tilde{\tau}_i))\big),$$

where $w_j(\cdot, \cdot)$ is a weight defined as $w_j(\tilde{x}, \tilde{\tau}) = \exp(2\hat{m}_j(\tilde{x}, \tilde{\tau}) + 2\hat{v}_j(\tilde{x}, \tilde{\tau}))$. Supplement B contains more details on the derivation. The weight depends only on data collected prior to the current data collection period. The second term in our formula depends on both the prior data and the new experimental data. We note that, although our criterion averages across all nodes here, it effectively gives more weight to high variance or a high mean nodes because these will have larger values of $w_j(\tilde{x}, \tilde{\tau})$. This focuses attention to nodes with the largest surges.

Comparing the IMSE in (4.1) and the proposed E-IMSE in (4.2), there are two key differences. First, $1 - \exp(-\hat{v}_j^{\text{prop}}(\tilde{x}_i, \tilde{\tau}_i))$ is not the same as $\hat{v}_j^{\text{prop}}(\tilde{x}_i, \tilde{\tau}_i)$. However, at values of $\hat{v}_j^{\text{prop}}(\tilde{x}_{ti})$ close to 0, where we hope our emulator lies, the first order behavior of these functions is the same. The major difference then lies in the weight function. The IMSE criterion does not give varying weight to different areas of the x space. The E-IMSE criterion, on the contrary, gives an increased weight to those xs with large mean values. In our context, minimizing E-IMSE naturally selects storms with higher predicted surge in addition to filling gaps in the data and adhering to the forecast distribution.

To illustrate the real effect of our criterion, Figure 2 shows the results for round 1 of the Michael study using IMSE and E-IMSE, each selecting 20 new storms by minimizing the respective criterion. For illustrative purposes, only the landfall characteristics that most affect the storm strength are modified: the maximum wind speed (MWS) and the radius of the 34 kt isotach (R34). Increasing either MWS or R34 gives an increased storm strength that, in turn, induces larger surge levels. The resulting designs in Figure 2 show a clear trend that IMSE emphasizes weaker storms compared to E-IMSE. Focusing on storms with larger, more catastrophic surge is a desired behavior when examining storm surge. This feature naturally occurs in importance sampling of storm surge (Dawson and Hall (2006)). Here, minimizing E-IMSE replicates this behavior by acknowledging that a reasonable statistical model has a

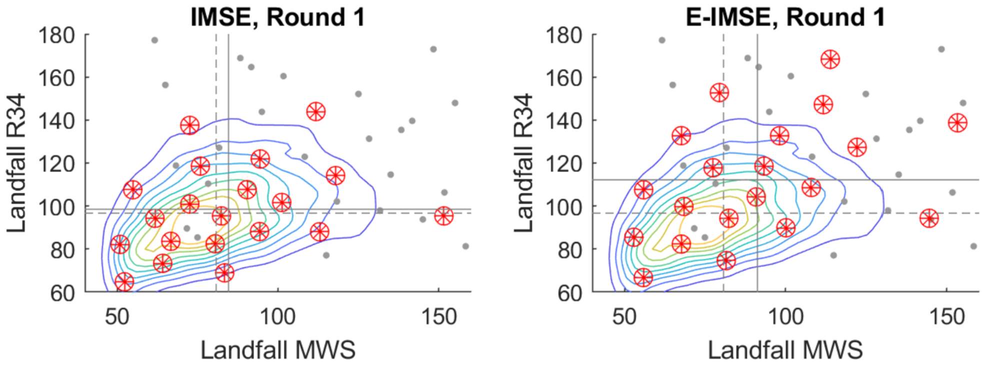

FIG. 2. *Illustration of the differences in optimal designs decided by criterion IMSE* (*left panel*) *and criterion E-IMSE* (*right panel*). *The small dots are the storm characteristics used in the round* 0 *experiment projected onto the dimensions MWS and R*34. *The circles with stars represent the* 20 *new chosen storm characteristics. The background contours represent the* 10%, 20%, . . . , 90% *highest density regions of the forecast as of round* 1. *The dashed lines are the marginal means for each landfall characteristic. The solid lines are the marginal sample averages of the selected landfall characteristics.*

structure where larger mean values also have higher variances. When we minimize E-IMSE, we place storms near regions with larger surges which have high-variance.

Computationally, we rely on a randomized search algorithm. Random search is common in design of experiments in a continuous domain as the criterion's surface is highly nonconvex (see, e.g., Gramacy and Lee (2009)). Our algorithm had to fit under the constraint of one hour of overhead computation. The whole procedure takes roughly 15 minutes on a desktop computer. A more precise description of the approach is found in Supplement B, and the exact code is included in Supplement D (Plumlee et al. (2021)). The unknown parameters that implicitly inform this function are set to their maximum likelihood values from the most recent optimization.

The resulting designs are presented in Figure 3. There are a few interesting features of Michael that become apparent. First, because of the complex coastline in this region, our round 0 approach resulted in off-coast simulations that did not align with the remainder of the selected storms, as demonstrated in the LAT/LONG plot of the samples. Second, the distributions indicate that Michael was originally forecast to be quite weak but rapidly strengthened closer to landfall. This poor forecast makes Michael a challenging test case. Improving the properties of the hurricane forecast distribution provided by the NHC is considered outside the scope of this article. Lastly, we note that as Michael gets closer to landfall, both forecast distributions and designed experiment runs get more concentrated. This means that the prediction region gets smaller, and thus the emulation problem gets easier closer to landfall.

5. Design and emulation performance analysis. This section describes the performance analysis of both the emulation strategy from Section 3 and of the designed experiment strategy from Section 4. One question is the effectiveness of the partial parallel approach, compared to the more popular decomposition-type approach, where the surge response is dimension reduced through a tool like principal component analysis (PCA). The reduced dimension components are then modeled with individual Gaussian processes (potentially with different parameters). Examples of this type of approach are Higdon et al. (2008) and Chang et al. (2014), and the exact method is described in Supplement B.

To assess emulator accuracy, we ran through extra storms with the same counts as designed storms for the last three efforts of data collection (see Table 1). These storms were drawn from the forecast distribution with random landfall characteristics. This block of 40 storms

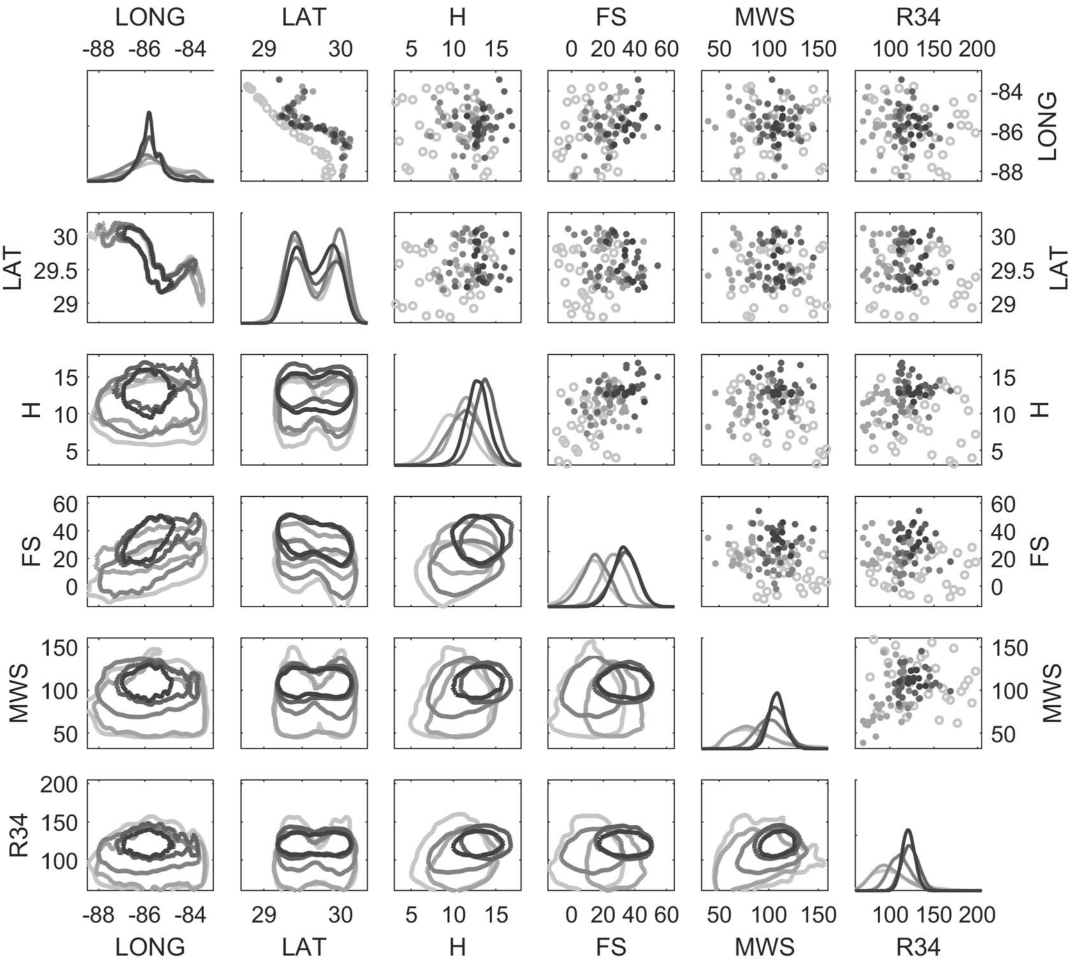

FIG. 3. *Diagram of the predictive forecast distributions of the six landfall characteristics alongside the storms chosen via the mechanism and described in Section* 4. *The bottom left panels show the* 90% *high-density region contours for the forecast distribution for forecast periods* 60,48,36,24 *and* 12 *from landfall*, *where darker lines imply closer to landfall. The diagonal panels show the one-dimensional marginal forecast densities for each landfall characteristic where darker lines imply a later round. The upper right panels show the selected design storms. The open circles are the round* 0 *storms. The smaller dots are the storms chosen by E-IMSE*, *where darker dots imply a later round.*

can be used as a holdout set with the 120 other storms being the training set. These random storm runs also permit a formal investigation of the relative value of designed experiments vs. randomly chosen storms. For the randomly chosen storms we can treat them as a (poorly) designed experiment and replace the original designed runs with their random counterparts in rounds 3, 4 and 5. If we leave out a single storm and build a new emulator, this would recreate an emulator built by choosing 39 storms at random in place of our carefully chosen 40 storms. This yields a benchmark to compare our experimental design approach, but understand there is also a small sample size discrepancy (39 storms in the random approach vs. 40 storms in our designed experiment approach).

Five different measures of accuracy are reported for completeness of understanding. In our comparison, only nodes with ground elevation greater than one meter are investigated. This focuses our analysis metrics closer to habitable areas that have a notable surge signal. For this study we are considering our ability to predict the imputed values, thus evaluating the emulation, not the quality, of the imputation. The study in the next section will evaluate the

TABLE 2

The accuracy of the emulator measures as described in Section 5. The number in parentheses is the p-value of a one-sided paired t-test with the null that the far left column has a larger mean

	E-IMSE Minimizing Storms		Randomly Selected Storms	
	Partial Parallel	PCA Based	Partial Parallel	PCA Based
Root mean squared error	0.1234 (–)	0.1245 (0.350)	0.1684 (0.007)	0.1567 (0.018)
Mean absolute error	0.0732 (–)	0.0780 (0.007)	0.0945 (0.009)	0.0930 (0.008)
Mean Dawid-Sebastiani score	−4.0516 (–)	−3.5220 (0.000)	−3.6650 (0.003)	−3.1852 (0.000)
95 % coverage	0.9086	0.8558	0.8899	0.8510
95 % interval score	0.5713 (–)	0.6827 (0.001)	0.8453 (0.015)	0.9475 (0.005)

overall performance that accounts for both imputation and emulation. The four metrics are the typical root mean squared error, mean absolute error, which is often used in meteorological forecasting (Willmott and Matsuura (2005)), the Dawid-Sebastiani score, which evaluates adequacy of both the forecast variance and the mean (Gneiting and Raftery (2007)), and the 95% coverage rate and interval score. More details of these evaluations are provided in the Appendix.

The results are presented in Table 2, where the proposed storms do better than randomly selected storms and the partial parallel approach does better than a PCA-based approach for this case. We conduct hypothesis testing using a t-test with a sample size of 40 and find significant results for chosen storms against randomly selected storms at the 0.05 level. The partial parallel approach does better than the PCA-based approach in general, often at a statistically significant level. Our explanation is that the PCA-based approach tends to focus the weights on high surge areas, meaning that it tends to mischaracterize variation in low surge areas. This is perhaps because the first few principal components can be estimated based on the sample covariance (a similar discussion can be found in Sansó and Forest (2009)). That our surge response (at least a log-transformed version) can reasonably be modeled with a separable covariance function implies that the partial parallel can get near optimal predictions without modelling the internode covariance structure which agrees with the findings of (Gu and Berger (2016)). In Table 2, both methods and designs yield confidence intervals that give slight undercoverage which can be interpreted as overconfidence. One reviewer for this article suggested this result could be due to inadequate smoothness of the covariance function or using plug-in maximum likelihood estimators of the covariance parameters. We have no evidence for a single cause but concur these are plausible and add three more potential causes: it is at least slightly incorrect to use the same covariance parameters across all nodes in the partial parallel approach; the log-transform only yields approximate Gaussian behavior, and/or this could be due to random chance since the confidence intervals are evaluated on correlated observations with a relatively small sample size (40 storms). More theoretical details are given in Supplement B.

6. Hindcast prediction performance analysis. This section discusses the efficacy of the approach in terms of accurate surge forecasts, not just accurate emulation. We also note that a comparison to P-Surge using a small number of real surge readings is summarized in Supplement C (Plumlee et al. (2021)) where the conclusions are similar. Our gold standard hindcast uses Michael’s operational best track data from the NHC because these data come from the same source as the forecasts, providing the most level comparison. The results are, overall, positive. This was somewhat surprising because the accuracy of the surge forecast can be corrupted by several factors outside of the statistical approach described in this ar-

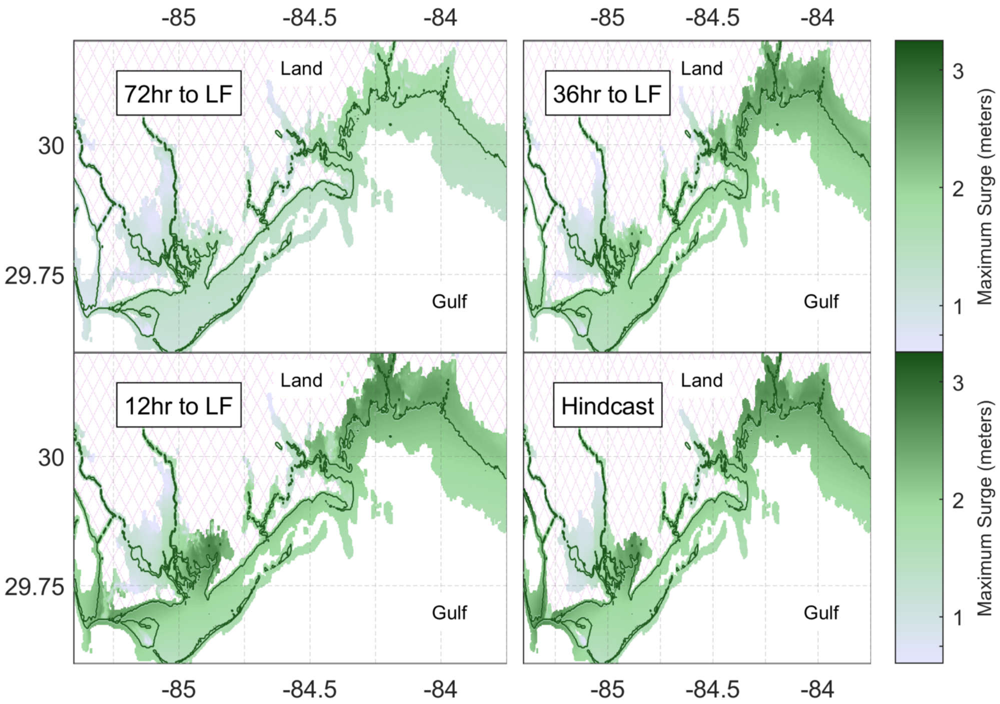

FIG. 4. *The ADCIRC hindcast peak surge (bottom right panel) and predictive median peak surge (other three panels) using the predictive distribution in (3.2) at the marked latitude and longitude. Color is shown only if the storm surge is above the node's ground elevation. Color is also shown only for nodes whose ground elevation is above* -4 *meters. The black line is the coastline.*

ticle. These include the NHC forecast accuracy, the forecast distribution and the algorithm that fills in missing information in the forecast. While the latter two have been designed and tested to ensure reasonableness, the first one remains outside of our control. Figure 4 shows the geographic layout near the eastern Florida Panhandle of our forecasted surge vs. the gold standard hindcast results. Nodes shown are in our M emulation sites, discussed in Section 2, so these figures include some surges over the ocean as well as land. The bend in the coastline on the right portion of the map represents a region exceptionally sensitive to storm surge due to funneling effects mentioned in the beginning of this article. One important observation from this plot is, despite the node independence assumption used for our emulator, the forecast surfaces remain smooth and reasonably behaved. This agrees with the observations from Gu and Berger (2016) that the predictive mean of the surface structure can be preserved from the emulator despite assuming node independence. As discussed in Section 1.4, the NHC underforecasted Michael's intensity and this propagates into our median forecast which is low throughout the storm, especially over 48 hours before landfall. Once the NHC's forecasted storm intensity increased, our median forecasted surge responded with a corresponding increase. Overall, this prediction appears to fall in line with reasonable surge forecasts.

To quantitatively understand the overall forecasting performance, we would like to determine not only how well we predict at wet nodes but also how well we predict dry nodes. Thus, we introduce two quantitative measures. Define a prediction that at j contains either a numerical water level if it is wet or the label "`dry`" if it is dry, call it h_j. Let the true hindcast run at j be labeled h_j^* have the same structure as h_j. The elevation at a node be labeled e_j.

Then, let

$$\text{Surge Score} = \frac{1}{\text{number of nodes}} \sum_{j \in \text{nodes}} \begin{cases} |h_j - h_j^*| & \text{if } h_j \neq \texttt{dry}, h_j^* \neq \texttt{dry}, \\ |h_j^* - e_j| & \text{if } h_j = \texttt{dry}, h_j^* \neq \texttt{dry}, \\ |h_j - e_j| & \text{if } h_j \neq \texttt{dry}, h_j^* = \texttt{dry}, \\ 0 & \text{if } h_j = \texttt{dry}, h_j^* = \texttt{dry}, \end{cases}$$

and

$$\text{Misclassification Rate} = \frac{1}{\text{number of nodes}} \sum_{j \in \text{nodes}} \begin{cases} 0 & \text{if } h_j \neq \texttt{dry}, h_j^* \neq \texttt{dry}, \\ 1 & \text{if } h_j = \texttt{dry}, h_j^* \neq \texttt{dry}, \\ 1 & \text{if } h_j \neq \texttt{dry}, h_j^* = \texttt{dry}, \\ 0 & \text{if } h_j = \texttt{dry}, h_j^* = \texttt{dry}. \end{cases}$$

To our understanding, the definition of Surge Score is unique to this article but was needed to study both the magnitude of error in our censored data environment and errors in wet/dry classification. Otherwise, for example, one could create a better numerical prediction by reporting the elevation at a node.

We also compared our method to the probabilistic storm surge forecasts from P-Surge, the NHC's forecasting surge forecasting tool, as a benchmark. We acknowledge that we are unfairly using our ADCIRC model as the gold standard, but this will give us an idea if the overall procedure meets the current practice. Exact comparisons of the potential model error between P-Surge's SLOSH and our ADCIRC are difficult. As mentioned in the Introduction, P-Surge's SLOSH has a coarser mesh compared to our ADCIRC. However, P-Surge leverages SLOSH simulations from multiple overlapping meshes of differing resolutions and extents; thus, its exact resolution cannot be ascertained. When examining P-Surge's data, we found P-Surge may not report a median and thus we performed analysis on nodes with elevation above *zero* meter where P-Surge reported some value during the comparison. We also only considered nodes that, at some point during our simulation, were wetted to remove trivial nodes.

The quantitative measures of our forecast, using the median as the prediction, are presented in Figure 5. The results are positive both in magnitude and trend over the course of our study. It should be noted that overland surge prediction at elevations above one meter is particularly challenging because these areas experience the largest variance in surge. Moreover, our imputation approach has significant impact at these nodes because they are often unwetted. Closer to landfall, 36 hours out, our accuracy measures get better as the NHC forecast improves. The

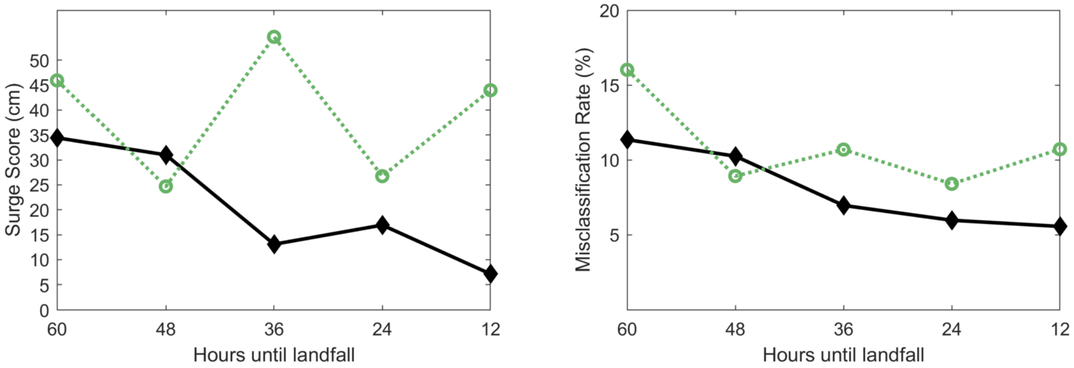

FIG. 5. *Accuracy quantification of the median of the predictive distribution for both the proposed approach* (*solid lines*) *and NHC forecast mechanism P-Surge* (*dashed lines*) *as described in Section* 6.

results are promising in favor of the proposed approach which better captures the anticipated surge earlier compared to P-Surge. An observation outside these figures is neither negative nor positive: the proposed surge forecasts are more consistent than P-Surge. The proposed approach appears less sensitive to forecast-to-forecast fluctuations, and one theory is that data reuse in our method induces some hysteresis. The medians of our predictive distributions are persistently below the actual measurements. This is expected due to the underprediction of Michael's intensity before landfall. When Michael strengthened, this resulted in rising surge at all sensors. As discussed in the previous subsection, this effectively is a sample size of one, a single storm. The lack of convergence of P-Surge in Figure 5 is at least partially due to the fact that our reference solution came from an ADCIRC model, and P-Surge employs multiple overlapping SLOSH model grids. We also suspect P-Surge's performance oscillates because of how it handles the irregular timing of NHC forecast information. Nonetheless, that our results have similar accuracy is encouraging.

7. Conclusions.

Storm surge forecasting for tropical cyclones is challenging because of the short turnaround time, substantial computational requirements and large forecast uncertainties. This article explains how, using a limited number of high-fidelity model runs with new and existing computer experiment technology, we can forecast storm-surge hazard probabilities. Our results indicate that using an emulator with a targeted experimental design is an effective strategy to leverage limited sampling of a high-fidelity computer model of storm surge. We have completed a case study on Hurricane Michael (2018) which required only 40 model runs per day to get adequate predictive performance. This mitigates the computational cost of running the high-fidelity model with only a small amount of overhead statistical computation. Our solution performed as well as or better than the primary data product currently available which uses more runs of a lower-fidelity model. This existing data product has over a decade of supporting development underneath it. The proposed method's performance, relative to an operational product, suggests that using a higher-fidelity model of storm surge with statistical tools could improve surge forecasting.

There is extensive potential future research for both forecasting of storm surge and general statistics. In terms of forecasting storm surge, there are potential emulator improvements, better storm forecast distributions and consideration of other storm parameterizations. In addition, it might be possible to reduce the current data-model discrepancies for ADCIRC through model calibration efforts (e.g., Chang et al. (2016), Gu and Wang (2018), Gu, Xie and Wang (2018), Plumlee (2017), Tuo and Wu (2015), Chang et al. (2019)). Another direction on the statistical side is the development of more powerful methods to handle the missing surge values at unwetted nodes. For in-surface imputation, we did try using Nearest Neighbor Gaussian process interpolation (Datta et al. (2016)) for imputation, and we saw a slight decrease in overall accuracy in addition to taking over 30 times longer. Other potential tools, like laGP (Gramacy (2016)), were too slow to be considered viable options. We imagine that significant improvement in the imputation of unwetted nodes requires a specialized method that uses some physical understanding of storm surge. Another possibility is a fully Bayesian approach that enables emulation and imputation based on a joint posterior distribution for all the unknowns (i.e., emulation parameters and missing values). Developing such a scheme without causing excessive computational burden poses a significant statistical challenge and merits more investigation. Some Markovian model that induces sparsity in the covariance structure (e.g., Lindgren, Rue and Lindström (2011)) might be useful for formulating a computationally feasible approach.

On the subject of designing sequential experiments, we have shown that a transformation of the response leads to a new experimental design criterion with desirable properties. This could lead to similar approaches with different transformations or generalizations to families of transformations. Our experimental design criterion is based on the predictive variance which agrees with our setting where the magnitude of the prediction error is critical. Reducing the predictive variance from 100 cm^2 to 10 cm^2 at one location provides more benefit than reducing the predictive variance from 10 cm^2 to 1 cm^2 at another location. Currin et al. (1991) popularized an alternative, entropy-based criterion (Lindley (1956)). Beck and Guillas (2016) demonstrated an iterative algorithm, like the one we use (Supplement B), was an effective approach but it may be slower than an algorithm that uses an entropy-based criterion. We selected a design algorithm that was sufficiently fast to be employed in our setting when the main computational cost was running the computer model (see Table 1). Perhaps a faster experimental design algorithm would be advantageous in a setting with larger sample sizes or a cheaper computer model.

APPENDIX: DETAILS ON QUANTITATIVE PREDICTION COMPARISONS

Let mean_j, med_j and var_j be the median and variance of the predictive distribution at node j. Let U_j and L_j be the 97.5% quantile and 2.5% quantile of the predictive distribution. Let a_j be the imputed peak surge value held out for testing. Root mean squared error is given by $\sqrt{M^{-1}\sum_{j=1}^{M}(a_j - \mathrm{mean}_j)^2}$. Mean absolute error is given by $M^{-1}\sum_{j=1}^{M}|a_j - \mathrm{med}_j|$. The Dawid–Sebastiani score is given by

$$M^{-1}\sum_{j=1}^{M}\left(\frac{(a_j - \mathrm{mean}_j)^2}{\mathrm{var}_j} + \log(\mathrm{var}_j)\right).$$

The 95% coverage rate is $M^{-1}\sum_{j=1}^{M} I(\mathrm{L}_j \le a_j \le \mathrm{U}_j)$ (where I is the indicator function), and the 95% interval score (Gneiting and Raftery (2007)) is given by

$$M^{-1}\sum_{j=1}^{M}\left(\mathrm{U}_j - \mathrm{L}_j + \frac{1}{0.025}(a_j - \mathrm{U}_j)I(a_j > \mathrm{U}_j) + \frac{1}{0.025}(\mathrm{L}_j - a_j)I(a_j < \mathrm{L}_j)\right).$$

Acknowledgments. This material was based upon work partially supported by the National Science Foundation under Grant Award Number DMS-1638521 (to the Statistical and Applied Mathematical Sciences Institute), Grant Award Number DMS-1953111, Grant Award Number ACI-1339723 and by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-ND0001-01.

The authors are partially supported by the Statistical and Applied Mathematical Sciences Institute.

SUPPLEMENTARY MATERIAL

Details on forecasting (DOI: 10.1214/20-AOAS1398SUPPA; .pdf). This contains a description of the conversion of NHC forecasts to forecast distributions as well as the updating process for NHC forecasts given different landfall characteristics.

Additional statistical and algorithmic details (DOI: 10.1214/20-AOAS1398SUPPB; .pdf). This contains a description of some additional statistical and algorithmic details that could not fit in the main article due to length constraints.

More predictive performance analysis (DOI: 10.1214/20-AOAS1398SUPPC; .pdf). This contains a description of a more prediction performance analysis and a comparison of our approach to P-Surge using surge readings from water level meters.

Exemplar code (DOI: 10.1214/20-AOAS1398SUPPD; .zip). This contains Matlab code and data that illustrate the predictive algorithm at a subset of nodes.

REFERENCES

ALTAF, M., BUTLER, T., MAYO, T., LUO, X., DAWSON, C., HEEMINK, A. and HOTEIT, I. (2014). A comparison of ensemble Kalman filters for storm surge assimilation. *Mon. Weather Rev.* **142** 2899–2914.

ASHER, T. G., LUETTICH, R. A. JR., FLEMING, J. G. and BLANTON, B. O. (2019). Low frequency water level correction in storm surge models using data assimilation. *Ocean Model.* **144** 101483.

BECK, J. and GUILLAS, S. (2016). Sequential design with mutual information for computer experiments (MICE): Emulation of a tsunami model. *SIAM/ASA J. Uncertain. Quantificat.* **4** 739–766. MR3507556 https://doi.org/10.1137/140989613

BECT, J., GINSBOURGER, D., LI, L., PICHENY, V. and VAZQUEZ, E. (2012). Sequential design of computer experiments for the estimation of a probability of failure. *Stat. Comput.* **22** 773–793. MR2909621 https://doi.org/10.1007/s11222-011-9241-4

BEVEN, J. L., BERG, R. and HAGEN, A. (2019). Hurricane Michael Tropical Cyclone Report No. AL142018 NOAA/National Hurricane Center.

BILSKIE, M. V., HAGEN, S. C. and MEDEIROS, S. C. (2019). Unstructured finite element mesh decimation for real-time hurricane storm surge forecasting. *Coastal Eng.* **156** 103622.

BILSKIE, M. V., COGGIN, D., HAGEN, S. C. and MEDEIROS, S. C. (2015). Terrain-driven unstructured mesh development through semi-automatic vertical feature extraction. *Adv. Water Resour.* **86** 102–118.

BILSKIE, M. V., HAGEN, S. C., MEDEIROS, S. C., COX, A. T., SALISBURY, M. and COGGIN, D. (2016). Data and numerical analysis of astronomic tides, wind-waves, and hurricane storm surge along the northern Gulf of Mexico. *J. Geophys. Res.* **121** 3625–3658.

BODE, L. and HARDY, T. A. (1997). Progress and recent developments in storm surge modeling. *J. Hydraul. Eng.* **123** 315–331.

CANGIALOSI, J. P. (2019). National Hurricane Center Forecast Verification Report: 2018 Hurricane Season Technical Report NOAA/National Hurricane Center.

CHANG, W., HARAN, M., OLSON, R. and KELLER, K. (2014). Fast dimension-reduced climate model calibration and the effect of data aggregation. *Ann. Appl. Stat.* **8** 649–673. MR3262529 https://doi.org/10.1214/14-AOAS733

CHANG, W., HARAN, M., APPLEGATE, P. and POLLARD, D. (2016). Calibrating an ice sheet model using high-dimensional binary spatial data. *J. Amer. Statist. Assoc.* **111** 57–72. MR3494638 https://doi.org/10.1080/01621459.2015.1108199

CHANG, W., KONOMI, B. A., KARAGIANNIS, G., GUAN, Y. and HARAN, M. (2019). Ice model calibration using semi-continuous spatial data. Preprint. Available at arXiv:1907.13554.

CONTI, S., GOSLING, J. P., OAKLEY, J. E. and O'HAGAN, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika* **96** 663–676. MR2538764 https://doi.org/10.1093/biomet/asp028

CRESSIE, N. (1990). The origins of kriging. *Math. Geol.* **22** 239–252. MR1047810 https://doi.org/10.1007/BF00889887

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. MR1239641 https://doi.org/10.1002/9781119115151

CRESSIE, N., SHI, T. and KANG, E. L. (2010). Fixed rank filtering for spatio-temporal data. *J. Comput. Graph. Statist.* **19** 724–745. MR2732500 https://doi.org/10.1198/jcgs.2010.09051

CURRIN, C., MITCHELL, T., MORRIS, M. and YLVISAKER, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.* **86** 953–963. MR1146343

CYRIAC, R., DIETRICH, J. C., FLEMING, J. G., BLANTON, B. O., KAISER, C., DAWSON, C. N. and LUETTICH, R. A. (2018). Variability in Coastal Flooding predictions due to forecast errors during Hurricane Arthur. *Coastal Eng.* **137** 59–78.

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. MR3538706 https://doi.org/10.1080/01621459.2015.1044091

DAWSON, R. and HALL, J. (2006). Adaptive importance sampling for risk analysis of complex infrastructure systems. *Proc. R. Soc. Ser. A Math. Phys. Eng. Sci.* **462** 3343–3362.

DIETRICH, J. C., ZIJLEMA, M., WESTERINK, J. J., HOLTHUIJSEN, L. H., DAWSON, C., LUETTICH, R. A. JR., JENSEN, R. E., SMITH, J. M., STELLING, G. S. et al. (2011). Modeling hurricane waves and storm surge using integrally-coupled, scalable computations. *Coastal Eng.* **58** 45–65.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

GRAMACY, R. B. (2016). laGP: Large-scale spatial modeling via local approximate Gaussian processes in R. *J. Stat. Softw.* **72** 1–46.

GRAMACY, R. B. and LEE, H. K. H. (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics* **51** 130–145. MR2668170 https://doi.org/10.1198/TECH.2009.0015

GRAMACY, R. B. and LEE, H. K. H. (2012). Cases for the nugget in modeling computer experiments. *Stat. Comput.* **22** 713–722. MR2909617 https://doi.org/10.1007/s11222-010-9224-x

GU, M. and BERGER, J. O. (2016). Parallel partial Gaussian process emulation for computer models with massive output. *Ann. Appl. Stat.* **10** 1317–1347. MR3553226 https://doi.org/10.1214/16-AOAS934

GU, M. and WANG, L. (2018). Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA J. Uncertain. Quantificat.* **6** 1555–1583. MR3875809 https://doi.org/10.1137/17M1159890

GU, M., XIE, F. and WANG, L. (2018). A theoretical framework of the scaled Gaussian stochastic process in prediction and calibration. Preprint. Available at arXiv:1807.03829.

HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994 https://doi.org/10.1198/016214507000000888

HUNG, Y., JOSEPH, V. R. and MELKOTE, S. N. (2015). Analysis of computer experiments with functional response. *Technometrics* **57** 35–44. MR3318347 https://doi.org/10.1080/00401706.2013.869263

JELESNIANSKI, C. P. (1966). Numerical computations of storm surges without bottom stress. *Mon. Weather Rev.* **94** 379–394.

JELESNIANSKI, C. P., CHEN, J. and SHAFFER, W. A. (1992). SLOSH: Sea, lake, and overland surges from hurricanes Technical Report No. 48 US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.

JIA, G. and TAFLANIDIS, A. A. (2013). Kriging metamodeling for approximation of high-dimensional wave and surge responses in real-time storm/hurricane risk assessment. *Comput. Methods Appl. Mech. Engrg.* **261/262** 24–38. MR3069864 https://doi.org/10.1016/j.cma.2013.03.012

JIA, G., TAFLANIDIS, A. A., NADAL-CARABALLO, N. C., MELBY, J. A., KENNEDY, A. B. and SMITH, J. M. (2016). Surrogate modeling for peak or time-dependent storm surge prediction over an extended coastal region using an existing database of synthetic storms. *Nat. Hazards* **81** 909–938.

JOHNSON, L. R., GRAMACY, R. B., COHEN, J., MORDECAI, E., MURDOCK, C., ROHR, J., RYAN, S. J., STEWART-IBARRA, A. M. and WEIKEL, D. (2018). Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: A Dengue case study. *Ann. Appl. Stat.* **12** 27–66. MR3773385 https://doi.org/10.1214/17-AOAS1090

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

LINDLEY, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27** 986–1005. MR0083936 https://doi.org/10.1214/aoms/1177728069

LIU, F. and WEST, M. (2009). A dynamic modelling strategy for Bayesian computer model emulation. *Bayesian Anal.* **4** 393–411. MR2507369 https://doi.org/10.1214/09-BA415

LOEPPKY, J. L., MOORE, L. M. and WILLIAMS, B. J. (2010). Batch sequential designs for computer experiments. *J. Statist. Plann. Inference* **140** 1452–1464. MR2592224 https://doi.org/10.1016/j.jspi.2009.12.004

MAK, S., SUNG, C.-L., WANG, X., YEH, S.-T., CHANG, Y.-H., JOSEPH, V. R., YANG, V. and WU, C. F. J. (2018). An efficient surrogate model for emulation and physics extraction of large eddy simulations. *J. Amer. Statist. Assoc.* **113** 1443–1456. MR3902221 https://doi.org/10.1080/01621459.2017.1409123

MOREY, S. L., BAIG, S., BOURASSA, M. A., DUKHOVSKOY, D. S. and O'BRIEN, J. J. (2006). Remote forcing contribution to storm-induced sea level rise during Hurricane Dennis. *Geophys. Res. Lett.* **33** L19603.

PARKER, K., RUGGIERO, P., SERAFIN, K. A. and HILL, D. F. (2019). Emulation as an approach for rapid estuarine modeling. *Coastal Eng.* **150** 79–93.

PENG, S.-Q. and XIE, L. (2006). Effect of determining initial conditions by four-dimensional variational data assimilation on storm surge forecasting. *Ocean Model.* **14** 1–18.

PLUMLEE, M. (2017). Bayesian calibration of inexact computer models. *J. Amer. Statist. Assoc.* **112** 1274–1285. MR3735376 https://doi.org/10.1080/01621459.2016.1211016

PLUMLEE, M., ASHER, T. G, CHANG, W. and BILSKIE, M. V (2021). Supplement A to "High-fidelity hurricane surge forecasting using emulation and sequential experiments." https://doi.org/10.1214/20-AOAS1398SUPPA

PLUMLEE, M., ASHER, T. G, CHANG, W. and BILSKIE, M. V (2021). Supplement B to "High-fidelity hurricane surge forecasting using emulation and sequential experiments." https://doi.org/10.1214/20-AOAS1398SUPPB

PLUMLEE, M., ASHER, T. G, CHANG, W. and BILSKIE, M. V (2021). Supplement D to "High-fidelity hurricane surge forecasting using emulation and sequential experiments." https://doi.org/10.1214/20-AOAS1398SUPPD

PLUMLEE, M., ASHER, T. G, CHANG, W. and BILSKIE, M. V (2021). Supplement C to "High-fidelity hurricane surge forecasting using emulation and sequential experiments." https://doi.org/10.1214/20-AOAS1398SUPPC

RANJAN, P., BINGHAM, D. and MICHAILIDIS, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics* **50** 527–541. MR2655651 https://doi.org/10.1198/004017008000000541

RESIO, D. T. and WESTERINK, J. J. (2008). Modeling the physics of storm surges. *Phys. Today* **61** 33–38.

ROHMER, J. and IDIER, D. (2012). A meta-modelling strategy to identify the critical offshore conditions for coastal flooding. *Nat. Hazards Earth Syst. Sci.* **12** 2943–2955.

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

SAKIA, R. (1992). The Box-Cox transformation technique: A review. *J. R. Stat. Soc., Ser. D Stat.* **41** 169–178.

SANSÓ, B. and FOREST, C. (2009). Statistical calibration of climate system properties. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 485–503. MR2750089 https://doi.org/10.1111/j.1467-9876.2009.00669.x

SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments. Springer Series in Statistics*. Springer, New York. MR2160708 https://doi.org/10.1007/978-1-4757-3799-8

SHEPARD, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the* 1968 23*rd ACM National Conference* 517–524. ACM, New York.

TAFLANIDIS, A. A., KENNEDY, A. B., WESTERINK, J. J., SMITH, J., CHEUNG, K. F., HOPE, M. and TANAKA, S. (2012). Rapid assessment of wave and surge risk during landfalling hurricanes: Probabilistic approach **139** 171–182.

TANAKA, S., BUNYA, S., WESTERINK, J. J., DAWSON, C. and LUETTICH, R. A. JR. (2011). Scalability of an unstructured grid continuous Galerkin based hurricane storm surge model. *J. Sci. Comput.* **46** 329–358. MR2765498 https://doi.org/10.1007/s10915-010-9402-1

TAYLOR, A. A. and GLAHN, B. (2008). Probabilistic guidance for hurricane storm surge. In 19*th Conference on Probability and Statistics* **74** 8.

TORO, G., NIEDORODA, A., REED, C. and DIVOKY, D. (2010). Quadrature-based approach for the efficient evaluation of surge hazard. *Ocean Eng.* **37** 114–124.

TUO, R. and WU, C. F. J. (2015). Efficient calibration for imperfect computer models. *Ann. Statist.* **43** 2331–2352. MR3405596 https://doi.org/10.1214/15-AOS1314

VERNON, I., GOLDSTEIN, M. and BOWER, R. (2014). Galaxy formation: Bayesian history matching for the observable universe. *Statist. Sci.* **29** 81–90. MR3201849 https://doi.org/10.1214/12-STS412

WALKER, A. M., TITLEY, D. W., MANN, M. E., NAJJAR, R. G. and MILLER, S. K. (2018). A fiscally based scale for tropical cyclone storm surge. *Weather Forecast.* **33** 1709–1723.

WELCH, W. J., BUCK, R. J., SACKS, J., WYNN, H. P., MITCHELL, T. J. and MORRIS, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics* **34** 15–25.

WESTERINK, J. J., LUETTICH, R. A., FEYEN, J. C., ATKINSON, J. H., DAWSON, C., ROBERTS, H. J., POWELL, M. D., DUNION, J. P., KUBATKO, E. J. et al. (2008). A basin- to channel-scale unstructured grid hurricane storm surge model applied to southern Louisiana. *Mon. Weather Rev.* **136** 833–864.

WILLMOTT, C. J. and MATSUURA, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **30** 79–82.