Neglecting model structural uncertainty underestimates upper tails of flood hazard

Environmental Research Letters

OPEN ACCESS

RECEIVED
10 September 2017

REVISED
3 June 2018

ACCEPTED FOR PUBLICATION
8 June 2018

PUBLISHED
5 July 2018

LETTER

Neglecting model structural uncertainty underestimates upper tails of flood hazard

Tony E Wong[1,6,7], Alexandra Klufas[2], Vivek Srikrishnan[3] and Klaus Keller[1,4,5]

[1] Earth and Environmental Systems Institute, Pennsylvania State University, University Park, PA 16802, United States of America
[2] Department of Mathematics, Wellesley College, Wellesley, MA 02481, United States of America
[3] Department of Energy and Mineral Engineering, Pennsylvania State University, University Park, PA 16802, United States of America
[4] Department of Geosciences, Pennsylvania State University, University Park, PA 16802, United States of America
[5] Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15289, United States of America
[6] Current address: Department of Computer Science, University of Colorado, Boulder, CO 80309, United States of America
[7] Author to whom any correspondence should be addressed.

E-mail: anthony.e.wong@colorado.edu

Keywords: coastal flooding, natural hazards, Bayesian statistics, uncertainty, extremes

Supplementary material for this article is available online

Abstract

Coastal flooding drives considerable risks to many communities, but projections of future flood risks are deeply uncertain. The paucity of observations of extreme events often motivates the use of statistical approaches to model the distribution of extreme storm surge events. One key deep uncertainty that is often overlooked is model structural uncertainty. There is currently no strong consensus among experts regarding which class of statistical model to use as a 'best practice'. Robust management of coastal flooding risks requires coastal managers to consider the distinct possibility of non-stationarity in storm surges. This increases the complexity of the potential models to use, which tends to increase the data required to constrain the model. Here, we use a Bayesian model averaging approach to analyze the balance between (i) model complexity sufficient to capture decision-relevant risks and (ii) data availability to constrain complex model structures. We characterize deep model structural uncertainty through a set of calibration experiments. Specifically, we calibrate a set of models ranging in complexity using long-term tide gauge observations from the Netherlands and the United States. We find that in both considered cases, roughly half of the model weight is associated with the non-stationary models. Our approach provides a formal framework to integrate information across model structures, in light of the potentially sizable modeling uncertainties. By combining information from multiple models, our inference sharpens for the projected storm surge 100 year return levels, and estimated return levels increase by several centimeters. We assess the impacts of data availability through a set of experiments with temporal subsets and model comparison metrics. Our analysis suggests that about 70 years of data are required to stabilize estimates of the 100 year return level, for the locations and methods considered here.

1. Introduction

Storm surges drive substantial risks to coastal communities (Nicholls and Cazenave 2010), but there remains deep structural uncertainty regarding how best to model this threat. Previous work has broken important new ground by considering process-based modeling (Fischbach *et al* 2017, Orton *et al* 2016, Johnson *et al* 2013) as well as statistical modeling approaches (Buchanan *et al* 2015, Grinsted *et al* 2013, Tebaldi *et al* 2012, Menéndez and Woodworth 2010). Recently, we have seen the advent of semi-empirical models for sea-level rise and their application to coastal risk management (Kopp *et al* 2017, Nauels *et al* 2017, Wong *et al* 2017a, 2017b, Mengel *et al* 2016). The total flood hazard depends on predictions of both sea-level rise and storm surge properties. In this case, it can be attractive to have flexible and efficient models to estimate storm surge hazards, with a formal statistical accounting of uncertainty and linked to accessible climate variables. This motivates our study's focus on the statistical modeling of storm surges.

Previous studies have provided important new insights by examining the potentially sizable impacts of non-stationarity in the treatment of storm frequency, distribution and intensity (e.g. Ceres *et al* 2017, Lee *et al* 2017, Cid *et al* 2016, Grinsted *et al* 2013, Haigh *et al* 2010b, Menéndez and Woodworth 2010). For example, Grinsted *et al* (2013) use a generalized extreme value (GEV) distribution to model extreme sea levels, and incorporate non-stationarity in the model parameters by allowing them to covary with global mean surface temperature. Other studies consider a hybrid statistical model wherein the frequency of extreme sea level events is governed by a Poisson process (PP) and the magnitude of these events follows a Generalized Pareto distribution (GPD) (Wahl *et al* 2017, Hunter *et al* 2017, Buchanan *et al* 2017, Cid *et al* 2016, Bulteau *et al* 2015, Marcos *et al* 2015, Arns *et al* 2013, Tebaldi *et al* 2012). Non-stationarity may be incorporated into the PP/GPD statistical model by covarying the PP/GPD parameters with climatic conditions (Marcos *et al* 2015, Haigh *et al* 2010b). Here, we follow and expand on the work of Haigh *et al* (2010b) and examine how non-stationarity—covarying with changing North Atlantic oscillation (NAO) index—affects projections of future storm surge return levels using a PP/GPD model.

Extreme events are, by definition, rare. It is hence important to use the relatively sparse data well. The GEV approach requires to bin observations into time blocks, processed in a manner so as to remove the interdependence of the observations, and take block maxima. Often, this is done using annual blocks (e.g. Wong and Keller 2017, Karamouz *et al* 2017), yielding a potentially limited amount of data with which to fit an extreme value statistical model (Coles 2001). Another option is to process data to achieve independence, then use shorter time lengths of blocks (Grinsted *et al* 2013), but the choice of processing procedure is nontrivial and the fidelity with which non-stationary behavior may be detected is uncertain (e.g. Ceres *et al* 2017, Lee *et al* 2017). The PP/GPD modeling approach is an attractive option because all events above a specified threshold are considered in fitting the model, leading to a richer set of data (e.g. Knighton *et al* 2017, Arns *et al* 2013). While we do not employ these methods, it is important to note that other approaches exist to analyze extreme sea levels; for example, those based on the joint probability method (McMillan *et al* 2011, Haigh *et al* 2010a, Tawn and Vassie 1989, Pugh and Vassie 1979). Previous work has examined how data availability affects model prediction (Dangendorf *et al* 2016), but this question remains largely open for longer tide gauge records (>90 years).

A related open question is how to select a statistical model of extreme storm surges. Relative to stationary models, the increased complexity of non-stationary models can lead to wider predictive intervals, and perhaps the dismissal of the more complex model—along with arguably decision-relevant tail behavior. Traditional approaches often favor parsimonious use of the limited data (e.g. Karamouz *et al* 2017, Lee *et al* 2017, Buchanan *et al* 2015, Tebaldi *et al* 2012). Bayesian model averaging (BMA), however, offers an avenue to combine a range of candidate model structures by allowing the data to inform the degree to which each model is to be trusted (Hoeting *et al* 1999). Models are a proxy for data not yet observed, and our BMA approach presents an opportunity to formally integrate multiple information streams (Moftakhari *et al* 2017).

Here, we combine the non-stationarity covarying with NAO index with a PP/GPD modeling approach to address the interrelated questions of how data length affects model choice, and how model choice impacts estimates of storm surge hazards. We employ the PP/GPD model because we are motivated by the need to examine how best to utilize the inherently limited data regarding extreme sea levels. We use two relatively long and complete tide gauge records to demonstrate that for both sites and all data lengths, non-stationary models receive considerable weight in a Bayesian model averaging sense. The major contributions of this study are: (i) to present a formal statistical framework to combine information across models and account for structural uncertainties through use of Bayesian model averaging, and (ii) to assess how the length of data record affects our model choices, and thus impacts estimates of future flood hazard.

2. Methods

2.1. Storm surge statistical modeling

We employ a peaks-over-thresholds (POT) approach, with a PP/GPD statistical model, to estimate the distribution of extreme storm surge events. We find similar conclusions in an experiment assessing the implications of our results using a block maxima approach in the region considered by Grinsted *et al* (2013) (see supplementary material available at stacks.iop.org/ERL/13/074019/mmedia). The POT approach makes use of only observational data that exceed a specified threshold to fit the PP/GPD model parameters. We follow previous work (e.g. Wahl *et al* 2017, Arns *et al* 2013) and process the data by: (i) using a constant threshold $\mu(t)$ equal to the 99th percentile of the daily maximum water levels, (ii) detrending by subtracting a moving window one-year average from the raw hourly data (or three-hourly for Delfzijl) to account for sea-level rise but retain sub-decadal variability, the effects of astronomical tides, and interannual variability, as well as the effects of storm surges, and (iii) using a declustering routine to isolate extreme events at least 72 hours apart. In coastal risk management applications, these methods would be used together with a set of local mean sea-level rise projections that would likely have an annual time step. Thus, it is important to retain these non-mean sea level signals. In a set of supplemental experiments, we also examine a declustering time-scale of 24 hours and POT thresholds of the 95th and 99.7th percentiles. The interested reader is referred to Arns *et al* (2013) for a careful review of key structural uncertainties.

The probability density function (pdf, f) and cumulative distribution function (cdf, F) for the potentially non-stationary form of the GPD used here are given by:

$$f\left(x\left(t\right);\mu\left(t\right),\sigma\left(t\right),\xi\left(t\right)\right) \\ =\frac{1}{\sigma(t)}\left(1+\xi\left(t\right)\frac{x(t)-\mu(t)}{\sigma(t)}\right)^{-\left({}^{1}\!/\xi\left(t\right)+1\right)} \tag{1}$$

$$F\left(x\left(t\right);\mu\left(t\right),\sigma\left(t\right),\xi\left(t\right)\right) \\ =1-\left(1+\xi\left(t\right)\frac{x(t)-\mu(t)}{\sigma(t)}\right)^{-{}^{1}\!/\xi\left(t\right)}, \tag{2}$$

where $x(t)$ is the processed daily maximum tide gauge level (meters), $\sigma(t)$ is the scale parameter (meters) and $\xi(t)$ is the shape parameter (unitless), all as functions of time t (days). σ governs the width of the distribution and ξ governs the heaviness of the distribution's tail. A Poisson process governs the probability g of observing $n(t)$ exceedances of threshold $\mu(t)$ during time interval Δt (days):

$$g(n(t);\lambda(t))=\frac{(\lambda(t)\Delta t)^{n(t)}}{n(t)!}\exp(-\lambda(t)\Delta t), \tag{3}$$

where $\lambda(t)$ is the Poisson rate parameter (exceedances day^{-1}).

We incorporate potential non-stationarity into the PP/GPD model following the approach of Grinsted *et al* (2013), by allowing the model parameters to covary with winter (DJF) average NAO index:

$$\begin{cases}\lambda\left(t\right)=\lambda_0+\lambda_1 NAO\left(t\right)\\ \sigma\left(t\right)=exp\left[\sigma_0+\sigma_1 NAO\left(t\right)\right]\\ \xi\left(t\right)=\xi_0+\xi_1 NAO\left(t\right).\end{cases} \tag{4}$$

λ_0, λ_1, σ_0, σ_1, ξ_0, and ξ_1 are uncertain model parameters, determined by fitting to the processed tide gauge record (detailed below). We assume the parameters are stationary within each year. The processing of tide gauge data into a surge index in Grinsted *et al* (2013) serves to (1) achieve independence among observations, and (2) increase the effective amount of data by pooling across sites. Regarding (1), we process our tide gauge data to achieve independence (see above). Regarding (2), we are investigating how data availability affects our ability to constrain storm surge statistical models, and what the impacts are on model projections relevant to managing local coastal risks. We use direct tide gauge data instead of a surge index because we are currently unaware of any method to map surge index back to a localized projection.

Finally, the joint likelihood function for the model parameters $\theta=(\lambda_0,\lambda_1,\sigma_0,\sigma_1,\xi_0,\xi_1)^{\mathrm{T}}$, given the time series of daily maxima threshold exceedances, x, is:

Table 1. Candidate model structures and their parameters.

Model structure	Non-stationary parameters	Model parameters to calibrate
ST	None	λ_0, σ_0, ξ_0
NS1	λ	λ_0, λ_1, σ_0, ξ_0
NS2	λ, σ	λ_0, λ_1, σ_0, σ_1, ξ_0
NS3	λ, σ, ξ	λ_0, λ_1, σ_0, σ_1, ξ_0, ξ_1

$$L(x|\theta)=\prod_{i=1}^{N}\left[g(n(y_i);\ \lambda(y_i))\right. \\ \left.\prod_{j=1}^{n(y_i)} f(x_j(y_i);\ \mu(y_i),\sigma(y_i),\xi(y_i))\right], \tag{5}$$

where $i=1, 2, \ldots, N$ indexes the years of tide gauge data and $j=1, 2, \ldots, n(y_i)$ indexes the exceedances $x_j(y_i)$ in year y_i. The product indexed by j in equation (5) is replaced by 1 for all i such that $n(y_i)=0$.

We consider four candidate models within the class of PP/GPD models, ranging from a stationary model (denoted by 'ST', in which $\lambda_1=\sigma_1=\xi_1=0$) to fully non-stationary ('NS3', in which all six parameters are considered). These models are summarized in table 1. We project future storm surge return levels to 2065. We focus on the 100 year return level, which is motivated by its common use in coastal risk management (e.g. Coastal Protection and Restoration Authority of Louisiana 2017), but results for other return periods are presented in the supplementary material.

2.2. Model calibration

2.2.1. Data

We fit the candidate models' parameters (table 1) using the tide gauge data record from two sites: Delfzijl, the Netherlands (Rijkswaterstaat 2017), and Sewells Point (Norfolk), Virginia, United States (NOAA 2017). We selected these sites because the lengths of the records (137 and 89 years, respectively) enable our set of experiments regarding the impacts of data length on surge level estimation, they are geographically well-separated and these tide gauge records are relatively complete (each site has three or fewer gaps longer than one month). We use time series of detrended daily block maxima for the POT approach (e.g. Arns *et al* 2013).

We use historical monthly NAO index data from Jones *et al* (1997). We use the sea level pressure projection of the MPI-ECHAM5 simulation under SRES scenario A1B as part of the ENSEMBLES project (www.ensembles-eu.org, Roeckner *et al* 2003). We calculate the winter mean (DJF) NAO index following Stephenson *et al* (2006) to use as input to the nonstationary models. We caution that these results do not account for model structural nor parametric uncertainty regarding future NAO index. An assessment of the impacts of these uncertainties on projected surge levels is another important avenue for future study.

We evaluate the impacts of data length on PP/GPD parameter estimates through a set of experiments. In these experiments, we employ only the 30, 50, 70, 90, 110 and 137 most recent years of data from the Delfzijl tide gauge site, and the 30, 50, 70 and 89 most recent years from Norfolk.

2.2.2. Bayesian calibration framework

We calibrate each of the four candidate models (table 1) using each of the two processed tide gauge records ($x(t)$) and winter NAO index series (NAO(t)). We employ a robust adaptive Markov chain Monte Carlo approach (Vihola 2012). The essence of this calibration approach is to update the prior probability distribution of the model parameters ($p(\theta)$) by quantifying the goodness-of-fit between the observational data and the Poisson process/generalized Pareto models given by candidate sets of model parameters. This goodness-of-fit is quantified by the likelihood function (equation 5). Bayes' theorem combines the prior knowledge regarding the model parameters with the information gained from the observational data (i.e. the likelihood function) into the posterior distribution of the model parameters, given the data ($p(\theta|x)$):

$$p\left(\theta|x\right) \propto L\left(x|\theta\right) p\left(\theta\right). \tag{6}$$

We represent prior knowledge regarding the parameters ($p(\theta)$) as follows. First, we obtain maximum likelihood parameter estimates (MLEs) for 28 tide gauge sites with at least 90 years of data available, as well as the two records on which this study focuses. These sites were selected using the University of Hawaii Sea Level Center's online database, and a spreadsheet utility we developed (and provide with the model codes in the repository accompanying this study) (Caldwell *et al* 2011). Details regarding these sites are provided in the supplementary material accompanying this article. Second, we fit either a normal or gamma distribution to the set of 30 MLEs for each parameter, depending on whether the parameter has infinite (normal: λ_1, σ_1, ξ_0, ξ_1) or half-infinite (gamma: λ_0, σ_0) support. The resulting prior distributions, MLEs, and an experiment using uniform prior distributions are shown in the supplementary material.

We initialize the Markov chains at the MLE parameters for each site and for each candidate model. We produce 500 000 iterations for 10 parallel Markov chains and remove the first 50 000 iterations for burn-in. Gelman and Rubin diagnostics are used to assess convergence and burn-in length (Gelman and Rubin 1992). For each site, for each of the four candidate models, and for each of the data length experiments, we draw an ensemble of 10 000 parameter sets for analysis from the remaining 4 500 000 Markov chain samples. We calibrate in this manner for each of the length of data experiments (see section 2.2.1).

We also conduct a preliminary experiment by binning the Delfzijl data into 11 overlapping 30 year blocks, spanning the 137 year range. We calibrate the stationary model (ST) to the data in each of the 11 blocks, and calculate the estimated 100 year return level for each block's ensemble. We examine changes in the quantiles of these 11 distributions to assess the potential need for a non-stationary approach.

2.2.3. Bayesian model averaging

Bayesian model averaging (BMA) (Hoeting *et al* 1999) is a method by which the storm surge return level estimates implied by the posterior parameters (obtained as in section 2.2.2) for each candidate model (table 1) may be combined and weighted by the model marginal likelihood, given the data, $p(M_k|x)$. Let RL(y_i $|x$, M_k) denote the return level in year y_i assuming model structure M_k∈{*ST, NS1, NS2, NS3*} and given the observational data x. Then the BMA-weighted return level in year y_i, integrating the estimates from all four candidate models, is

$$\mathrm{RL}\left(y_i|x\right) = \sum_{k=1}^{4} \mathrm{RL}\left(y_i|M_k\right) p(M_k|x). \tag{7}$$

The BMA weights, $p(M_k|x)$, are given by

$$p\left(M_k|x\right) = \frac{p\left(x|M_k\right) p\left(M_k\right)}{\sum_{l=1}^{4} p\left(x|M_l\right) p\left(M_l\right)}, \tag{8}$$

where the denominator marginalizes the probability of the data, $p(x)$, over the four model structures considered. We make the assumption that all model structures are equally likely *a priori* (i.e. $p(M_k) = p(M_l)$, $\forall M_k$, M_l ∈ {*ST, NS1, NS2, NS3*}). The probabilities $p(x|M_k)$ are determined by integration over the posterior distributions of the model parameters:

$$p\left(x|M_k\right) = \int_\theta p\left(x|\theta, M_k\right) p\left(\theta\right) d\theta, \tag{9}$$

where the integral is over the relevant parameters for model M_k. The probabilities $p(x|\theta,M_k)$ are the likelihood function (equation 5) with conditional dependence on the model structure made explicit. These and the prior probabilities ($p(\theta)$) are sampled as described in section 2.2.2.

From equation (9), $p(x|M_k)$ is the normalizing constant (or *marginal likelihood*) for the probability density function associated with model M_k. We use bridge sampling (Meng and Wing 1996) to estimate the marginal likelihoods of the models under consideration, using a normal approximation to the joint posterior as the importance density.

2.2.4. Model comparison metrics

We employ several metrics for model comparison. They are motivated by the balance between model goodness-of-fit, model complexity, and the availability of data. The first metric is the Akaike information criterion (AIC) (Akaike 1974):

$$\mathrm{AIC} = -2\log\left(L_{\max}\right) + 2N_p, \tag{10}$$

where $L_{\max}$ is the maximum value of the likelihood function (equation 5) within the posterior model ensemble and N_p is the number of model parameters.

Table 2. Model selection criteria for the four candidate models. Lower is better for AIC, BIC and DIC; higher is better for BMA weight. Shaded cells denote the model choice indicated by each metric.

Tide gauge	Model structure	AIC	BIC	DIC	BMA weight
Delfzijl, the Netherlands	ST	5545.62	5557.26	13855.07	0.42
	NS1	5546.07	5561.59	13852.99	0.33
	NS2	5547.70	5567.10	13853.27	0.22
	NS3	5548.21	5571.50	13851.72	0.04
Norfolk, Virginia, USA	ST	2883.20	2893.21	7198.87	0.51
	NS1	2884.39	2897.74	7198.76	0.24
	NS2	2886.27	2902.96	7199.84	0.20
	NS3	2886.73	2906.75	7198.11	0.05

The second metric is the Bayesian information criterion (BIC) (Schwarz 1978):

$$\mathrm{BIC} = -2\log\left(L_{\max}\right) + N_p \log\left(N_{\mathrm{obs}}\right), \tag{11}$$

where N_{obs} is the number of observational data used to fit the model. Thus, for $N_{\mathrm{obs}} > e^2$, BIC penalizes over parameterization more harshly than AIC.

The third metric is the deviance information criterion (DIC) (Spiegelhalter *et al* 2002). For a given model structure, define the *deviance* for a given set of model parameters as $D(\theta) = -2\log(L(x|\theta))$. Denote by $\bar{D}$ the expected value of $D(\theta)$ over θ, and let $\bar{\theta}$ refer to the expected value of θ. The *effective* number of parameters is calculated as $p_D = \bar{D} - D\left(\bar{\theta}\right)$. DIC is then:

$$\mathrm{DIC} = p_D + \bar{D}. \tag{12}$$

The final metric we employ for model comparison is the BMA weights themselves (equation 8). Note that AIC and BIC are calculated based on the performance of the maximum likelihood ensemble member, whereas DIC and BMA weight are based on the entire ensemble.

In addition to the four ensembles corresponding to each of the candidate models (table 1), we construct a BMA-weighted ensemble of estimated return levels as follows. We draw 10 000 sets of parameters from each of the four candidate models. The number of samples was selected to match the number of samples used for each individual model. For each of these 10 000 concomitant sets of BMA parameters, we calculate the return period according to equation (7).

3. Results

3.1. Hindcast test

The Delfzijl site displays evidence for non-stationary behavior in the 100 year return level (figure 1). We determine the distributions shown in figure (1) by binning the data into overlapping 30 year blocks and fitting the stationary (ST) model using the Bayesian approach outlined in section 2.2.2. The estimated median 100 year return level ranges from 412–490 cm across the 11 blocks, and widths of the 5%–95% credible interval range from 146–285 cm. This motivates the need for a non-stationary approach.

We find that the more complex models (NS1, NS2 and NS3) generally result in somewhat lower traditional model performance metrics (table 2). However, we note that differences of $O(1)$ in AIC or BIC may not be sufficient evidence to dismiss the more complex models (Kass and Raftery 1995). For both sites, the BMA weights associated with the non-stationary models NS1 and NS2 are roughly between 20% and 30%, indicating the value of multi-model approaches over single-model or stationary modeling approaches.

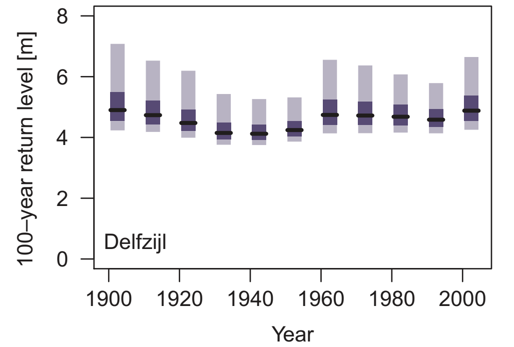

Figure 1. Distributions of the 100 year return levels at Delfzijl, the Netherlands, using the stationary (ST) model and 30 year blocks of data, centered at the locations of the vertical bars. The black horizontal lines denote the ensemble median; the dark and light bars denote the 25%–75% and 5%–95% quantile ranges, respectively.

3.2. Estimates of current and future surge levels

The resulting predictive distributions for 2016 and projected 2065 surge levels demonstrate the impacts of integrating across model structures (figure 2; see supplementary material for these results in tabular form). Interestingly, the NS3 model displays a *reduction* in 100 year return level for both sites by 2065, but also receives the lowest BMA weight (about 5%). The fact that the ST, NS1 and NS2 models' projections are in relative agreement and match the data well (see table 2) lends confidence to their results. This agreement, characterized by quite similar posterior pdfs, leads to a tighter credible range in the BMA projection (figure 2). While the sharpened inference in the BMA pdf in this case may seem counterintuitive, this follows from the fact that the BMA return levels are *averages* of the return levels from the four candidate models. Averaging is a smoothing operation, so extreme behavior is dampened (see also supplementary material for a note describing this phenomenon). Indeed, a key

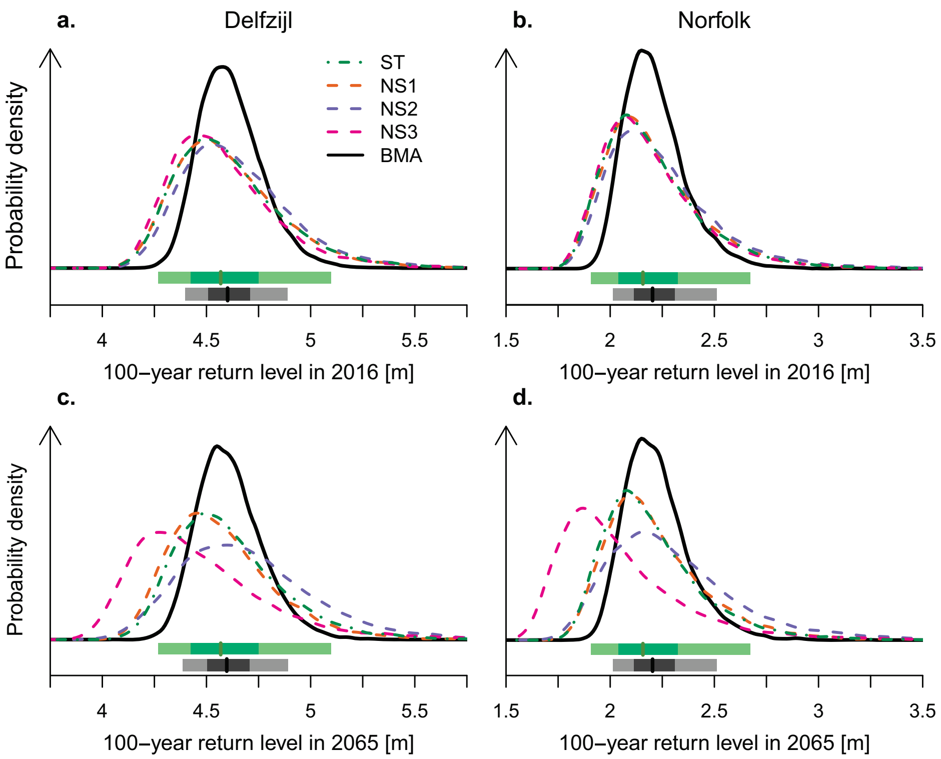

Figure 2. Posterior probability density function (pdfs) of the 100 year return levels using the BMA-weighted ensemble and four individual model structures for 2016 conditions (top row) and projected 2065 surge levels (bottom row), at Delfzijl (left column) and Norfolk (right column). Below the pdfs are boxplots for the stationary model (ST, green horizontal bars) and the BMA-weighted ensemble (gray/black horizontal bars). The bold vertical lines denote the ensemble medians; the dark and light bars denote the 25%–75% and 5%–95% quantile ranges, respectively.

strength of our BMA approach is to formally quantify the degree of belief in each model structure, informed by the quality of model match to data.

We find that a stationary PP/GPD approach underestimates projected 100 year surge levels in 2065 by 3 and 4 cm for Delfzijl and Norfolk, respectively, relative to the BMA approach (ensemble medians, figures 2(*c*) and (*d*); see also tables S2 and S3). While 3 cm may not seem like a substantial increase in hazard, it is ultimately up to the decision-maker to assess the relevant hazards for themselves, and our BMA approach incorporates model specification uncertainty into the projections presented. In any case, these results serve as a proof of concept of the use of Bayesian model averaging in a statistical treatment of extreme sea levels, and characterize the model structural uncertainty.

3.3. Reliability of estimated surge levels

We assess the impacts of data length on the distributions of PP/GPD parameters for the four candidate models (figure 3). With a relatively short record (30–50 years of data), 5%–95% credible intervals for the 100 year surge level are much wider than when 70 or more years of data are available. While it is beyond the scope of this study, future work might consider developing a formal convergence metric using (for example) Kolmogorov and Smirnov statistics (Smirnov 1948, Kolmogorov 1933).

The BMA weights change for each site as more data become available, but once 70 years of data are available, the ordering of the models' BMA weights remains stable (figure 4). Across all of the data length experiments, the stationary model has the largest BMA weight for both sites, at about 40%–50%. As more data become available at Delfzijl, the stationary model receives less than 50% weight and models NS1 and NS2 receives roughly 30 and 20% weight, respectively. We find similar results for Norfolk. It is consistent and clear across sites and data lengths, however, that the non-stationary models receive about half of the model weight. This result is also robust to changes in the selected POT threshold (see supplementary material). This illustrates the potential limitations of single-model approaches.

4. Discussion

Our analysis (i) showcases a new framework to integrate decision-relevant information (i.e. non-stationarity) into storm surge projections and (ii) uses this framework to demonstrate practical implications of neglecting key modeling uncertainties. Our analysis, of course, is subject to several caveats. For example, our BMA approach weights each model according to its posterior probability (under a uniform prior over the model space), thereby implicitly using

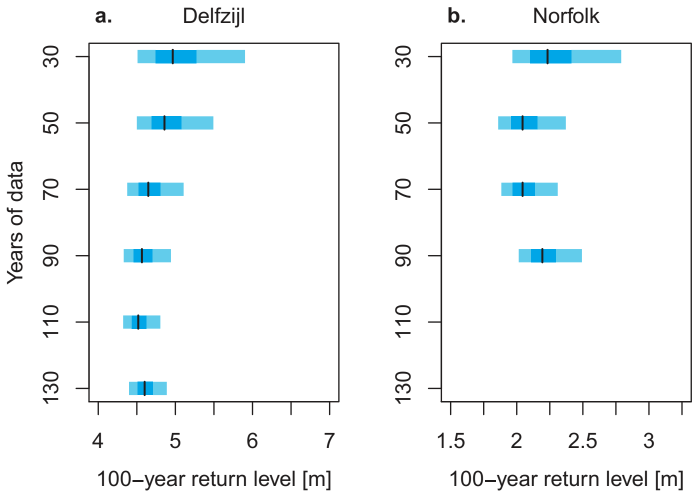

Figure 3. Distributions of return levels from the BMA-weighted ensemble in year 2016 for varying time lengths of data employed for (*a*) Delfzijl and (*b*) Norfolk. The black vertical lines denote the ensemble medians; the extent of the darker boxes gives the 25%−75% credible range; the extent of the lighter boxes gives the 5%−95% credible range.

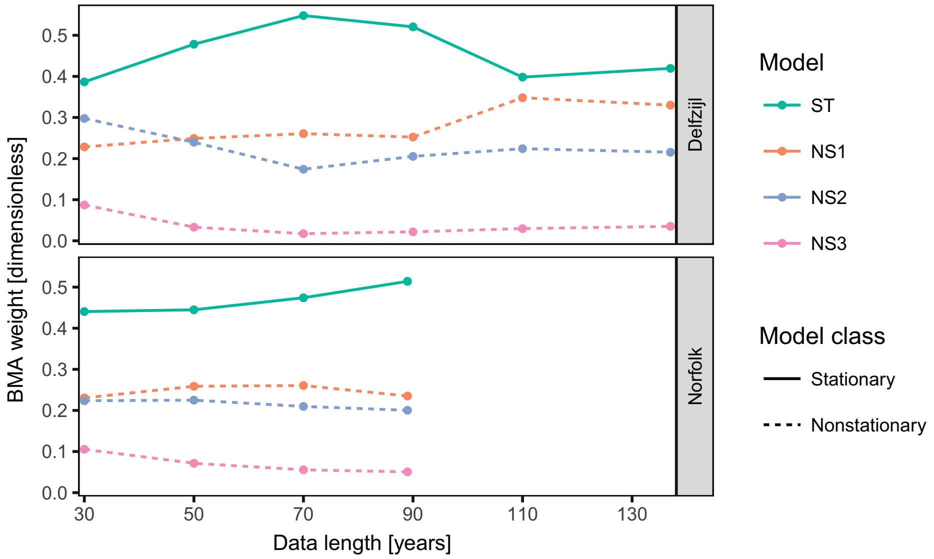

Figure 4. Bayesian model averaging weights (equation 8) for the four candidate models, using varying lengths of data from the tide gauge stations at Delfzijl (137 years of data total) and Norfolk (89 years). Higher values imply better model-data match.

a quadratic loss function with respect to the choice of model (Robert 2007). The quadratic loss function may not be the most appropriate loss for all applications using storm surge distributions. A fruitful future study might assess the impacts of alternative loss functions tailored to specific decision problems. Additionally, other applications may require sampling approaches other than the bridge sampling employed here (e.g. Yao *et al* 2017).

We caution that our analysis focuses on NAO index as a covariate for the storm surge statistical model parameters, but there may well be other useful predictors for modulating surge. It is a, perhaps, counterintuitive result that two sites on opposite sides of the Atlantic display similar behavior in storm surge return levels response to changing NAO index (cf figure 2) when one might expect to see opposite effects for the two sites. We hypothesize that this is due to the fact that in our simple single covariate model, any non-stationarity must be attributed to NAO index. Future work might consider incorporating other potential predictors, to test for additional drivers of storm surge non-stationarity.

We focus on the 100 year return level, but provide results for other return periods in the supplementary material. Higher return levels likely require more data for the same constraint, and fewer data for lower return levels. Furthermore, we find tighter constraint on the 100 year return level in the BMA ensemble, as a result of convergent projections from the four candidate models. This may not always be the case, and implementing our BMA approach with more diverse sets of candidate models is an important avenue for future work. Combining information across model structures using BMA can be of use in decision-making by more efficiently integrating the available information, and tighter constraint on projected flood hazard can help to avoid potential over-/under-protection regrets. Finally, many previous statistical treatments of storm surge hazard have made a single-model assumption (e.g. Grinsted *et al* 2013). Our results suggest that this may yield an overestimate of the range in projected flood hazard, so it is important to formally assess the impacts of those assumptions.

5. Conclusions

We present a framework for incorporating model structural uncertainty into estimates of coastal surge level probability, using two long tide gauge records to evaluate the impacts of data availability on our results. Our analysis indicates that previous work using a stationary Poisson process/generalized Pareto distribution modeling approach may underestimate the upper tails of flood hazards, and overestimate the uncertainty range. Discarding models on the basis of performance metrics (table 2) or by assuming a single model structure neglects model structural uncertainty that may be captured through our BMA approach. Our results highlight the impacts of neglecting key modeling uncertainties on estimates of storm surge return levels, and are of practical use to provide a more complete picture of decision-relevant information for the management of coastal flood risks.

Acknowledgments

We gratefully acknowledge Benjamin Lee, Alexander Bakker, David Johnson, Jordan Fischbach, Debra Knopman, Neil Berg, Nathan Urban, Ryan Sriver, Nancy Tuana, Robert Lempert, Murali Haran, Robert Ceres, Robert Nicholas, Chris Forest and Jesse Nusbaumer for valuable inputs. This work was co-supported by the National Science Foundation through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507; the National Oceanic and Atmospheric Administration (NOAA) through the Mid-Atlantic Regional Integrated Sciences and Assessments (MARISA) program under NOAA grant NA16OAR4310179; and the Penn State Center for Climate Risk Management. The ENSEMBLES data used in this work was funded by the EU FP6 Integrated Project ENSEMBLES (Contract number 505539) whose support is gratefully acknowledged.

We are not aware of any real or perceived conflicts of interest for any authors. Any conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. Any errors and opinions are, of course, those of the authors. All results, model codes, analysis codes, data and model outputs used for analysis are freely available from https://github.com/tonyewong/MESS and are distributed under the GNU general public license.

ORCID iDs

Tony E Wong https://orcid.org/0000-0002-7304-3883
Vivek Srikrishnan https://orcid.org/0000-0003-0049-3805
Klaus Keller https://orcid.org/0000-0002-5451-8687

References

Akaike H 1974 A new look at the statistical model identification *IEEE Trans. Automat. Contr.* **19** 716–23

Arns A, Wahl T, Haigh I D, Jensen J and Pattiaratchi C 2013 Estimating extreme water level probabilities: a comparison of the direct methods and recommendations for best practise *Coast. Eng.* **81** 51–66

Buchanan M K, Kopp R E, Oppenheimer M and Tebaldi C 2015 Allowances for evolving coastal flood risk under uncertain local sea-level rise *Clim. Change* **137** 347–62

Buchanan M K, Oppenheimer M and Kopp R E 2017 Amplification of flood frequencies with local sea level rise and emerging flood regimes *Environ. Res. Lett.* **12** 64009

Bulteau T, Idier D, Lambert J and Garcin M 2015 How historical information can improve estimation and prediction of extreme coastal water levels: application to the Xynthia event at la Rochelle (France) *Nat. Hazards Earth Syst. Sci.* **15** 1135–47

Caldwell P C, Merrifield M A and Thompson P R 2011 Sea level measured by tide gauges from global oceans as part of the joint archive for sea level (JASL) from 1846 to 2017 (NODC Accession 0019568) (National Oceanographic Data Center, NOAA) (https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:0019568) (Accessed: 24 July 2017)

Ceres R, Forest C E and Keller K 2017 Understanding the detectability of potential changes to the 100 year peak storm surge *Clim. Change* **145** 221–35

Cid A, Menéndez M, Castanedo S, Abascal A J, Méndez F J and Medina R 2016 Long-term changes in the frequency, intensity and duration of extreme storm surge events in southern Europe *Clim. Dyn.* **46** 1503–16

Coastal Protection and Restoration Authority of Louisiana 2017 *Louisiana's Comprehensive Master Plan for a Sustainable Coast* (Baton Rouge, LA: Coastal Protection and Restoration Authority of Louisiana) p 168 (http://coastal.la.gov/wp-content/uploads/2017/04/2017-Coastal-Master-Plan_Web-Book_CFinal-with-Effective-Date-06092017.pdf)

Coles S, Bawa J, Trenner L and Dorazio P 2001 *An Introduction to Statistical Modeling of Extreme Values* vol 208 (London: Springer)

Dangendorf S, Arns A, Pinto J G, Ludwig P and Jensen J 2016 The exceptional influence of storm 'Xaver' on design water levels in the German Bight *Environ. Res. Lett.* **11** 054001

Fischbach J R, Johnson D R and Molina-Perez E 2017 *Reducing Coastal Flood Risk with a Lake Pontchartrain Barrier* (Santa Monica, CA: RAND Corporation) (www.rand.org/pubs/research_reports/RR1988.html)

Gelman A and Rubin D B 1992 Inference from iterative simulation using multiple sequences *Stat. Sci.* **7** 457–511

Grinsted A, Moore J C and Jevrejeva S 2013 Projected Atlantic hurricane surge threat from rising temperatures *Proc. Natl Acad. Sci. USA* **110** 5369–73

Haigh I D, Nicholls R and Wells N 2010a A comparison of the main methods for estimating probabilities of extreme still water levels *Coast. Eng.* **57** 838–49

Haigh I, Nicholls R and Wells N 2010b Assessing changes in extreme sea levels: application to the English Channel, 1900–2006 *Cont. Shelf Res.* **30** 1042–55

Hoeting J A, Madigan D, Raftery A E and Volinsky C T 1999 Bayesian model averaging: a tutorial *Stat. Sci.* **14** 382–417

Hunter J R, Woodworth P L, Wahl T and Nicholls R J 2017 Using global tide gauge data to validate and improve the representation of extreme sea levels in flood impact studies *Glob. Planet. Change* **156** 34–45

Johnson D R, Fischbach J R and Ortiz D S 2013 Estimating surge-based flood risk with the coastal Louisiana risk assessment model *J. Coast. Res.* **67** 109–26

Jones P D, Jonsson T and Wheeler D 1997 Extension to the North Atlantic oscillation using early instrumental pressure observations from Gibraltar and south-west Iceland *Int. J. Climatol.* **17** 1433–50

Karamouz M, Ahmadvand F and Zahmatkesh Z 2017 Distributed hydrologic modeling of coastal flood inundation and damage: nonstationary approach *J. Irrig. Drain. Eng.* **143** 1–14

Kass R and Raftery A 1995 Bayes factors *J. Am. Stat. Assoc.* **90** 773–95

Knighton J, Steinschneider S and Walter M T 2017 A vulnerability-based, bottom-up assessment of future riverine flood risk using a modified peaks-over-threshold approach and a physically based hydrologic model *Water Resour. Res.* **53** 10043–64

Kolmogorov A 1933 Sulla determinazione empirica di una lgge di distribuzione *Inst. Ital. Attuari. Giorn.* **4** 83–91

Kopp R, DeConto R M, Bader D, Hay C C, Horton R M, Kulp S, Oppenheimer M, Pollard D and Strauss B H 2017 Evolving understanding of Antarctic ice-sheet physics and ambiguity in probabilistic sea-level projections *Earth's Future* **5** 1217–33

Lee B S, Haran M and Keller K 2017 Multi-decadal scale detection time for potentially increasing Atlantic storm surges in a warming climate *Geophys. Res. Lett.* **44** 10617–23

Marcos M, Calafat F M, Berihuete Á and Dangendorf S 2015 Long-term variations in global sea level extremes *J. Geophys. Res. Ocean.* **120** 8115–34

McMillan A, Batstone C, Worth D, Tawn J, Horsburgh K and Lawless M 2011 *Coastal Flood Boundary Conditions for UK Mainland and Islands. Project SC060064/TR2: Design sea levels* (Bristol: Environment Agency) (www.gov.uk/government/uploads/system/uploads/attachment_data/file/291216/scho0111btki-e-e.pdf)

Menéndez M and Woodworth P L 2010 Changes in extreme high water levels based on a quasi-global tide-gauge data set *J. Geophys. Res. Ocean.* **115** 1–15

Meng X L and Wing H W 1996 Simulating ratios of normalizing constants via a simple identity: a theoretical exploration *Stat. Sin.* **6** 831–60

Mengel M, Levermann A, Frieler K, Robinson A, Marzeion B and Winkelmann R 2016 Future sea level rise constrained by observations and long-term commitment *Proc. Natl Acad. Sci. USA* **113** 2597–602

Moftakhari H, AghaKouchak A, Sanders B F, Matthew R A and Mazdiyasni O 2017 Translating uncertain sea level projections into infrastructure impacts using a Bayesian framework *Geophys. Res. Lett.* **44** 11,914–21

Nauels A, Meinshausen M, Mengel M, Lorbacher K and Wigley T M L 2017 Synthesizing long-term sea level rise projections—the MAGICC sea level model v2.0 *Geosci. Model Dev.* **10** 2495

Nicholls R J and Cazenave A 2010 Sea level rise and its impact on coastal zones *Science* **328** 1517–20

NOAA 2017 NOAA Tides and Currents: Sewells Point, VA–Station ID: 8638610 *National Oceanic and Atmospheric Administration (NOAA)* (https://tidesandcurrents.noaa.gov/stationhome.html?id=8638610) (Accessed: 17 February 2017)

Orton P M, Hall T M, Talke S A, Blumberg A F, Georgas N and Vinogradov S 2016 A validated tropical-extratropical flood hazard assessment for New York Harbor *J. Geophys. Res. Ocean.* **121** 8904–29

Pugh D T and Vassie J M 1979 Extreme sea levels from tide and surge probability *Proc. 16th Conf. Coastal Engineering (Hamburg)* vol 1 pp 911–30

Rijkswaterstaat 2017 *Ministry of the Interior* (the Netherlands) (Accessed: 16 December 2018) (http://live.waterbase.nl/)

Robert C P 2007 *The Bayesian Choice* (New York: Springer)

Roeckner E *et al* 2003 The atmospheric general circulation model ECHAM5. *Part I: Model description, Report* 349 (Hamburg: Max Planck Institute for Meteorology)

Schwarz G 1978 Estimating the dimension of a model *Ann. Stat.* **6** 461–4

Smirnov N 1948 Table for estimating the goodness of fit of empirical distributions *Ann. Math. Stat.* **19** 279–81

Spiegelhalter D J, Best N G, Carlin B P and van der Linde A 2002 Bayesian measures of model complexity anf fit *J. Roy. Stat. Soc. B* **64** 583–639

Stephenson D B, Pavan V, Collins M, Junge M M and Quadrelli R 2006 North Atlantic oscillation response to transient greenhouse gas forcing and the impact on European winter climate: a CMIP2 multi-model assessment *Clim. Dyn.* **27** 401–20

Tawn J A and Vassie J M 1989 Extreme sea levels: the joint probabilities method revisited and revised *Proc. Inst. Civ. Eng.* **87** 429–42

Tebaldi C, Strauss B H and Zervas C E 2012 Modelling sea level rise impacts on storm surges along US coasts *Environ. Res. Lett.* **7** 14032

Vihola M 2012 Robust adaptive metropolis algorithm with coerced acceptance rate *Stat. Comput.* **22** 997–1008

Wahl T, Haigh I D, Nicholls R J, Arns A, Dangendorf S, Hinkel J and Slangen A B A 2017 Understanding extreme sea levels for broad-scale coastal impact and adaptation analysis *Nat. Commun.* **8** 16075

Wong T E, Bakker A M R and Keller K 2017a Impacts of Antarctic fast dynamics on sea-level projections and coastal flood defense *Clim. Change* **144** 347–64

Wong T E, Bakker A M R, Ruckert K L, Applegate P, Slangen A and Keller K 2017b BRICK0.2, a simple, accessible and transparent model framework for climate and sea-level projections *Geosci. Model Dev.* **10** 2741–60

Wong T E and Keller K 2017 Deep uncertainty surrounding coastal flood risk projections: a case study for New Orleans *Earth's Future* **5** 1015–26

Yao Y, Vehtari A, Simpson D and Gelman A 2017 Using stacking to average Bayesian predictive distributions (http://arxiv.org/abs/1704.02030)