Emulation as an Approach for Rapid Estuarine Modeling

Kai Parker [a, *], Peter Ruggiero [b], Katherine A. Serafin [c], David F. Hill [d]

*[a, *] School of Civil and Construction Engineering, Oregon State University, Corvallis, OR, USA 97330. parkerk@oregonstate.edu*
[b] College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA, 97330. pruggier@ceoas.oregonstate.edu
[c] Department of Geophysics, Stanford University, Stanford, CA, USA 94305-2215. kserafin@stanford.edu
[d] School of Civil and Construction Engineering, Oregon State University, Corvallis, OR, USA, 97330. david.hill@oregonstate.edu

* Corresponding Author
Permanent Address: 1717 7th Street, Los Osos, CA, USA, 93402. kaiparker@gmail.com

Abstract:

Probabilistic flood hazard assessment is a promising methodology for estuarine risk assessment but currently remains limited by prohibitively long simulation times. This study addresses this problem through the development of an emulator, or surrogate model, which replaces the simulator (in this case the coupled ADCIRC+SWAN model) with a statistical representation that is able to rapidly predict estuarine variables relevant to flooding. Emulation of water levels (WLs), non-tidal residual, and significant wave height, is explored at Grays Harbor, Washington (WA) USA using Gaussian process regression. The effectiveness of the methodology is validated at various model simplification levels to determine where error is being sourced. Emulated WLs are found to be skillful when compared to over a decade of tide gauge observations (root mean square error, RMSE, <15 cm). The largest loss of skill in the method originates with ADCIRC+SWAN attempting to reproduce observations, even when the majority of relevant physics are included. Subsequent simplifications to the simulator (input reduction techniques) and the emulator itself are found to introduce a trivial amount of error (average increase in RMSE of 1 cm). Emulated WLs are also compared to spatially varying observations and found to be equally skillful throughout the estuary. An example emulation application is explored by decomposing the relative forcing contributions to extreme WLs across the study site. Results show a compound nature of extreme estuarine WLs in that all forcing dimensions contribute to extremes, with streamflow having the least influence and tides the largest. Overall the approach is shown to be both skillful and efficient at reproducing critical hydrodynamic variables, suggesting that emulation may play a key role in improving our ability to probabilistically assess flood risk in complex environments as well as being promising in a range of other applications.

Keywords: Emulation; Gaussian Process Regression; ADCIRC+SWAN; Estuary; Probabilistic Modeling; Water Levels

1. Introduction

Modeling estuarine hydrodynamics remains both a challenge and a goal for the scientific community. Estuaries and bays are often densely populated with significant economic and cultural investment [Pendleton, 2010]. They are also subject to a unique flood hazard environment, with high water levels (WLs) driven by numerous contributing processes including both offshore and local waves, storm surge, and river inflows, among others. Over the past several decades, research efforts have led to improved computational models and increased physical understanding of estuarine flood dynamics [Bode and Hardy, 1997; Kantha and Clayson, 2000; Ganju et al., 2015]. However, increasing hydrodynamic model predictive skill is generally coupled to increasing complexity within numerical models and a correspondingly larger computational load. This has led to computational time, rather than a physical understanding of the problem, being a limiting control on our ability to answer questions about estuarine flooding.

Increasing computer processing power and code parallelization has pushed the boundary for what can be explored with complex computer codes. However, even with these advances, many questions still cannot be comprehensively addressed due to computational limitations. One example is the recent focus by the scientific community on uncertainty in model results [Mastrandrea et al., 2010; Green et al., 2011]. In the field of flood hazards, a major thrust area has been probabilistic assessments, which brings the benefits of uncertainty quantification, utility as a stakeholder-centered decision making tool, better handling of extreme events, and more skillful flooding estimates [Cloke and Pappenberger, 2009; Di Baldassarre et al., 2010; Dale et al., 2014]. However, the combination of multiple model iterations (required for probabilistic modeling) and large per-run computational costs has remained a barrier for moving forward.

Often the solution to long simulation times is a compromise, such as simplifying or eliminating various forcing components [Purvis et al., 2008; Lin et al., 2010]; using smaller ensemble sizes [Davis et al., 2010]; or simplifying model physics [Dawson et al., 2005; Moel et al., 2012]. A promising recent development has been to implement variable model complexity, with a fast model determining relevant or extreme events and a more highly-resolved, accurate model being used to simulate the extremes [Lin et al., 2010, 2012; Orton et al., 2016]. This technique has been successfully demonstrated for hurricane-induced flooding but is potentially problematic for other regions. For example, environments not dominated by tropical cyclones often are defined by compound events where combinations of non-extreme forcings can combine to create extremes [Leonard et al., 2014; Wahl et al., 2015; Moftakhari et al., 2017; Zscheischler et al., 2018]. In addition, event based techniques can still be considered computationally limited as the full parameter space cannot usually be explored. There remains a need for a modeling technique that can bridge the gap between time-intensive, complex models and fast simulation times.

This paper investigates emulation as a technique for the efficient prediction of estuarine hydrodynamic variables in Grays Harbor, Washington (WA) USA. The foundational idea of emulation (also referred to as surrogate modeling, response surface modeling, and meta-modeling, among others) is the replacement of a slower processes-based model (a simulator) with a fast, statistical model (an emulator) [O'Hagan, 2006; Razavi et al., 2012]. In the standard modeling paradigm, the map between simulator inputs and outputs is based on the laws of physics as implemented within a process-based model [Castelletti et al., 2012]. In emulation, this map is approximated using a statistical model. The benefit is that, following an upfront computational expense to create a training dataset and train the emulator, applying the emulator is nearly instantaneous. Thus, emulation represents a tradeoff between short simulation times and errors associated with the approximation. This tradeoff suggests that emulation may be ideal for probabilistic flood modeling along with many other potential applications including assessments of model uncertainty, model optimization, sensitivity analysis, real time forecasting, and extreme event analysis [Oakley, 1999; Kennedy et al., 2006; Levy and Steinberg, 2010].

The general concept of emulation originated in the 1980s through the idea of computer experiments [Sacks et al., 1989]. Since then, emulation ideas have spread widely resulting in a rich literature of applications, emulator formulations, and theories from numerous fields. Razavi et al. [2012] reviews emulation in the field of water resources, with over 30 studies revealing a wide range of applications and emulation approaches. As a brief overview of coastal applications, Gouldby et al [2014], Malde et al. [2016b] and Rueda et al. [2016] successfully implemented emulators for wave prediction problems using SWAN [Booij et al., 1997] as a simulator. The pairing of SWAN and emulation was extended to delineating offshore conditions causing wave induced coastal flooding by Rohmer and Idier, [2012] by using kriging and an adaptive sampling technique. Timmermans [2015] used emulation to explore how tuning parameters affect uncertainty in results from the Wave Watch III [Tolman, 2009] wave model. Liu and Guillas [2017] investigated the effect of uncertainty in bathymetry on tsunami height predictions using a novel merging of Gaussian process regression (GPR) emulation with dimensional reduction techniques.

In the context of flooding, emulation has been applied to river channel flooding [Apel et al., 2008] and coastal dyke systems [Moel et al., 2012], although from the relatively simplistic perspective of lookup tables. Surge response functions [SRF; Resio et al., 2009; Song et al., 2012] can be considered a specific case of emulation through regression of dimensionless cyclone scaling terms. However, SRFs are limited in application to tropical cyclones, and have been shown to perform poorly in complex environments [Taylor et al., 2015]. As an alternative to SRFs, Kim et al. [2015] used an artificial neural network to emulate coupled ADCIRC+STWAVE calculated surge from tropical cyclones. This approach was enhanced by Bass and Bedient [2018] who used a similar strategy but with the addition of a coupled hydrologic model and GPR as the emulator formulation. Jia et al. [2013; 2016] used GPR emulation for predicting tropical cyclone surges.

Overall, multiple studies have demonstrated the potential of emulation in a coastal hazard setting. Surge from tropical cyclones has, in particular, seen a variety of successful emulator implementations. This study builds on these recent efforts but explores an estuary in the USA Pacific Northwest (PNW) that does not experience tropical cyclone forcing. This results in a unique challenge in terms of handling diverse forcings and a potentially larger input parameter space, since there is no dominant forcing dimension. Other studies focused on predicting WLs, such as those by Jia et al. [2013; 2016], Kim et al. [2015], and Bass and Bedient [2018], reduce input dimensionality through considering only cyclones and using discrete cyclone characteristics as input dimensions. This study, however, considers a general application of emulating the coupled ADCIRC+SWAN [ADCSWAN; Dietrich et al., 2011] simulator in which any combination of forcings can be used to calculate WLs. This paper is intended as a rigorous investigation into the applicability of emulation in this new context. Therefore, the focus here is primarily on describing the methodology and validation and only a single application, decomposing extreme estuarine water levels, is presented.

2. Study Sites and Observations

2.1 Study Site

Grays Harbor, WA (Figure 1) is an excellent candidate for testing emulation as it exhibits many of the complexities that make estuarine modeling difficult. Grays Harbor is predominantly shallow, dominated by depths averaging less than 5 meters, but also contains a maintained (United States Army Corps of Engineers; USACE) deep-water navigation channel giving it significant depth variability (Figure 1). The bay exhibits spatial variability in WLs [Cialone et al., 2001] as a result of its size [approximately 235.3 km^2, Engle et al., 2007], shape, and gradients in forcing. Grays Harbor is located in the PNW (Figure 1) and is therefore subject to an energetic storm and wave climate. A Global Ocean Wave 2 (GOW2) reanalysis [Perez et al., 2017] near the study site (see Figure 1) reveals a mean offshore significant wave height (Hs) of 2.5 meters with events exceeding 7.5 meters annually. Extreme storm events are generally associated with extratropical cyclones that can produce strong winds, pressure differentials, and precipitation [Allan and Komar, 2002b; Mass and Dotson, 2010]. These events are often associated with significant non-tidal residuals (NTR) [Allan and Komar, 2002b, 2006; Allan et al., 2011; Serafin et al., 2017], although of a smaller magnitude than locations impacted by tropical cyclones or with broader continental shelves [Zhang et al., 1999]. Within this study, NTR is defined as an observed or modeled WL with tides removed (with the specifics of how NTR is calculated detailed in section 4.3). Grays Harbor has significant hydrological input from the Chehalis, Humptulips, Hoquiam, Elk, and Johns Rivers which collectively drain a watershed of over 7,000 km^2 for an average monthly runoff volume of 22 million m^3/month [Engle et al., 2007].

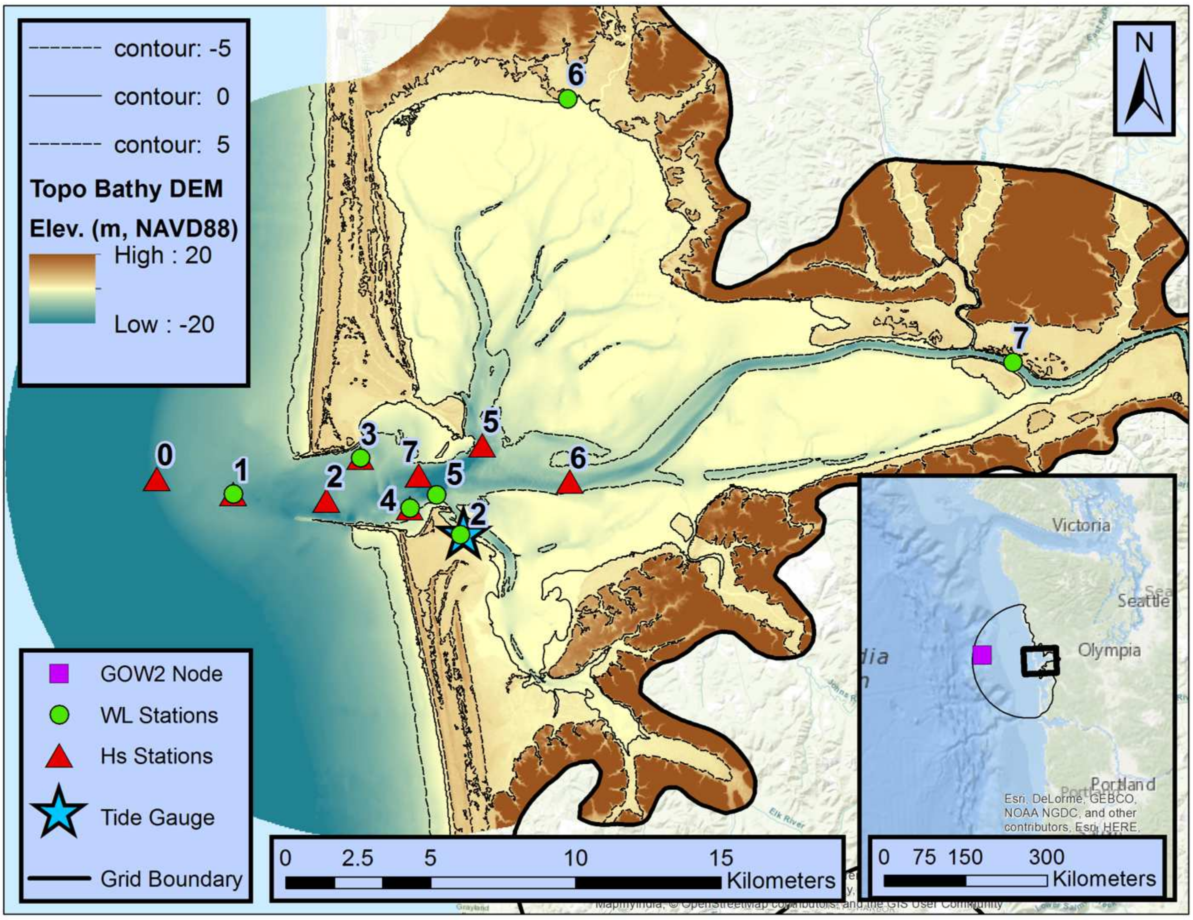

Figure 1: Grays Harbor, WA study site and locations of observational datasets. Circles and triangles represent USACE deployments with co-located instruments labeled with a single number representing both WL and Hs stations. The main panel shows the bathymetry and topography of the estuary in the NAVD88 Datum. The inset panel shows the larger geographical context of the estuary with the thin black line delineating the domain of the hydrodynamic model. The purple square within the inset is the location of the utilized GOW2 node (located at 47° N, 125° W).

2.2 Observational Data

This study utilizes a variety of observational datasets ranging from instrument deployments to reanalysis products. Forcing and model development datasets are explained in the following section (2.2.1), while section 2.2.2 details observations specifically used for model validation.

2.2.1 Forcing and Model Development Datasets

Wave forcing for the model was obtained from the GOW2 reanalysis of Perez et al. [2017] with output selected from a node located at (Lat: 47° N, Lon: 125° W; Figure 1). Atmospheric forcing was provided by the North American Regional Reanalysis (NARR) [Mesinger et al., 2006]. NARR provides a wide range of gridded atmospheric variables from which the 3-hourly 10m wind fields and 3-hourly surface pressure fields were utilized. Streamflow was obtained from USGS river gauges with total estuary inflow constructed as the sum of three gauged rivers, the Chehalis, Satsop, and Wynoochee (USGS stations 12031000, 12035000, and 12037400 respectively). The Satsop and Wynoochee rivers are tributaries to the Chehalis river which join the Chehalis below the Chehalis gauge. Therefore, the sum of these three gauges reproduces the majority of the Chehalis flow into the Grays Harbor estuary. While Grays Harbor has other river inlets, the majority of the input flow is concentrated at the Chehalis River which captures around 80% of the watershed area. For simplicity, as well as due to temporal availability of gauge data, only the Chehalis input (as constructed from the three gauged rivers) is included in the study with all other streamflow inputs assumed to be minimal with only local influences on variables of interest.

The bathymetry data for the simulator grid were developed by blending two National Oceanic and Atmospheric Administration (NOAA) digital elevation models (DEMs): the Astoria, OR tsunami DEM (1/3 arc second) and the coastal relief model (3 arc seconds) [NOAA National Centers for Environmental Information, 2003; Love et al., 2012]. Bay topography was sourced from Oregon Department of Geology and Mineral Industries (DOGAMI) LiDAR [DOGAMI, 2010].

2.2.2 Validation Datasets

In addition to forcing, a series of observational datasets were used to validate simulated and emulated variables within the study site. The first dataset is the Westport, WA tide gauge (NOAA station ID # 9441102) which provides continuous hourly WL data beginning in 2006. WL observations were decomposed into constituent components (e.g., deterministic tide, monthly mean sea level anomalies (MMSLA), storm surge etc.) using the approach described in

Serafin and Ruggiero [2014]. The five largest NTR events on record were extracted for testing model skill. A brief summary of these storm events is provided in Table 1.

Table 1: Summary of forcing for the five largest NTR events at Grays Harbor, WA. Forcing values are reported at the occurrence of maximum NTR.

	Storm 1	**Storm 2**	**Storm 3**	**Storm 4**	**Storm 5**
Date	12/15/06	12/3/07	01/12/14	12/12/14	12/11/15
Non-Tidal Residual (m)	0.73	0.93	0.66	0.58	1.11
Significant Wave Height (m)	8.0	11.1	8.5	6.2	11.1
Peak Wave Period (sec)	12.7	15.4	14.3	12.2	18.2
Wave Direction (deg.)	243	195	259	229	248
Surface Pressure (HPa.)	977	989	989	986	984
Wind Speed (m/s)	17.5	19.2	14.9	9.1	12.3
Wind Direction (deg.)	217	201	238	218	198
Streamflow (m^3/s)	1270	2020	860	670	1510

Water level observations at the tide gauge were supplemented by a field campaign carried out by the USACE from September-December 1999 [Figure 1, Cialone et al., 2001; 2002]. This dataset includes seven locations near the inlet with bottom mounted tripods measuring wave characteristics, WLs, tidal currents, and suspended sediment concentrations. Additionally, five surface stations were distributed throughout the bay measuring WLs, conductivity, and temperature. The USACE field campaign was broken up into two deployments (with a small maintenance/data collection break between the two). Instruments were replaced in approximately the same location except for Hs station 0 which was moved to location 7 for the second deployment [Cialone et al., 2002]. Figure 1 illustrates the spatial distribution of the various observation stations which have been renamed in this paper for clarity.

3. Methods

3.1 Simulator Configuration

This study utilizes the coupled Advanced Circulation [ADCIRC; Luettich and Westerink, 1992] and unstructured Simulating Waves Nearshore [SWAN, Zijlema, 2010] simulator

[ADCSWAN; Dietrich et al., 2011]. ADCSWAN has seen extensive validation and success in predicting WLs and NTR at various estuaries around the world [Dietrich et al., 2012; Bhaskaran et al., 2013; Krien et al., 2015]. Recently, ADCSWAN has been successfully implemented in the PNW with good agreement between simulator output and observations of WLs, NTR, and currents [Cialone et al., 2002; Cheng et al., 2015b]. ADCSWAN is implemented in the 2D depth-integrated barotropic mode which has been shown to perform with acceptable error for computing WLs and depth integrated currents in estuaries [Resio and Westerink, 2008; Weaver and Luettich, 2010]. ADCIRC is run in the fully 2-way coupled implementation with SWAN, which has been shown to be critical for resolving interactions between waves and nearshore hydrodynamics [Cialone et. al., 2002; Funakoshi et al., 2008; Dietrich et al., 2010, 2011;]. ADCSWAN is run on an unstructured mesh that extends beyond the continental shelf (approximately 115 kilometers offshore; Figure 1). Unstructured meshes provide flexibility in simulator resolution with the utilized model grid having element sizes ranging from around 7,000 meters offshore to under 20 meters within the inner Grays Harbor channel.

3.2 Dimensional Reduction and Levels of Simplification

Emulator construction requires sampling the full input parameter space. This constraint dictates that the number of times the simulator must be run to create the training dataset is proportional to the number of dimensions included as inputs. In general, process-based hydrodynamic simulators are based on many inputs making some form of dimensional reduction necessary. Emulator construction thus requires finding a balance between minimizing the number of inputs and maintaining sufficient complexity to acceptably resolve output variables of interest.

Figure 2 provides a conceptual model of the dimensional reduction approach taken in this study (through simplifications), transforming the full process-based simulator (ADCSWAN) into an emulator. Each of the simplifications, noted on the right side of Figure 2, theoretically introduces some level of error into the output, noted on the left side of Figure 2. These errors are discussed in this paper both as individual contributions, and in the cumulative (sum of all errors up to a given level) sense. When discussed explicitly in this paper, simplification levels will be capitalized. For example, a comparison of model output from the level 3 simplification (Stationary Simulator) to Observations (no simplification) quantifies the cumulative level 3 error.

The following sections (3.2.1 – 3.2.3) explain each simplification in this hierarchy while corresponding error is quantified in the Results section.

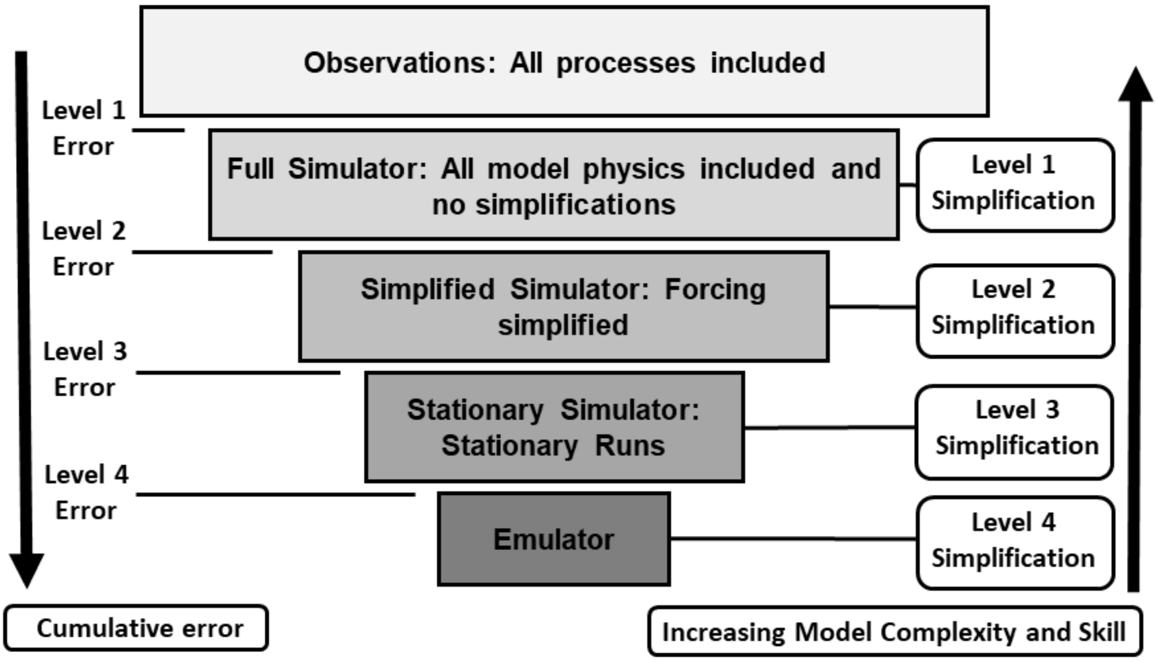

Figure 2: Hierarchy of model simplifications between observations (top) and emulator output (bottom). Each simplification is associated with a level (right side of the figure) and some amount of error (defined on the left side of the figure).

3.2.1 Simulator Simplifications

The first level of simplification is simply that of using a process-based simulator. Simulators are unable to exactly reproduce observations for a variety of reasons ranging from incorrect or unresolved physics (e.g., assumptions, parameterizations, etc.) to numerical approximations (truncation errors, etc.) to incorrect or biased input forcing. The x induced by this simplification is primarily a function of the chosen model, model tuning, and the quality of forcing/bathymetric information. Research has shown that errors in model inputs such as bathymetry and mesh resolution [Bunya et al., 2010; Weaver and Slinn, 2010] and forcing fields [Madsen and Jakobsen, 2004; Lewis et al., 2013; Lakshmi et al., 2017] are significant sources of model error. Therefore, the specific configuration and choice of ADCSWAN (section 3.1) and the quality of observational data (section 2.2.1) are the primary controls on the impact of this simplification.

This study considers emulation of a specific implementation of the ADCSWAN model and therefore the model grid (bathymetry, resolution, etc.) is held constant. Additionally, ADCSWAN contains a large number of input switches, tuning parameters, forcing options, numerical configurations, and other choices [Westerink et al., 1992]. This study holds all general model configuration parameters constant leaving the various forcing components of WL variability as the sole driver of input dimensionality within the emulator.

3.2.2 Forcing Simplifications

Even with the simplification of holding the model configuration fixed, the input dimensionality remains high, due to the numerous physical forcing mechanisms. Below we describe simplifications that reduce the model dimensionality to 16. This reduction is desirable since it requires a smaller training dataset and therefore produces a more efficient emulator construction.

3.2.2.1 Wave Simplification

It is well known that offshore wave energy can impact water levels within bays such as Grays Harbor [Olabarrieta et al., 2011; Cheng et al., 2015b]. Wave forcing is implemented in the simulator using a JONSWAP spectrum fitted to peak wave period (Tp), Hs, mean wave direction (MWD), and directional spread parameters. While research has shown the importance of forcing with full directional spectra for reproducing wave observations [Rogers et al., 2007; Montoya et al. 2013], most studies accounting for wave influence on WLs use simpler bulk parameter-based formulations. Therefore, a fitted JONSWAP spectrum is used for both the Full (level 1) and Simplified Simulator (level 2) comparisons. Based on previous research in the PNW [Cheng et al., 2015a], directional spread is held constant at 20 degrees, and wave forcing is applied uniformly along the Full Simulator open boundary (Figure 1). With these simplifications, wave forcing is included in the emulator as three dimensions: Hs, Tp, and MWD.

3.2.2.2 Atmospheric Simplification

Atmospheric forcing represents a unique challenge for emulation due to the spatial variability of wind and pressure fields. Gridded inputs represent a high degree of dimensionality, with every node potentially representing an input dimension. For this reason, a sensitivity study was undertaken to see if spatially constant atmospheric forcing could be used as an approximation of the full forcing fields. WL output from simulator runs with full gridded forcing were compared to runs with spatially constant forcing. Results indicated (not shown) that the error introduced in predicted WLs by the spatially constant assumption was acceptable in comparison to the corresponding reduction in dimensionality. This error is quantified in the Results section (along with other simulator simplifications) as level 2 error. Adopting the spatially constant assumption, atmospheric forcing is reduced in the emulator framework to three dimensions: wind speed, wind direction, and sea level atmospheric pressure.

3.2.2.3 Tidal Simplification

Tidal forcing is generally represented in hydrodynamic models through harmonic constituents. Many studies using ADCIRC are forced with eight or fewer constituents, mainly because global databases of tidal constituents (e.g., TPXO [Dushaw et al., 1997], or LeProvost [Le Provost et al., 1994]) are typically limited to that number. Despite this, simulations using this small number of constituents are typically found to agree well with both harmonic analysis derived and observed tidal elevations [Westerink et al., 1992; Blain and Rogers, 1998; Blain et al., 2001]. ADCIRC simulates tidal forcing as a boundary elevation time series [Luettich et al., 1992] determined by a spatially variable, temporally constant phase and amplitude and a temporally variable, spatially constant equilibrium argument and nodal factor. Amplitudes and phases are determined by the simulator boundary location and are therefore not an emulator input dimension when considering a fixed study site. The nodal factor represents adjustments of the amplitude/phase of each constituent that results from the nodal tide cycle. The equilibrium argument (deterministic based on date and time) controls the timing of the harmonic.

While tides are deterministic, they are included within the emulator as forcing for a variety of reasons which will be described in section 3.2.3. In approaching simplifications, a sensitivity test was performed to determine the tidal dimensionality required for accurately reproducing maximum WLs during storm events (Table 1). It was found that removing the nodal factor did not significantly change simulated WLs. After this simplification, results showed that eight harmonics (without nodal factors) were sufficient for accurately producing WLs. This allows tides to be included in the emulator as eight input dimensions: 8 harmonic equilibrium arguments, each ranging from 0 to 360 degrees.

3.2.2.4 Streamflow Simplification

Streamflow is represented in ADCIRC as a flux of water into the domain (specified as a normal flow per unit width of boundary). This allows the simulation of large rivers that have significant cross-channel velocity profiles and for calibration where data on these cross-channel profiles are available. For this study, we instead specify a laterally constant velocity profile across each river boundary. This simplification is common [Bunya et al., 2010; McKay and Blain, 2010], especially if the boundary is far enough away from the area of interest that a natural flow profile can develop. This allows streamflow to be represented as a single input dimension (the total volumetric flow rate) for each river inlet.

3.2.2.5 Base Water Level Simplifications

A final input dimension is considered within the emulator framework as a "Base WL" parameter. This is included to account for large scale changes to estuary sea level, as is experienced through MMSLAs, seasonal variability, and sea level rise (SLR) [Serafin and Ruggiero, 2014]. These forcing dimensions are defined in the simulator simply as a static change to mean sea level and are therefore included in the emulator as a single input dimension.

3.2.3 Simulator Stationarity Simplification

ADCSWAN and other process-based hydrodynamic simulators are dynamic in that both inputs and outputs are functions of time and the simulator state is determined, in part, by previous states. Seeking simplicity, this study makes the assumption that the dynamic system can be approximated using a series of stationary simulations. Precedents for such an assumption exist for coastal systems, including spectral evolution in wave modeling (SWAN) approximated using a series of steady-state simulations [Rogers et al., 2007; Rusu and Pilar, 2008].

Simplifying tidal forcing with stationary simulations is difficult since there is no tidal equilibrium in WLs. One approach would be to consider tides as a series of horizontal water surfaces of different elevations (corresponding to tidal phases). This would reduce tidal forcing dimensionality to a single value (tidal WL), but at the cost of losing spatial variability. Testing showed that, for the Grays Harbor study site, tidal wave evolution and propagation across the estuary results in significant spatial variability in tidally forced WLs. A second approach would be to decouple NTR and tidal WLs and add the two as a linear summation. However, further testing confirmed that this simplification results in significant error. Therefore, a hybrid solution was developed in which all non-tidal forcing is stationary, but tides are computed dynamically with model output recorded only at the specific moment of interest. This approach is appropriate since tides are deterministic and, for a specific set of equilibrium arguments, the previous state of tide induced WLs will always be the same. This approach allows tidal forcing to be simplified but retains the spatial variability in tidal WLs and the nonlinear interactions with other processes.

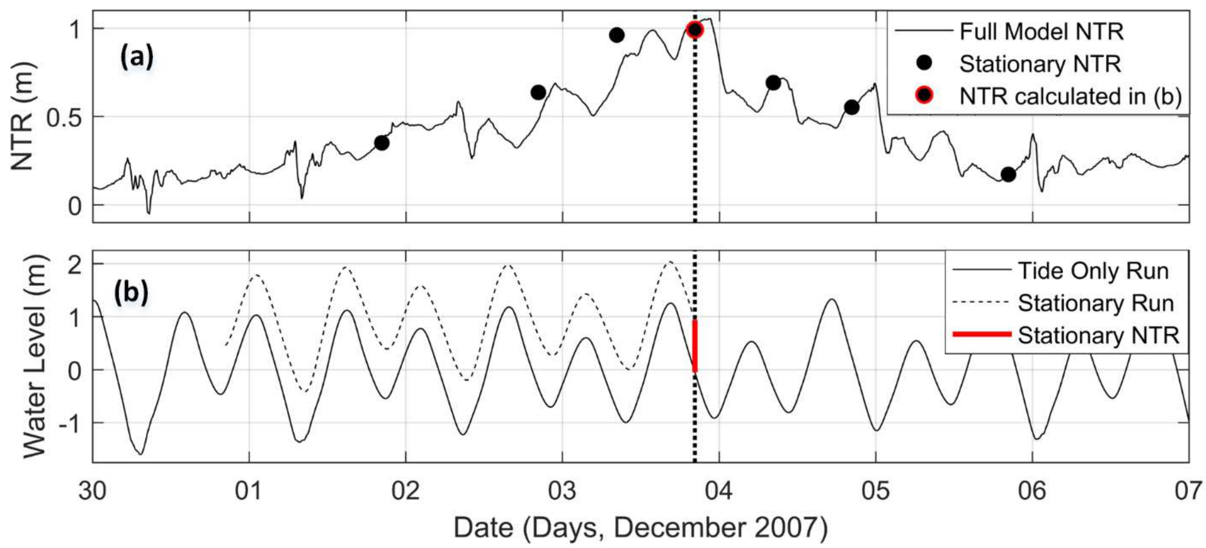

Figure 3: Panel (a): Comparison of NTR during storm 2 from a fully dynamic simulation (black line) and simplified stationary simulations (black dots). Panel (b): Example stationary run (at the peak of the storm) showing how the stationary NTR is calculated. The horizontal bold dotted line represents the time of the stationary run. At this time, NTR is calculated by subtracting the value of a tide only run from the value of the stationary run (Bold red line). This NTR value is plotted as a red outlined dot in panel (a).

Figure 3 illustrates how stationary runs compare to the Full Simulator (dynamic). Figure 3a compares NTR from the fully forced ADCSWAN (simplification level 1; black line) and seven stationary ADCSWAN runs (simplification level 3; black dots) during storm 2 (Table 1). NTRs are computed for both cases by subtracting a ‘tides only’ simulation from the fully forced model. Figure 3b demonstrates how the stationary NTR is computed for the peak of storm 2. This NTR value is plotted in Figure 3a as a red outlined dot. The agreement between the fully dynamic run and the seven stationary runs was found to be sufficient, with an RMSE error for storm 2 of 11 cm.

3.3 Experimental Design

A conceptual overview of the process used for constructing an emulator, in the context of this study, is provided in Figure 4.

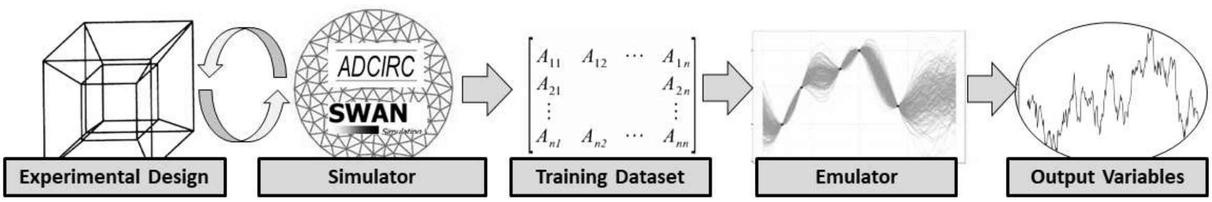

Figure 4: Conceptual framework for developing an estuarine hydrodynamic emulator.

The first step in building an emulator is the selection of design points (experimental design) to create the training dataset. This study implements a design from the commonly utilized Latin Hypercube sampling (LHS) family of schemes first explored by McKay et al. [1979]. LHS is one of the oldest and most popular experimental designs and has been found to perform well for complex simulators [Jones and Johnson, 2009]. The specific experimental design for this study was created using a “maximin” LHS design [Johnson, 1990; Morris and Mitchell, 1995] from the LHS package in R [Carnell, 2017].

Parameters required for a LHS design are the number of dimensions to be included, the range of each dimension, and the number of design points. As detailed in section 3.2, this study used an input parameter dimensionality of 16, including wind speed and direction, sea surface pressure, Hs, Tp, MWD, streamflow, base WL, and eight tidal equilibrium arguments. LHS considers only the maximum and minimum values of each dimension with design points spaced approximately uniformly across dimensions. Ranges were chosen for each parameter in an attempt to span all plausible forcing scenarios. This was determined by looking at 100-year return level events as calculated from the observational records. The size of the training dataset is typically controlled by the cost of running the simulator, but Loeppky et al. [2009] provide the general guidance that the training dataset should be approximately 10 times the number of dimensions of the input space. Given the 16 input dimensions of this study, this suggests a theoretical training dataset size of 160 runs. To explore the relationship between training dataset size and emulator skill and to validate the emulator's overall effectiveness, this study conservatively developed a larger training dataset consisting of 480 ADCSWAN runs.

3.4 Emulator Configuration

A variety of formulations have previously been used in an emulation context, including support vector machines, artificial neural networks, radial basis functions, and many others [Jin et al., 2001; Gano et al., 2006; Razavi et al., 2012]. This study uses GPR, (also referred to as Kriging), a Bayesian statistical non-parametric regression model well suited to this particular application as it scales well to high-dimensional input and intrinsically considers model uncertainty [O'Hagan, 2006; Levy and Steinberg, 2010]. Furthermore, GPR is a general and flexible framework that can be optimized for a variety of modeling problems [Rasmussen and Williams, 2006]. For example, many other common emulator formulations, such as neural networks [Rasmussen and Williams, 2006] and radial basis functions [Anjyo and Lewis, 2011], can be shown to be equivalent to GPR under specific conditions.

The foundational definition of a Gaussian process is that of an infinite collection of variables for which any finite subset is described by a multivariate Gaussian distribution. Every point in the input space can be modeled as a random variable (due to uncertainty about the functional response to inputs). A Gaussian process governs how these variables are related. A common way of thinking about GPR is as a distribution over functions [Rasmussen and Williams, 2006]. This is mathematically tractable as a GPR can be completely defined by a mean and covariance function (due to being modeled as a multivariate Gaussian distribution). From a Bayesian perspective, this means a GPR is specified using a prior mean and covariance function.

The data then updates this prior, using Bayesian inference, with information about the true form of the function to develop the posterior. The mean posterior function is then the most probable function (considering all possible functions) given the data that has been observed.

This process is conceptualized for a one-dimensional case in Figure 5. The effect of the Bayesian conditioning on the emulator can be seen as "anchoring" the posterior sample functions (and uncertainty) at locations of observations. This limits the possible functions to those that go through these observed points. Uncertainty is quantified by considering the possible functions that pass through these training points.

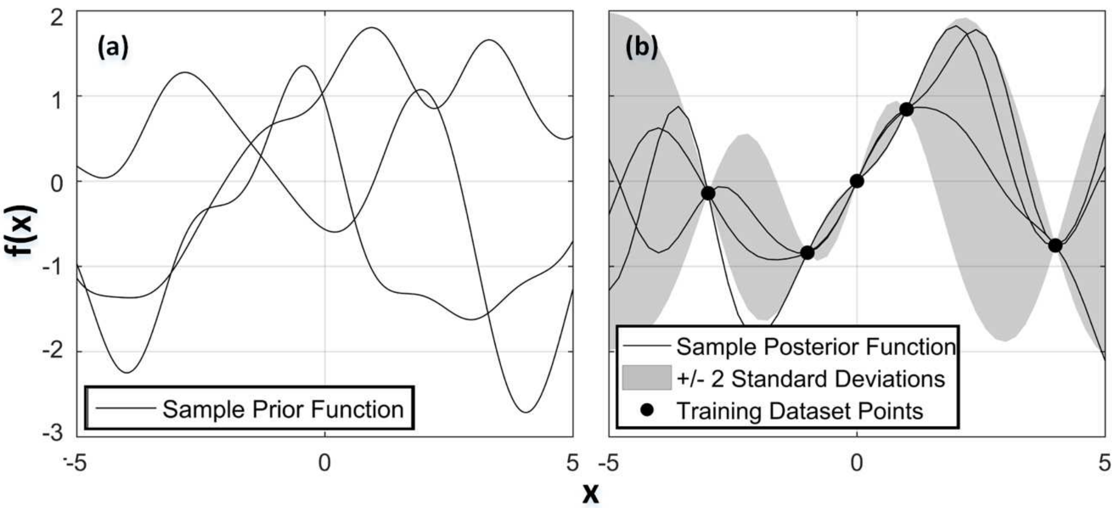

Figure 5: Example 1-D application of GPR for determining f(x) from observations. Panel (a) shows three random sample functions drawn from the prior distribution. A non-informative prior is specified so the average over functions has a zero mean. Panel (b) shows 3 random sample functions drawn from the posterior distribution after 4 training observations (dark black points). The effect of training is to constrain possible functions to only those that go through observation points. In panel (b), the shaded region represents plus and minus 2 standard deviations from the mean posterior prediction. Figure after Rasmussen and Williams [2006].

The first component of a Gaussian process is the mean function, which defines the mean of the infinite set of functions that are being considered. A common choice is to set the prior mean to zero, which can be thought of as a non-informative prior where the form of the function between inputs and outputs is unknown. This is demonstrated in Figure 5, as shown by the approximate mean of the sample prior functions being zero. As an alternative, this study follows the methodology of Timmermans [2015] who used a simple linear regression to obtain information about the mean function's form. Residual analysis of our data showed a cubic relationship for the tidal equilibrium argument terms, a somewhat expected result due to the cyclic nature of tides. Based on this result, the prior mean function was defined with a cubic term for all tidal equilibrium argument inputs and a linear term for all other inputs. A k-fold cross-validation (see section 4.2) was performed to evaluate emulator skill with and without the modified mean function. Results showed a significant gain in skill (both in terms of RMSE and Determination, R^2) by using the modified mean function.

The covariance function of a Gaussian process is the second necessary component for defining the emulator. The covariance function (often called the kernel) can be thought of as describing the relationship between points in the process. Practically this describes the smoothness of the resulting GPR. In general, the covariance function contains hyper-parameters describing the details of the relationship between points (e.g., parameters such as length-scale, signal variance, etc.). These parameters can be inferred from the data, which is commonly done through maximizing the marginal likelihood rather than full Bayesian inference [Schulz et al., 2018]. This was the approach used for this study.

A comparison of model performance was performed using 3 commonly used covariance functions: the Gaussian, squared exponential, and Matern [Rasmussen and Williams, 2006]. The Matern covariance function was tested with ν (a parameter controlling smoothing) equal to 1.5 and 2.5. The best performing model was evaluated using k-fold cross-validation and comparing model RMSE values [Kohavi, 1995; Arlot and Celisse, 2010]. K-fold cross-validation breaks the total training dataset into k segments and cycles through every possible combination of withholding one segment for validation and training the emulator with the remaining segments. This results in an ensemble of skill metrics for which the mean is less biased and more robust to the training period than a standard validation methodology [Arlot and Celisse, 2010]. It was found that the Matern (ν =2.5) performed the best and therefore was utilized for all results found in the following section. The training of the emulator was performed using the Managing Uncertainty in Complex Models (MUCM) package in R [Malde et al., 2016a].

4. Results

4.1 Error Introduced by Model Simplifications

With the construction of the emulator being hierarchical (Figure 2), it becomes important to assess the skill of the emulator at multiple simplification levels to determine where errors are being introduced. This was investigated for simulator simplifications by looking at the 5 largest storms, in terms of NTR, on record (Table 1). Storm events were chosen for this analysis as it is expected that the strong forcing gradients and rapidly changing dynamics of these events would provide the most robust test of simulator simplifications. For each storm, WLs were calculated using the Full Simulator (dynamic, non-simplified) and the Stationary Simulator (all simplifications except for emulation) to quantify the sum of level 2 and level 3 errors. This comparison was performed at six or seven (seven except for storm 2) temporally random points distributed across each storm. Figure 6 shows the difference between calculated WLs (level 1 simplification minus level 3 simplification) at two locations: the tide gauge (Figure 6a) and WL station 7 (Figure 6b). This difference is denoted here as the "error" resulting from simplifying the simulator. Two locations are plotted to visually sample how error is affected by location within the estuary.

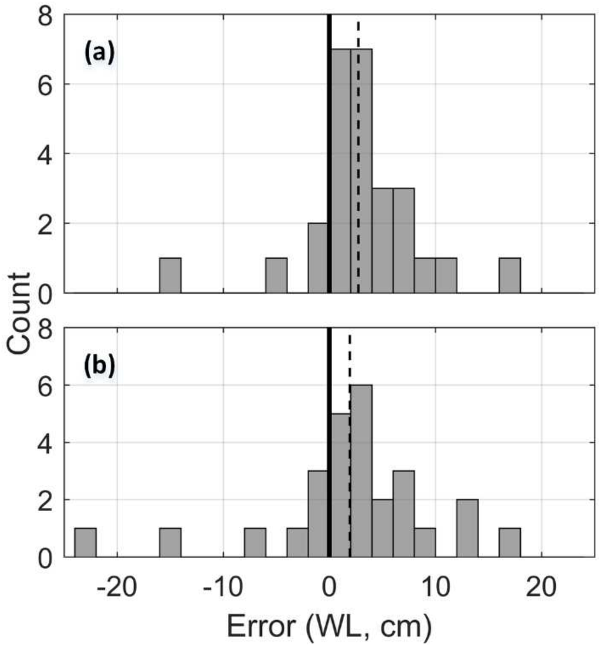

Figure 6: Error between Full and Stationary (Level 1 and Level 3) Simulator calculated WLs. The histogram includes comparisons for storms 1-5 (Table 1). The bold line represents zero error while the

dotted line is the mean error of the ensemble. Subplot (a) is computed at the tide gauge location while subplot (b) is at WL station 7 (located deeper within in the estuary; Figure 1).

The error computed via this test was found to have a max of 25 cm with a RMSE of 6 cm at the gauge location. The maximum RMSE for approximately 100 test stations randomly scattered across the estuary domain was found to be 9 cm. The Simplified Simulator was found to be only slightly biased with a mean approximately 2 cm lower than the Full Simulator (represented as a positive mean error in Figure 6). Additional snapshot runs were performed to examine the model's error for non-storm conditions (not shown). Results confirmed that the Simplified Simulator, on average, performs better for non-storm conditions, suggesting that the results in Figure 6 are likely conservative.

4.2 Emulator Validation

The ability of the emulator to replicate the Stationary Simulator (level 4 error) was quantified using a k-fold cross-validation. Figure 7 shows the results of the cross-validation with 5 segments comparing emulated WLs and simplified simulator WLs.

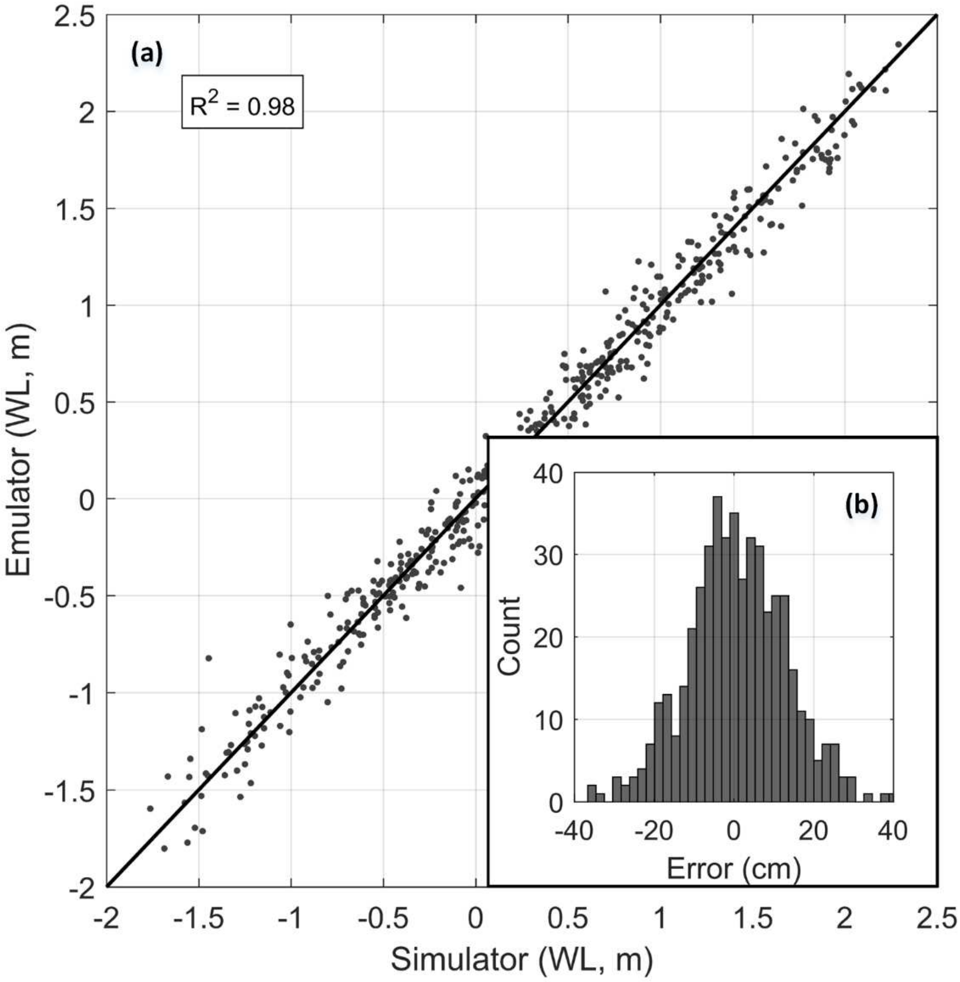

Figure 7: Panel (a): Simplified Simulator vs. emulator WLs for the full training dataset. The comparison was performed using a 5-fold validation procedure. Panel (b): histogram of the error between Simplified Simulator and emulator WLs.

Overall the emulator was found to perform well at this comparison level with a high level of skill. The emulator shows little bias (mean of the residuals is less than 1 cm), and relatively even variance in residuals across WL magnitude (Figure 7a). However, the width of the histogram in Figure 7b suggests that this step introduces more error than simulator simplifications. The RMSE was found to be around 13 cm (level 4 error) which is significantly larger than the calculated simulator simplification error (Figure 6, sum of level 2 and level 3 error) of approximately 6 cm. However, this comparison of RMSEs is imperfect as the level 4 error assessment is based on a larger sample size and more rigorous k-fold validation and the level 2 and level 3 error assessment only examined performance during storm events.

4.3 Emulator Performance: Westport, WA Tide Gauge

The next stage of quantifying the skill of the emulator is to compare emulated WLs to observations at the tide gauge. This test provides a measure of the cumulative level 4 error, or the total integrated error from predicting observed WLs using emulation. This analysis was performed by using the emulator to hindcast hourly WLs at the location of the tide gauge for the entire period of record (2006-2016). Comparison between tide gauge observations and hourly emulated WLs for a randomly chosen month long segment are shown in Figure 8. Overall, hourly emulated WLs (for the over 10-year long record) compare favorably to the tide gauge with an R^2 value of greater than 0.96, RMSE of approximately 15 cm, and a bias of less than 1 cm.

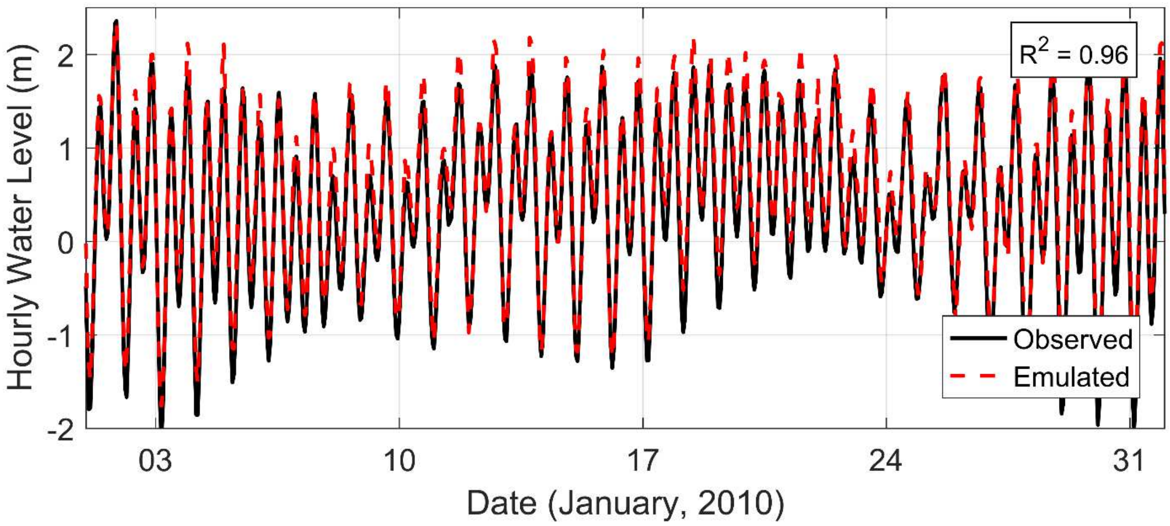

Figure 8: Comparison of emulated (red dashed line) and observed (black line) hourly WLs at the Grays Harbor tide gauge for January 2010. Coefficient of determination is calculated using the entire tidal record (2006 - 2016). WLs are plotted in reference to mean sea level.

As with tide gauge records, WL output from an emulator can be considered as the sum between two components: tides and NTRs. In the PNW, tides are the dominant source of WL variability [Allan and Komar, 2002a] and so the skill of the emulator in predicting WLs is primarily controlled by its ability to reproduce the deterministic tides. However, coastal hazard research often considers NTR individually as the driver of extreme WLs on top of regular, and well predicted, tidal cycles. Therefore, it is additionally important to test the emulator's skill at reproducing NTR signals. Furthermore, this provides a more robust test of emulator performance as NTR is not explicitly modeled as an output by the emulator.

NTRs are often calculated at tide gauges by subtracting the predicted tide (determined through harmonic analysis) from the measured WL. This procedure can be problematic since NTR may affect tidal phase, resulting in a false NTR signal from out of phase tidal signals [Pugh, 1996; Haigh et al., 2014; Serafin and Ruggiero; 2014]. Therefore, for this study NTRs were calculated from tide gauge data using the procedure of Serafin and Ruggiero [2014] (modeled after Bromirski et al., 2003), which uses spectral filtering to remove energy from tidal bands.

The Bromirski et al. [2003] methodology is not used to determine NTRs from the ADCSWAN simulations. Storm simulations are on the scale of weeks which is too short temporally to recover energy across all tide bands of interest in the frequency domain. Instead, NTRs from the ADCSWAN simulations (at all simplification levels) and emulator simulations were calculated as a full forcing run minus a tidal only run. Emulated NTRs were then subsequently smoothed with a loess filter to reduce noise associated with tidal phase mismatches between the tide-only and full forcing emulated time series.

Comparison of observed and hindcasted NTRs for storms 1-5 (Figure 9) show a good overall performance of the emulator. To contextualize this comparison, the Full, Simplified, and Stationary Simulator calculated NTR time series are all plotted. A quantitative comparison of error between observed and modeled NTRs found that all simplification levels (from full ADCSWAN to emulator) have an RMSE of approximately 14 cm plus or minus 1 cm. The similar error across all simplification levels suggest that the largest source of error for NTR is in Full Simulator itself (level 1). For example, in Figure 9b it is clear that the Full Simulator (red line) is unable to reproduce the peak NTR signal in storm 2.

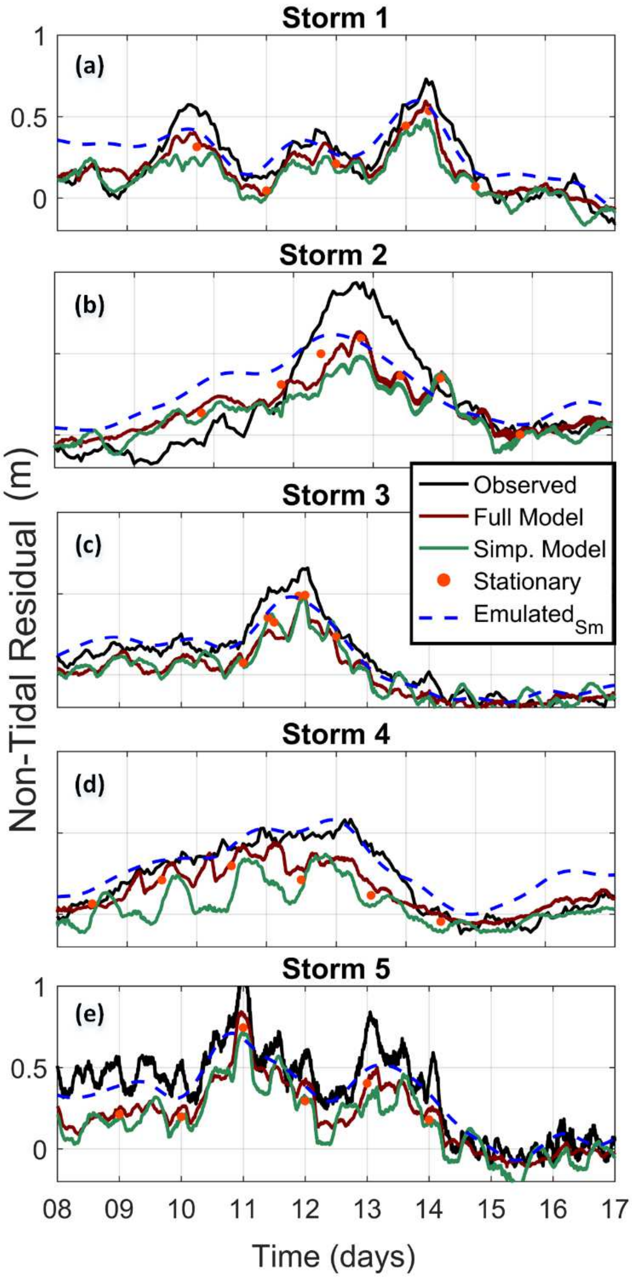

Figure 9: Tide Gauge comparison of observed and modeled NTR at various simplification levels. Each subpanel (a through e) is one of the top 5 storms of record (see Table 1). Due to windowing for spectral filtering, storm 5's observed NTR is calculated using the subtraction method rather than the Bromirski method. All panels have the same y-axis scaling. The specific dates on the x axes vary by storm, but each tick represents a day.

4.4 Emulator Performance: USACE Field Campaign

The tide gauge provides a rich dataset for validating the emulator due to its record length but is spatially limited to a single comparison point within the study area. One key strength of emulation, in comparison to a fully data driven methodology, is the ability to provide WL information across study sites where observational information may not be available. An emulator can be constructed at any location within the ADCSWAN model domain where output is provided. Figure 10 evaluates the spatial performance of emulation through a comparison of emulated and observed WL time series for a 1999 field campaign led by the USACE in Grays Harbor [Figure 1, Cialone et al., 2001; 2002].

Figure 10 shows good performance between observed and modeled WLs across most locations. The main exception to this is WL station 6 which displays comparatively poor agreement between the observed and emulated WLs. This lack of skill is equally shown by the Full Simulator and is therefore not a result of the emulation procedure. Table 2 gives RMSE values for a comparison between observations and modeled output at various levels of simplification (level 1 simulations and level 4 emulations are shown in Figure 10). The levels described in the column headers are cumulative error, or a comparison of the model at that simplification level (Figure 2) to observations. No level 3 skill estimate was developed due to the computational constraints of simulating sufficient stationary point runs to accurately quantify model skill.

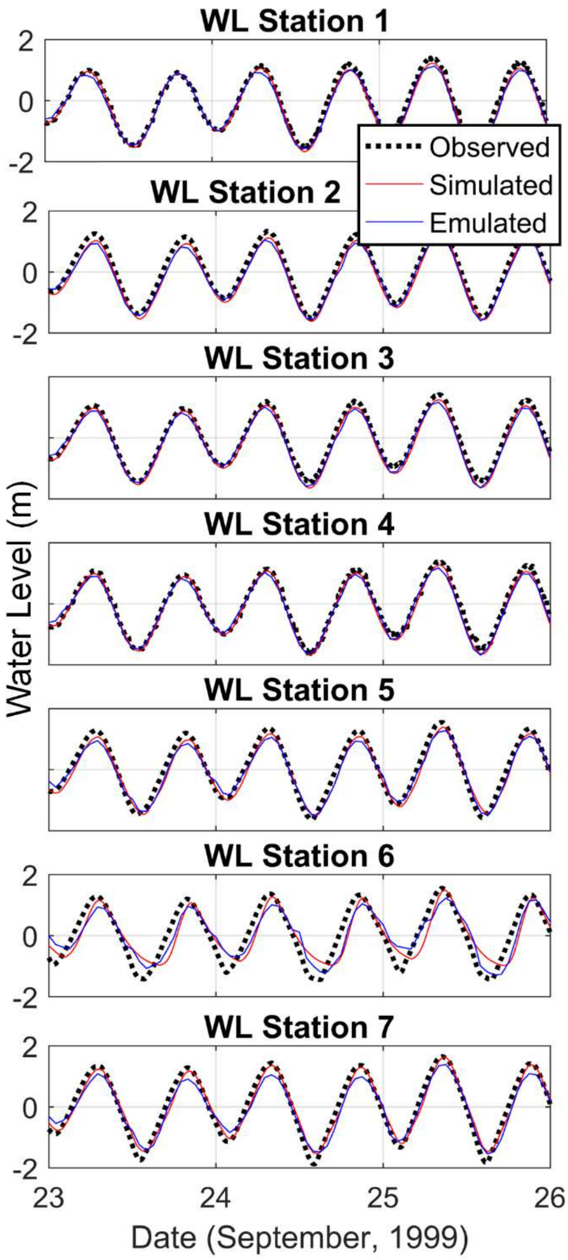

Figure 10: Comparison of observations (USACE deployments, Figure 1), Full Simulator and emulated WLs. All panels have the same y-axis and x-axis scaling.

Table 2 indicates that the largest drop in skill is at the level 1 simulator simplification. This corresponds to the full ADCSWAN model's inability to perfectly reproduce observations. Level 2 simplifications are found to only nominally impact modeled WLs with a small (1 cm) increase in RMSE for some stations. Level 4 simplifications additionally produce very little loss of skill.

Table 2: RMSE values comparing WL model output to observations at various simplification levels. Rows are locations / variables while the two column groupings represent instrument deployments. The length of the time series comparison varies depending on station deployment length.

	Deployment 1: RMSE (cm)			Deployment 2: RMSE (cm)		
Station	**Level 1**	**Level 2**	**Level 4**	**Level 1**	**Level 2**	**Level 4**
WL 1	24	24	25	21	21	22
WL 2	19	20	22	16	17	19
WL 3	26	26	27	37	37	29
WL 4	19	19	21	22	23	20
WL 5	21	22	22	19	19	20
WL 6	44	44	41	36	36	39
WL 7	21	22	28	23	21	27

5. Discussion

5.1 Effect of Simulator Simplifications

The hierarchical validation used in this study provides a unique approach to quantifying the error budget as sourced from multiple simplification levels. A comparison of model performance at various simplification levels (Table 2) found that the primary source of lost skill is from the Full Simulator rather than from simulator simplifications or from emulation. Averaging across stations and deployment periods, all simplifications and emulation only increased RMSE by 1 cm relative to the level 1 error.

This result is of particular interest when compared to the quantification of discrete error from only emulation (see the histogram in Figure 7) which shows that emulation introduces comparatively significant error into WL estimates. A cumulative level 4 comparison at the tide gauge (Figure 8) found a RMSE of 15 cm while Level 4 itself (Figure 7) had a RMSE of 13 cm. This result suggests that the error from each simplification during emulator construction is not independent. In other words, the cumulative error variance is not the sum of the discrete error variances. Practically, the dominance of level 1 error is found to mask that of the other levels.

This may not be true if level 1 error was able to be significantly reduced by improvements in process-based modeling at which point other simplifications may become relevant to the error budget.

In terms of quantifying model skill, the RMSE for comparisons to the 1999 USACE field data is overall larger than the RMSE in comparison to the tide gauge. A close examination of the USACE WL time series shows significant high frequency noise that is likely the cause of the overall larger RMSE values. The USACE data exhibits more high frequency variability due to a shorter averaging period (6 minutes for the tide gauge and 3 minutes for the USACE data).

This study suggests that the most effective action to improve emulated WL predictions is to reduce level 1 error. One option would be optimized tuning, a process which can be accomplished by including tuning parameters within the emulator framework [Kennedy et al., 2006; Hall et al., 2011;]. ADCSWAN could also be replaced with a different simulator or physics implementation, for example ADCSWAN in 3D baroclinic mode. This would come at the cost of drastically increasing computation time and requiring additional input dimensionality in the form of density, temperature, and salinity fields.

A level 1 error reduction could also be accomplished through improving the quality of model input data, both in terms of bathymetry and forcing. Incorrect bathymetry is likely the source of errors for WL station 6. Figure 10 shows WLs at this station having an asymmetric tidal signal indicative of shallow water while the observations have less asymmetry. This suggests that the water depth at the time of deployment was greater than the depth in the compiled bathymetric dataset used to generate the ADCSWAN grid. Therefore, investment in more accurate or more recent bathymetry is another viable step for decreasing level 1 error. Similarly, level 1 error integrates error as a function of poor-quality forcing, making improvements to forcing another promising avenue for error reduction.

Level 2 error could be reduced by making less aggressive simplifications of forcing inputs. It is conceptually straight forward to include other input dimensionality such as spatially variable atmospheric forcing or full spectral wave forcing. A promising strategy for including field variables as input dimensions is through decomposing the field into principal components [Higdon et al., 2008; Liu and Guillas, 2017].

There are additionally a range of options for avoiding the stationarity assumption made in this study, which would eliminate or reduce level 3 error. The incorporation of temporal variability in emulators is reviewed by Reichert et al. [2011] who suggest the following strategies:

1) Apply a standard emulation methodology but with time as an additional degree of dimensionality. [Conti et al., 2005]
2) Describe the time series using basis functions and then apply emulation to the basis function coefficients. [Bayarri et al., 2007; Higdon et al., 2008]
3) Emulate the difference from one time point to the next. [Bhattacharya, 2007; Conti et al., 2009]
4) Use a Gaussian stochastic process as a Bayesian prior. [Liu and West, 2009]
5) Develop a hybrid dynamic/emulated model, or a "Mechanistic dynamic emulator." [Reichert et al., 2011; Albert C., 2012]

In the context of this study, strategy 1 is conceptually the simplest but it is not clear a priori how far into the past the system's memory extends and each included time step multiplies the dimensionality of the input space. Strategy 2 is complicated by identifying basis functions that adequately capture the various contributing signals. For example, a Fourier transformation is a natural solution except that storm surge is non-periodic and an important contributor to estuarine WLs. Strategies 3-5 all have potential advantages but bring additional complexity to an already complex methodology so were not explored further.

For reducing level 4 error, GPR is a flexible framework and there are likely gains to be made through a more exhaustive approach for emulator specification. In particular, handling of the periodic nature of tides within the covariance function [Roberts et al., 2013] is a promising research direction.

5.2 Computational Cost Considerations

Emulation is an approach to dramatically reduce simulation times and is therefore most valuable in situations where the simulator must be run for very long periods or for multiple iterations (e.g., probabilistic risk assessment). Emulation requires an upfront cost, through the running of multiple simulations to construct a training dataset, but is comparatively instantaneous after this initial investment. As the nature of the trade-off is computation time, it is useful to review the costs of building the training dataset.

The first control on computational cost for the training set is the number of input dimensions. The simplifications implemented in this study managed to reduce the input space to 16 dimensions. Each design point took approximately 66 core hours to run in parallel on a server with Intel Xeon E52450 CPUs (2.1 GHz). With this setup, a full experimental design of 160 points would require over 10.5 thousand core hours (although with parallelization the actual time was much less). This study developed a larger experimental design (over 400 points) but this was primarily for validation rather than emulator skill (see discussion below). Full ADCSWAN required approximately 18 core hours per day of simulation time. Based off these computational costs, emulation becomes an efficient option if approximately one and a half years of simulation are required. This limit is highly situationally dependent and is controlled by processor speed, simulator, emulator, etc. and is intended only as an order of magnitude reference. Furthermore, emulation is primarily targeted at probabilistic methodologies, rather than hindcasting, for which multiple iterations of time series quickly sum to very large total simulation times.

The above analysis is based on a LHS design and the Loeppky et al. [2009] guideline that a training dataset should be around 10 times the number of input dimensions. However, LHS is one of many possible experimental designs [Levy and Steinberg, 2010]. Significant research has focused on optimizing experimental designs beyond LHS and it is possible that a more complex design could reduce the size of the training dataset. For example, LHS does not consider the probability that a particular combination of input parameters may occur. Therefore, some design points are likely poorly utilized exploring space that is physically impossible or highly improbable (for example, high wave heights associated with low wave periods).

Finally, the above analysis did not consider the effect of training dataset size on skill. This relationship was tested by quantifying emulator performance at a variety of training dataset sizes. For this analysis, the total body of simulations (480) was partitioned into smaller dataset sizes ranging from 50 to 450 simulations for testing. For each smaller dataset, a k-fold validation with 5 segments was performed (Figure 11) to quantify emulator skill at this smaller training dataset size. This analysis is identical to that described in section 4.2 but with an artificially decreased training dataset size.

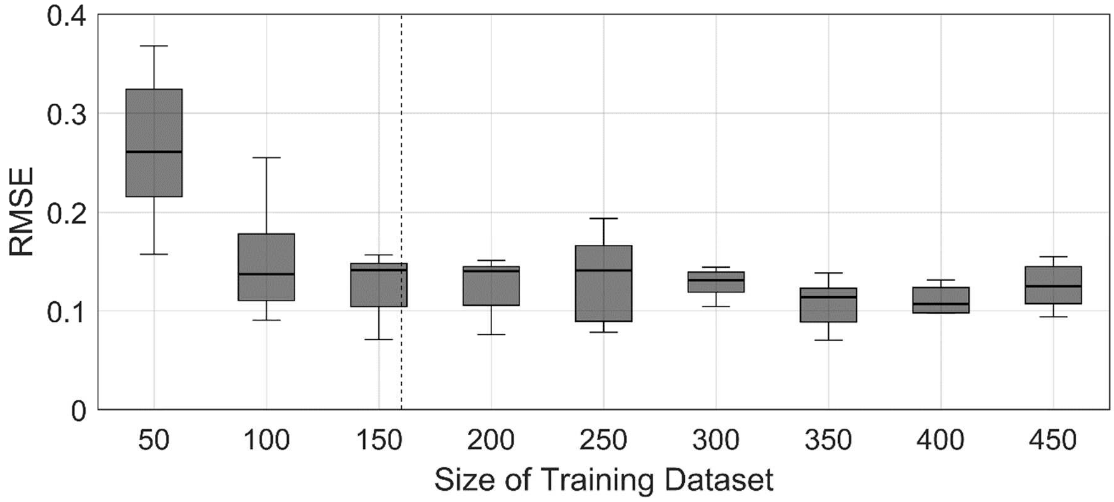

Figure 11: Impact of training dataset size on emulator skill. RMSE is calculated with a 5-fold cross-validation and represented by a standard boxplot. The dotted vertical line represents the theoretical training dataset size from Loeppky et al. [2009].

Results from this analysis are in good agreement with the guidance of Loeppky et al. [2009] in that ten times the number of input dimensions is sufficient for building a skillful emulator (Figure 11). Beyond this limit, only very small gains in skill are realized, suggesting that it is not efficient to over build the training dataset.

It is worth considering the cumulative computational cost of developing multiple emulators. This study takes the approach of building individual emulators at each location of interest. While emulator training and simulation is rapid for individual emulators, the sum computational cost of constructing many emulators can be significant. This is especially true considering that large estuarine hydrodynamic models can be of very high output dimensionality (the utilized Grays Harbor ADCSWAN grid has over 29,000 nodes). A common solution is to dimensionally reduce model outputs via approaches such as principal component analysis [Chen et al., 2011; Jia and Taflanidis, 2013; Jia et al., 2016; Bass and Bedient, 2017]. An alternative option uses the multivariate Gaussian process to generalize the standard GPR case to a “multi-output emulator” [Conti and O’Hagan, 2010; Fricker et al., 2010]. While not considered in this study, which is primarily concerned with point assessments, these approaches could result in significant computational savings for a larger output dimensionality. Further, considering emulators individually implicitly assumes independence of output variables and ignores the inherent correlation between output variables [Rasmussen and Williams, 2006].

5.3 Emulation Beyond Water Levels

While this study has focused primarily on emulating WLs, emulation can easily be extended to other variables in a coastal hazards framework. To explore this possibility, Hs was emulated at the observational Hs stations from the 1999 USACE field campaign (Figure 1) [Cialone et al., 2001; 2002]. Hs emulators were developed using an identical approach to that of WLs except that Hs emulation was found to not need cubic terms for the prior mean function. A comparison to observations (Figure 12) shows that GPR emulators perform well for Hs with the highest Hs (around September 26, 1999) being well reproduced by the emulator at stations 0, 1, 2, 3, and 4 (Figure 12). Performance is comparatively poor at stations 5 and 6, which are further within the estuary and less influenced by offshore waves. These results are further quantified in Table 3 which shows poor skill for interior Hs stations. It should be noted, however, that the Hs signals at these two stations have low variance and are barely above the noise floor.

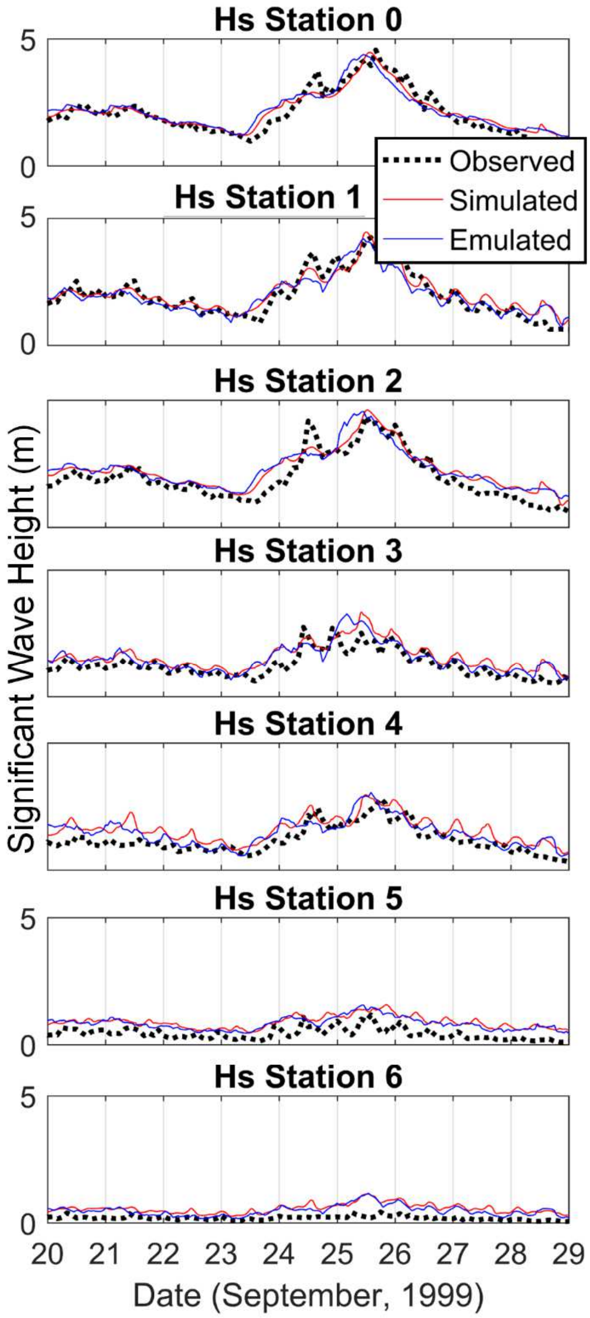

Figure 12: Comparison of observations (USACE deployments), Full Simulator and emulated Hs time series. Symbols are used for the observations for the sake of visual clarity. All panels have the same y-axis scaling.

Results show that, similarly to WLs, the largest loss of skill is at simplification level 1. Simplified Simulator and emulator results are found to closely track the Full Simulator. This is most evident at the bay interior stations where the Full Simulator and emulator are both found to over-predict Hs. Table 3 reveals that level 2, 3 and 4 simplifications produce little loss of skill for calculated Hs (average increase in RMSE of 2 cm). An exception to this is Hs station 5 which shows a significant increase in RMSE at the emulation level. The cause of poor emulator performance at this one location is unclear but it is likely due to a poor emulator model fit.

Table 3: RMSE values comparing Hs model output to observations at various simplification levels. Rows represent station locations while the two column groupings represent instrument deployments. Hs station 0/7 was relocated so deployment 1 values represent the Hs station 0 location and deployment 2 values representing the station 7 location.

	Deployment 1: RMSE (cm)			Deployment 2: RMSE (cm)		
Station	**Level 1**	**Level 2**	**Level 4**	**Level 1**	**Level 2**	**Level 4**
Hs 0/7	32	32	37	73	73	75
Hs 1	38	39	37	61	61	69
Hs 2	45	46	51	48	49	61
Hs 3	45	45	35	52	52	50
Hs 4	49	50	40	148	148	42
Hs 5	47	47	104	136	136	209
Hs 6	35	35	28	40	40	38

Overall these results suggest that emulation could be integrated into many parts of an estuary modeling system. However, a key assumption of emulation with GPR is smoothness in response characteristics, suggesting that GPR may be sub-optimal for “jumpy” variables. Not shown are similar results for Tp which can exhibit discontinuities within estuaries as Tp switches from one wave spectrum component to another. The emulator is qualitatively able to capture Tp characteristics but cannot resolve these instantaneous jumps. For this reason, it is important to carefully consider the form of the output variable being emulated and its relation to the input parameters.

5.4 Example Emulator Application: Extreme Water Level Decomposition

Outside of validation, it is illustrative to explore an example application of emulation. For this purpose, a decomposition of the relative forcing contributions responsible for extreme WLs within Grays Harbor was performed. Seven versions of a 31-year time series (1984-2015) were emulated under different forcing scenarios. As a baseline, a "full forcing" case time series was emulated with the observed forcing at Grays Harbor (comparable to the hindcasts in Figure 8 and Figure 10). Each additional forcing scenario was emulated with one forcing contribution excluded (waves, wind, pressure, base WL, streamflow, and tides) to isolate the relative contribution of individual forcings to WLs. A particular forcing contribution was calculated as the emulator output WL with full forcing minus the emulator output WL with all forcing except the component of interest. The exception to this is tides (which cannot be turned off due to how they are included in the emulator) which were calculated simply as emulator output with no other forcing but tides. WL contributions were calculated at the time of the 31 annual maximum WL events as determined by the full forcing time series. The average relative contribution of each forcing component over the 31 annual maxima are plotted along East-West and North-South transects in Figure 13.

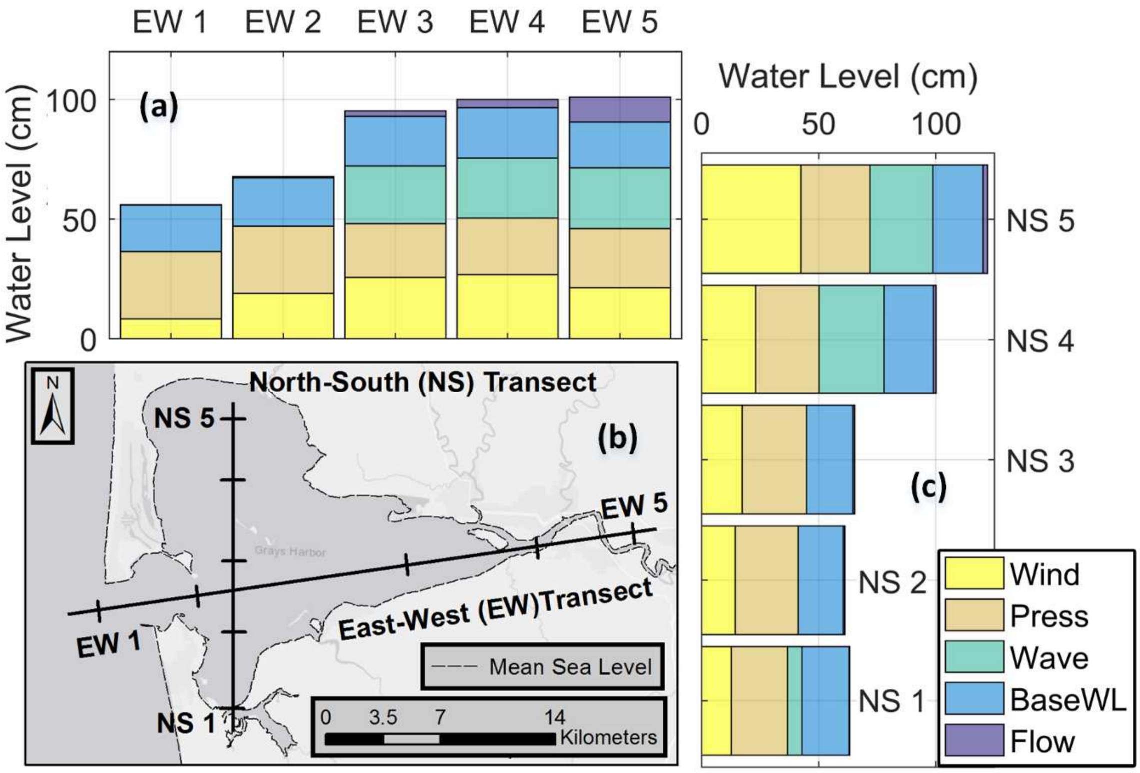

Figure 13: Average WL contribution from forcing components during extreme events (maximum annual WLs). Two transects are plotted with subplot (b) showing plotted transects, (East-West, EW) and (North-South, NS), with station locations marked as ticks. Tick locations are approximate (within 1 km) to scattered station locations. Subplot (a) is the East-West transect and subplot (c) is the North-South transect.

The diverse mix of contributions for each bar in Figure 13 shows that extreme WL events are compound in nature. This conclusion is further supported by the variance in emulated extreme WL contributions (not shown) which reveals that the composition of each individual annual maximum event varies widely across the timeseries. The mean contribution of each forcing is found to be significant providing evidence that all included forcing processes are important for properly quantifying extreme water levels. The only exception is streamflow which is found to be nominally important except near the streamflow boundary. This result is likely specific to the Grays Harbor estuary and would be different for a more hydrologically dominated estuary system [Svensson and Jones, 2004; Lavery and Donovan, 2005; Chen et al., 2014].

The mix of contributions is found to be spatially variable across the estuary domain, leading to both an East-West and North-South gradient in contributions to WLs. For example, the streamflow contribution is found to increase moving west towards the estuary's streamflow inlet. Wave influence is found to have a significant contribution to the annual maxima but only at stations in the northern and eastern reaches of the bay. This result is likely due to breaking induced setup not occurring at the bay's entrance channel. The influence of wind increases to the north, due to the mean wind direction emanating from the south during storm events. The influence of pressure anomalies on extreme WLs is found to be uniform but this result is likely from the spatial simplification of sea level pressure fields.

Not shown in Figure 13 is the contribution from tidal forcing. This is primarily for scale reasons as the tidal component is an order of magnitude larger than that from other forcing (average of 140 cm). Tides also show a gradient across the estuary although with the opposite pattern as that shown in Figure 13. The tidal component of annual maxima WLs decreases by about 30 cm moving from the center of the estuary moving North or East. As WLs are the sum of these two components (tides and forcing driven NTR), the calculated gradient in total WLs is less than that shown in Figure 13 (under 20 cm across the two transects).

6. Conclusions

This paper has presented an application of emulation, or surrogate modeling, to the problem of rapidly simulating hydrodynamic variables within the Grays Harbor, WA estuary. This methodology is targeted towards a variety of computationally constrained problems including probabilistic modeling, uncertainty quantification, model optimization, and non-parametric extreme event analysis. To facilitate efficiently achieving these goals, this study has focused on validating and quantifying the error induced by emulation. Additionally, a variety of simplifications to the simulator have been suggested for reducing input dimensionality, and therefore the size of the emulator training dataset.

The results from this study suggest that the Gaussian Process regression (GPR) derived emulator is skillful for calculating a variety of model output variables (WL, NTR, and Hs). A decadal-scale comparison of emulated WLs to tide gauge data showed the emulator having a RMSE of 15 cm. Emulator performance is evaluated at multiple observation points across the estuary domain providing confidence that emulation is skillful across spatial extents. Decomposing the error from different emulator construction simplification levels shows that the largest source of unexplained variance in emulator hindcasts is from ADCSWAN itself. Of particular interest, strong simulator simplifications (including that of stationarity) are a relatively low contributor to losses in emulator performance (average increase in RMSE of 1 cm). Therefore, future efforts to improve emulator performance should focus on improving the Full Simulator before reducing simplifications or optimizing the emulator.

Emulation is additionally found to be very efficient after the construction of the training dataset. Using an LHS experimental design, analysis shows that the training dataset size guidance of 10 times the number of input dimensions [Loeppky et al., 2009] is optimal in the case examined here. Overall emulation is found to have the same order of magnitude skill as process-based models as well as showing significant gains in computational efficiency. Therefore, emulation is shown to be a viable path for exploring estuarine hydrodynamic modeling problems.

Finally, the emulator was applied to investigate the relative contributions of different forcing variables to annual maxima WLs and NTR at the study site. Results show a diverse mix of forcing contributing to annual extreme WLs, indicating the importance of considering compound events for flood hazard assessments in the PNW. All forcing components, along with WL itself, were found to exhibit significant spatial variability hinting at important information for flood vulnerability assessment. Tides were found to be the largest contributor to extreme WLs with other components being of the same order of magnitude. The exception to this is streamflow which was found to be, on average, a relatively minor contributor to extremes except near the river's mouth. Additionally, waves were found to only contribute to WLs at stations near the edge of the estuary domain, a result that is likely tied to wave penetration into the estuary. While only a single example application of emulation to estuary hydrodynamics questions was explored, results signal the significant potential of emulation to a broad range of applications.

Acknowledgments

Tide gauge records are available through the National Oceanic and Atmospheric Administration (NOAA) National Ocean Service (NOS) website. River discharge is available from the USGS through the National Water Information System. The NARR climate dataset is available through NOAA's Earth System Research Laboratory website. Bathymetric and topographic data were obtained from NOAA's Bathymetric Data viewer (DEMs) and DOGAMI's LiDAR download portal. We thank Melisa Menendez and Jorge Perez at the

Environmental Hydraulics Institute of the Universidad de Cantabria (IH Cantabria) for providing the Global Ocean Wave 2 (GOW2) data. We also thank Mary Cialone and Dave Michalson of the U.S. Army Corps of Engineers for providing the Grays Harbor field observations from 1999.This work was funded by the NOAA Regional Integrated Sciences and Assessments Program (RISA) [Grant Number NA15OAR4310145] and a contracted grant with the Quinault Treaty Area (QTA) tribal governments (Quinault Indian Nation, Hoh Indian Tribe, and Quileute Tribe).

References

Albert, C., 2012. A mechanistic dynamic emulator. Nonlinear Anal. Real World Appl. 13, 2747–2754. https://doi.org/10.1016/j.nonrwa.2012.04.003

Allan, J.C., Komar, P.D., 2006. Climate Controls on US West Coast Erosion Processes. J. Coast. Res. 223, 511–529. https://doi.org/10.2112/03-0108.1

Allan, J.C., Komar, P.D., 2002. Extreme Storms on the Pacific Northwest Coast during the 1997-98 El Niño and 1998-99 La Niña. J. Coast. Res. https://doi.org/10.2307/4299063

Allan, J.C., Komar, P.D., 2002. Wave Climate Change and Coastal Erosion in the US Pacific Northwest, in: Ocean Wave Measurement and Analysis. American Society of Civil Engineers, Reston, VA, pp. 680–689. https://doi.org/10.1061/40604(273)70

Allan, J.C., Komar, P.D., Ruggiero, P., 2011. Storm Surge Magnitudes and Frequency on the Central Oregon Coast, in: Solutions to Coastal Disasters 2011. American Society of Civil Engineers, Reston, VA, pp. 53–64. https://doi.org/10.1061/41185(417)6

Anjyo, K., Lewis, J.P., 2011. RBF interpolation and Gaussian process regression through an RKHS formulation. J. Math-for-Industry 3, 63–71.

Apel, H., Merz, B., Thieken, A.H., 2008. Quantification of uncertainties in flood risk assessments. Int. J. River Basin Manag. 6, 149–162. https://doi.org/10.1080/15715124.2008.9635344

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. https://doi.org/10.1214/09-SS054

Bass, B., Bedient, P., 2018. Surrogate modeling of joint flood risk across coastal watersheds. J. Hydrol. 558, 159–173. https://doi.org/10.1016/j.jhydrol.2018.01.014

Bayarri, M.J., Berger, J.O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R.J., Paulo, R., Sacks, J., Walsh, D., 2007. Computer Model Validation with Functional Output. Ann. Stat. 35, 1874–1906. https://doi.org/10.1214/009053607000000163

Bhaskaran, P.K., Nayak, S., Bonthu, S.R., Murty, P.L.N., Sen, D., 2013. Performance and validation of a coupled parallel ADCIRC–SWAN model for THANE cyclone in the Bay of Bengal. Environ. Fluid Mech. 13, 601–623. https://doi.org/10.1007/s10652-013-9284-5

Bhattacharya, S., 2007. A simulation approach to Bayesian emulation of complex dynamic computer models. Bayesian Anal. 2, 783–815. https://doi.org/10.1214/07-BA232

Blain, C.A., Preller, R.H., Rivera, A.P., 2001. Tidal Prediction Using the Advanced Circulation Model (ADCIRC) and a Relocatable PC-based System. Oceanography 15.

Blain, C.A., Rogers, W.E., 1998. Coastal Tide Prediction Using the ADCIRC-2DDI Hydrodynamic Finite Element Model: Model Validation and Sensitivity Analyses in the Southern North Sea/English Channel.

Bode, L., Hardy, T.A., 1997. Progress and Recent Developments in Storm Surge Modeling. J. Hydraul. Eng. 123, 315–331. https://doi.org/10.1061/(ASCE)0733-9429(1997)123:4(315)

Booij, N., Holthuijsen, L.H., Ris, R.C., 1997. The “SWAN” Wave Model for Shallow Water, in: Coastal Engineering 1996. American Society of Civil Engineers, New York, NY, pp. 668–676. https://doi.org/10.1061/9780784402429.053

Bromirski, P.D., Flick, R.E., Cayan, D.R., 2003. Storminess Variability along the California Coast: 1858–2000. J. Clim. 16, 982–993. https://doi.org/10.1175/1520-0442(2003)016<0982:SVATCC>2.0.CO;2

Bunya, S., Dietrich, J.C., Westerink, J.J., Ebersole, B.A., Smith, J.M., Atkinson, J.H., Jensen, R., Resio, D.T., Luettich, R.A., Dawson, C., Cardone, V.J., Cox, A.T., Powell, M.D., Westerink, H.J., Roberts, H.J., 2010. A High-Resolution Coupled Riverine Flow, Tide, Wind, Wind Wave, and Storm Surge Model for Southern Louisiana and Mississippi. Part I: Model Development and Validation. Mon. Weather Rev. 138, 345–377. https://doi.org/10.1175/2009MWR2906.1

Carnell, R., 2017. LHS: Latin Hypercube Samples.

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2012. Data-driven dynamic emulation modelling for the optimal management of environmental systems. Environ. Model. Softw. 34, 30–43. https://doi.org/10.1016/j.envsoft.2011.09.003

Chen, T., Hadinoto, K., Yan, W., Ma, Y., 2011. Efficient meta-modelling of complex process simulations with time–space-dependent outputs. Comput. Chem. Eng. 35, 502–509. https://doi.org/10.1016/J.COMPCHEMENG.2010.05.013

Chen, W.-B., Liu, W.-C., Chen, W.-B., Liu, W.-C., 2014. Modeling Flood Inundation Induced by River Flow and Storm Surges over a River Basin. Water 6, 3182–3199. https://doi.org/10.3390/w6103182

Cheng, T.K., Hill, D.F., Beamer, J., García-Medina, G., 2015. Climate change impacts on wave and surge processes in a Pacific Northwest (USA) estuary. J. Geophys. Res. Ocean. 120, 182–200. https://doi.org/10.1002/2014JC010268

Cheng, T.K., Hill, D.F., Read, W., 2015. The Contributions to Storm Tides in Pacific Northwest Estuaries: Tillamook Bay, Oregon, and the December 2007 Storm. J. Coast. Res. 313, 723–734. https://doi.org/10.2112/JCOASTRES-D-14-00120.1

Cialone, M.A., Kraus, N.C., 2001. Engineering Study of Inlet Entrance Hydrodynamics: Grays Harbor, Washington, USA, in: Coastal Dynamics ’01. American Society of Civil Engineers, Reston, VA, pp. 413–422. https://doi.org/10.1061/40566(260)42

Cialone, M.A., Militello, A., Brown, M.E., Kraus, N.C., 2002. Coupling of Wave and Circulation Numerical Models at Grays Harbor Entrance, Washington, USA, in: Proceedings 28th Coastal Engineering Conference. World Scientific Publishing Company, pp. 1279–1291. https://doi.org/10.1142/9789812791306_0108

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. J. Hydrol. 375, 613–626. https://doi.org/10.1016/J.JHYDROL.2009.06.005

Conti, S., Gosling, J.P., Oakley, J.E., O ’Hagan, A., 2009. Gaussian process emulation of dynamic computer codes. Biometrika 96, 663–676. https://doi.org/10.1093/biomet/asp028

Conti, S., Anderson, C.W., Kennedy, M.C., O ’Hagan, A., 2005. A Bayesian Analysis of Complex Dynamic Computer Models. Sensit. Anal. Model Output.

Conti, S., O’Hagan, A., 2010. Bayesian emulation of complex multi-output and dynamic computer models. J. Stat. Plan. Inference 140, 640–651. https://doi.org/10.1016/j.jspi.2009.08.006

Dale, M., Wicks, J., Mylne, K., Pappenberger, F., Laeger, S., Taylor, S., 2014. Probabilistic flood forecasting and decision-making: an innovative risk-based approach. Nat. Hazards 70, 159–172. https://doi.org/10.1007/s11069-012-0483-z

Davis, J.R., Paramygin, V. a., Forrest, D., Sheng, Y.P., 2010. Toward the Probabilistic Simulation of Storm Surge and Inundation in a Limited-Resource Environment. Mon. Weather Rev. 138, 2953–2974. https://doi.org/10.1175/2010MWR3136.1

Dawson, R.J., Hall, J.W., Bates, P.D., Nicholls, R.J., 2005. Quantified Analysis of the Probability of Flooding in the Thames Estuary under Imaginable Worst-case Sea Level Rise Scenarios Quantified Analysis of the Probability of Flooding in the Thames Estuary under

Imaginable Worst-case Sea Level Rise Scenarios. Int. J. Water Resour. Dev. 21, 577–591. https://doi.org/10.1080/07900620500258380

Di Baldassarre, G., Schumann, G., Bates, P.D., Freer, J.E., Beven, K.J., 2010. Flood-plain mapping: a critical discussion of deterministic and probabilistic approaches. Hydrol. Sci. J. 55, 364–376. https://doi.org/10.1080/02626661003683389

Dietrich, J.C., Bunya, S., Westerink, J.J., Ebersole, B.A., Smith, J.M., Atkinson, J.H., Jensen, R., Resio, D.T., Luettich, R.A., Dawson, C., Cardone, V.J., Cox, A.T., Powell, M.D., Westerink, H.J., Roberts, H.J., Dietrich, J.C., Bunya, S., Westerink, J.J., Ebersole, B.A., Smith, J.M., Atkinson, J.H., Jensen, R., Resio, D.T., Luettich, R.A., Dawson, C., Cardone, V.J., Cox, A.T., Powell, M.D., Westerink, H.J., Roberts, H.J., 2010. A High-Resolution Coupled Riverine Flow, Tide, Wind, Wind Wave, and Storm Surge Model for Southern Louisiana and Mississippi. Part II: Synoptic Description and Analysis of Hurricanes Katrina and Rita. Mon. Weather Rev. 138, 378–404. https://doi.org/10.1175/2009MWR2907.1

Dietrich, J.C., Tanaka, S., Westerink, J.J., Dawson, C.N., Luettich, R.A., Zijlema, M., Holthuijsen, L.H., Smith, J.M., Westerink, L.G., Westerink, H.J., 2012. Performance of the Unstructured-Mesh, SWAN+ADCIRC Model in Computing Hurricane Waves and Surge. J. Sci. Comput. 52, 468–497. https://doi.org/10.1007/s10915-011-9555-6

Dietrich, J.C., Zijlema, M., Westerink, J.J., Holthuijsen, L.H., Dawson, C., Luettich, R.A., Jensen, R.E., Smith, J.M., Stelling, G.S., Stone, G.W., 2011. Modeling hurricane waves and storm surge using integrally-coupled, scalable computations. Coast. Eng. 58, 45–65. https://doi.org/10.1016/J.COASTALENG.2010.08.001

DOGAMI, 2010. Lidar Remote Sensing Data Collection: Southwest Washington.

Dushaw, B.D., Egbert, G.D., Worcester, P.F., Cornuelle, B.D., Howe, B.M., Metzger, K., 1997. A TOPEX/POSEIDON global tidal model (TPXO.2) and barotropic tidal currents determined from long-range acoustic transmissions. Prog. Oceanogr. 40, 337–367. https://doi.org/10.1016/S0079-6611(98)00008-1

Engle, V.D., Kurtz, J.C., Smith, L.M., Chancy, C., Bourgeois, P., 2007. A Classification of U.S. Estuaries Based on Physical and Hydrologic Attributes. Environ. Monit. Assess. 129, 397–412. https://doi.org/10.1007/s10661-006-9372-9

Funakoshi, Y., Hagen, S.C., Bacopoulos, P., 2008. Coupling of Hydrodynamic and Wave Models: Case Study for Hurricane Floyd (1999) Hindcast. J. Waterw. Port, Coastal, Ocean Eng. 134, 321–335. https://doi.org/10.1061/(ASCE)0733-950X(2008)134:6(321)

Ganju, N.K., Brush, M.J., Rashleigh, B., Aretxabaleta, A.L., Del Barrio, P., Grear, J.S., Harris, L.A., Lake, S.J., Mccardell, G., O'donnell, J., Ralston, D.K., Signell, R.P., Testa, J.M., Vaudrey, J.M.P., 2015. Progress and Challenges in Coupled Hydrodynamic-Ecological Estuarine Modeling. Estuaries and Coasts. https://doi.org/10.1007/s12237-015-0011-y

Gano, S., Kim, H., Brown, D., 2006. Comparison of Three Surrogate Modeling Techniques: Datascape, Kriging, and Second Order Regression, in: 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference. American Institute of Aeronautics and Astronautics, Reston, Virigina. https://doi.org/10.2514/6.2006-7048

Gouldby, B., Méndez, F.J., Guanche, Y., Rueda, A., Mínguez, R., 2014. A methodology for deriving extreme nearshore sea conditions for structural design and flood risk analysis. Coast. Eng. 88, 15–26. https://doi.org/10.1016/J.COASTALENG.2014.01.012

Green, C., Viavattene, C., Thompson, P., 2011. Guidance for assessing flood losses CONHAZ Report.

Haigh, I.D., Wijeratne, E.M.S., MacPherson, L.R., Pattiaratchi, C.B., Mason, M.S., Crompton, R.P., George, S., 2014. Estimating present day extreme water level exceedance probabilities around the coastline of Australia: tides, extra-tropical storm surges and mean sea level. Clim. Dyn. 42, 121–138. https://doi.org/10.1007/s00382-012-1652-1

Hall, J.W., Manning, L.J., Hankin, R.K.S., 2011. Bayesian calibration of a flood inundation model using spatial data. Water Resour. Res. 47, W05529. https://doi.org/10.1029/2009WR008541

Higdon, D., Gattiker, J., Williams, B., Rightley, M., 2008. Computer Model Calibration Using High-Dimensional Output. J. Am. Stat. Assoc. 103, 570–583. https://doi.org/10.1198/016214507000000888

Jia, G., 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. Eng. 261–262, 24–38. 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., 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. https://doi.org/10.1007/s11069-015-2111-1

Jin, R., Chen, W., Simpson, T.W., 2001. Comparative studies of metamodelling techniques under multiple modelling criteria. Struct. Multidiscip. Optim. 23, 1–13. https://doi.org/10.1007/s00158-001-0160-4

Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. J. Stat. Plan. Inference 26, 131–148. https://doi.org/10.1016/0378-3758(90)90122-B

Jones, B., Johnson, R.T., 2009. Design and analysis for the Gaussian process model. Qual. Reliab. Eng. Int. 25, 515–524. https://doi.org/10.1002/qre.1044

Kantha, L.H., Clayson, C.A., 2000. Numerical models of oceans and oceanic processes., Vol. 66. ed. Elsevier.

Kennedy, M.C., Anderson, C.W., Conti, S., O'Hagan, A., 2006. Case studies in Gaussian process modelling of computer codes. Reliab. Eng. Syst. Saf. 91, 1301–1309. https://doi.org/10.1016/J.RESS.2005.11.028

Kim, S.W., Melby, J. a., Nadal-Caraballo, N.C., Ratcliff, J., 2015. A time-dependent surrogate model for storm surge prediction based on an artificial neural network using high-fidelity synthetic hurricane modeling. Nat. Hazards 76, 565–585. https://doi.org/10.1007/s11069-014-1508-6

Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in: International Joint Conference on Artificial Intelligence.

Krien, Y., Dudon, B., Roger, J., Zahibo, N., 2015. Probabilistic hurricane-induced storm surge hazard assessment in Guadeloupe, Lesser Antilles. Nat. Hazards Earth Syst. Sci. 15, 1711–1720. https://doi.org/10.5194/nhess-15-1711-2015

Lakshmi, D.D., Murty, P.L.N., Bhaskaran, P.K., Sahoo, B., Kumar, T.S., Shenoi, S.S.C., Srikanth, A.S., 2017. Performance of WRF-ARW winds on computed storm surge using hydodynamic model for Phailin and Hudhud cyclones. Ocean Eng. 131, 135–148. https://doi.org/10.1016/J.OCEANENG.2017.01.005

Lavery, S., Donovan, B., 2005. Flood risk management in the Thames Estuary looking ahead 100 years. Philos. Trans. A. Math. Phys. Eng. Sci. 363, 1455–74. https://doi.org/10.1098/rsta.2005.1579

Le Provost, C., Genco, M.L., Lyard, F., Vincent, P., Canceil, P., 1994. Spectroscopy of the world ocean tides from a finite element hydrodynamic model. J. Geophys. Res. 99, 24777. https://doi.org/10.1029/94JC01381

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., Stafford-Smith, M., 2014. A compound event framework for understanding extreme impacts. Wiley Interdiscip. Rev. Clim. Chang. 5, 113–128. https://doi.org/10.1002/wcc.252

Levy, S., Steinberg, D.M., 2010. Computer experiments: a review. Adv. Stat. Anal. 94, 311–324. https://doi.org/10.1007/s10182-010-0147-9

Lewis, M., Bates, P., Horsburgh, K., Neal, J., Schumann, G., 2013. A storm surge inundation model of the northern Bay of Bengal using publicly available data. Q. J. R. Meteorol. Soc. 139, 358–369. https://doi.org/10.1002/qj.2040

Lin, N., Emanuel, K.A., Smith, J.A., Vanmarcke, E., 2010. Risk assessment of hurricane storm surge for New York City. J. Geophys. Res. 115, D18121. https://doi.org/10.1029/2009JD013630

Lin, N., Emanuel, K.A., Oppenheimer, M., Vanmarcke, E., 2012. Physically-based Assessment of Hurricane Surge Threat under Climate Change. Nat. Clim. Chang. 2.6, 462–467.

Liu, F., West, M., 2009. A dynamic modelling strategy for Bayesian computer model emulation. Bayesian Anal. 4, 393–411. https://doi.org/10.1214/09-BA415

Liu, X., Guillas, S., 2017. Dimension Reduction for Gaussian Process Emulation: An Application to the Influence of Bathymetry on Tsunami Heights. SIAM/ASA J. Uncertain. Quantif. 5, 787–812. https://doi.org/10.1137/16M1090648

Loeppky, J.L., Sacks, J., Welch, W.J., 2009. Choosing the Sample Size of a Computer Experiment: A Practical Guide. Technometrics 51, 366–376. https://doi.org/10.1198/TECH.2009.08040

Love, M.R., Friday, D.Z., Grothe, P.R., Carignan, K.S., Eakins, B.W., Taylor, L.A., 2012. Digital Elevation Model of Astoria, Oregon: Procedures, Data Sources and Analysis.

Luettich, R A, J., Westerink, J.J., Scheffner, N.W., 1992. ADCIRC: An Advanced Three-Dimensional Circulation Model for Shelves, Coasts, and Estuaries. Report 1. Theory and Methodology of ADCIRC-2DDI and ADCIRC-3DL.

Madsen, H., Jakobsen, F., 2004. Cyclone induced storm surge and flood forecasting in the northern Bay of Bengal. Coast. Eng. 51, 277–296. https://doi.org/10.1016/J.COASTALENG.2004.03.001

Malde, S., Oakley, J., Wyncott, D., 2016. MUCM: Gaussian Process Emulator.

Malde, S., Wyncoll, D., Oakley, J., Tozer, N., Gouldby, B., 2016. Applying emulators for improved flood risk analysis. E3S Web Conf. 7, 04002. https://doi.org/10.1051/e3sconf/20160704002

Mass, C., Dotson, B., 2010. Major Extratropical Cyclones of the Northwest United States: Historical Review, Climatology, and Synoptic Environment. Mon. Weather Rev. 138, 2499–2527. https://doi.org/10.1175/2010MWR3213.1

Mastrandrea, M.D., Field, C.B., Stocker, T.F., Edenhofer, O., Ebi, K.L., Frame, D.J., Held, H., Kriegler, E., Mach, K.J., Matschoss, P.R., Plattner, G.-K., Yohe, G.W., Zwiers, F.W., 2010. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties IPCC Cross-Working Group Meeting on Consistent Treatment of Uncertainties.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics 21, 239–245. https://doi.org/10.1080/00401706.1979.10489755

Mckay, P., Blain, C.A., 2010. Toward Developing a Hydrodynamic Flow and Inundation Model of the Lower Pearl River.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E.H., Ek, M.B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., Shi, W., Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E.H., Ek, M.B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., Shi, W., 2006. North American Regional Reanalysis. Bull. Am. Meteorol. Soc. 87, 343–360. https://doi.org/10.1175/BAMS-87-3-343

Moel, H., Asselman, N., Aerts, J., 2012. Uncertainty and sensitivity analysis of coastal flood damage estimates in the west of the Netherlands. Nat. Hazards Earth Syst. Sci 12, 1045–1058. https://doi.org/10.5194/nhess-12-1045-2012

Moftakhari, H.R., AghaKouchak, A., Sanders, B.F., Matthew, R.A., 2017. Cumulative hazard: The case of nuisance flooding. Earth's Futur. 5, 214–223. https://doi.org/10.1002/2016EF000494

Montoya, R.D., Osorio Arias, A., Ortiz Royero, J.C., Ocampo-Torres, F.J., 2013. A wave parameters and directional spectrum analysis for extreme winds. Ocean Eng. 67, 100–118. https://doi.org/10.1016/J.OCEANENG.2013.04.016

Morris, M.D., Mitchel, T.J., 1995. Exploratory Designs for Computational Experiments. J. Stat. Plan. Inference 43, 381–402. https://doi.org/10.1016/0378-3758(94)00035-T

NOAA National Centers for Environmental Information, 2003. Coastal Relief Model.

O'Hagan, A., 2006. Bayesian analysis of computer code outputs: A tutorial. Reliab. Eng. Syst. Saf. 91, 1290–1300. https://doi.org/10.1016/j.ress.2005.11.025

Oakley, J., 1999. Bayesian uncertainty analysis for complex computer codes. Thesis.

Orton, P.M., Hall, T.M., Talke, S.A., Blumberg, A.F., Georgas, N., Vinogradov, S., 2016. A validated tropical-extratropical flood hazard assessment for New York Harbor. J. Geophys. Res. Ocean. 121, 8904–8929. https://doi.org/10.1002/2016JC011679

Pendleton, L.H., 2010. The economic and market value of coasts and estuaries: what's at stake? Restore America's Estuaries, Arlington.

Perez, J., Menendez, M., Losada, I.J., 2017. GOW2: A global wave hindcast for coastal applications. Coast. Eng. 124, 1–11. https://doi.org/10.1016/J.COASTALENG.2017.03.005

Pugh, D.T., 1996. Tides, Surges and Mean Sea-Level. John Wiley & Sons.

Purvis, M.J., Bates, P.D., Hayes, C.M., 2008. A probabilistic methodology to estimate future coastal flood risk due to sea level rise. Coast. Eng. 55, 1062–1073. https://doi.org/10.1016/j.coastaleng.2008.04.008

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning. Cambridge: MIT press.

Razavi, S., Tolson, B. a., Burn, D.H., 2012. Review of surrogate modeling in water resources. Water Resour. Res. 48. https://doi.org/10.1029/2011WR011527

Reichert, P., White, G., Bayarri, M.J., Pitman, E.B., 2011. Mechanism-based emulation of dynamic simulation models: Concept and application in hydrology. Comput. Stat. Data Anal. 55, 1638–1655. https://doi.org/10.1016/j.csda.2010.10.011

Resio, D.T., Irish, J., Cialone, M., 2009. A surge response function approach to coastal hazard assessment – part 1: basic concepts. Nat. Hazards 51, 163–182. https://doi.org/10.1007/s11069-009-9379-y

Resio, D.T., Westerink, J.J., 2008. Modeling the Physics of Storm Surges. Phys. Today 61. https://doi.org/10.1063/1.2982120

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., Aigrain, S., 2013. Gaussian processes for time-series modelling. Philos. Trans. A. Math. Phys. Eng. Sci. 371, 20110550. https://doi.org/10.1098/rsta.2011.0550

Rogers, W.E., Kaihatu, J.M., Hsu, L., Jensen, R.E., Dykes, J.D., Holland, K.T., 2007. Forecasting and hindcasting waves with the SWAN model in the Southern California Bight. Coast. Eng. 54, 1–15. https://doi.org/10.1016/J.COASTALENG.2006.06.011

Rohmer, J., Idier, D., 2012. A meta-modelling strategy to identify the critical offshore conditions for coastal flooding. Nat. Hazards Earth Syst. Sci. 12, 2943–2955. https://doi.org/10.5194/nhess-12-2943-2012

Rueda, A., Gouldby, B., Mendez, F.J., Tomas, A., Losada, I.J., Lara, J.L., Diaz-Simal, P., 2016. The use of wave propagation and reduced complexity inundation models and metamodels for coastal flood risk assessment. J. Flood Risk Manag. 9, 390–401. https://doi.org/10.1111/jfr3.12204

Rusu, L., Pilar, P., 2008. Hindcast of the wave conditions along the west Iberian coast. Coast. Eng. 55, 906–919. https://doi.org/10.1016/J.COASTALENG.2008.02.029

Sacks, J., Schiller, S.B., Welch, W.J., 1989. Designs for Computer Experiments. Technometrics 31, 41–47. https://doi.org/10.1080/00401706.1989.10488474

Schulz, E., Speekenbrink, M., Krause, A., 2018. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. J. Math. Psychol. 85, 1–16. https://doi.org/10.1016/J.JMP.2018.03.001

Serafin, K.A., Ruggiero, P., 2014. Simulating extreme total water levels using a time-dependent, extreme value approach. J. Geophys. Res. Ocean. 119, 6305–6329. https://doi.org/10.1002/2014JC010093

Serafin, K.A., Ruggiero, P., Stockdon, H.F., 2017. The relative contribution of waves, tides, and non-tidal residuals to extreme total water levels on US West Coast sandy beaches. Geophys. Res. Lett. 44, 1839–1847. https://doi.org/10.1002/2016GL071020

Song, Y.K., Irish, J.L., Udoh, I.E., 2012. Regional attributes of hurricane surge response functions for hazard assessment. Nat. Hazards 64, 1475–1490. https://doi.org/10.1007/s11069-012-0309-z

Svensson, C., Jones, D.A., 2004. Dependence between sea surge, river flow and precipitation in south and west Britain. Hydrol. Earth Syst. Sci. 8, 973–992. https://doi.org/10.5194/hess-8-973-2004

Taylor, N.R., Irish, J.L., Udoh, I.E., Bilskie, M. V., Hagen, S.C., 2015. Development and uncertainty quantification of hurricane surge response functions for hazard assessment in coastal bays. Nat. Hazards 77, 1103–1123. https://doi.org/10.1007/s11069-015-1646-5

Timmermans, B., 2015. Uncertainty in Numerical Wind-wave Models. University of Southhampton.

Tolman, H.L., 2009. User manual and system documentation of WAVEWATCH III TM version 3.14 †.

Wahl, T., Jain, S., Bender, J., Meyers, S.D., Luther, M.E., 2015. Increasing risk of compound flooding from storm surge and rainfall for major US cities. Nat. Clim. Chang. 5, 1093–1097. https://doi.org/10.1038/nclimate2736

Weaver, R.J., Luettich, Jr., R.A., 2010. 2D vs. 3D Storm Surge Sensitivity in ADCIRC: Case Study of Hurricane Isabel, in: Estuarine and Coastal Modeling (2009). American Society of Civil Engineers, Reston, VA, pp. 762–779. https://doi.org/10.1061/41121(388)44

Weaver, R.J., Slinn, D.N., 2010. Influence of bathymetric fluctuations on coastal storm surge. Coast. Eng. 57, 62–70. https://doi.org/10.1016/J.COASTALENG.2009.09.012

Westerink, J.J., Luettich, R.A., Baptists, A.M., Scheffner, N.W., Farrar, P., 1992. Tide and Storm Surge Predictions Using Finite Element Model. J. Hydraul. Eng. 118, 1373–1390. https://doi.org/10.1061/(ASCE)0733-9429(1992)118:10(1373)

Zhang, K., Douglas, B.C., Leatherman, S.P., 1999. Twentieth-Century Storm Activity along the U.S. East Coast. J. Clim. 13.

Zijlema, M., 2010. Computation of wind-wave spectra in coastal waters with SWAN on unstructured grids. Coast. Eng. 57, 267–277. https://doi.org/10.1016/J.COASTALENG.2009.10.011

Zscheischler, J., Westra, S., van den Hurk, B.J.J.M., Seneviratne, S.I., Ward, P.J., Pitman, A., AghaKouchak, A., Bresch, D.N., Leonard, M., Wahl, T., Zhang, X., 2018. Future climate risk from compound events. Nat. Clim. Chang. 8, 469–477. https://doi.org/10.1038/s41558-018-0156-3