Supplementary Information for

Trends in Europe storm surge extremes match the rate of sea-level rise

Francisco M. Calafat, Thomas Wahl, Michael Getachew Tadesse, Sarah N. Sparrow

Contents

S1. Low reliability of single-site analyses

Here, we show that the low statistical power inherent to site-by-site analyses will often lead to unreliable results when estimating trends in extremes by: 1) reducing the chance of detecting a true trend due to high estimation uncertainty; 2) overestimating the magnitude of the trends when these are detected (the so-called winner's curse[28]); and 3) detecting spurious trends (i.e., finding a trend that in fact does not exist). Here, the term detection is used in the sense of reaching statistical significance, whichever the threshold. Inflated trend estimates arise because single-site analyses generally only have the power to detect large trends and, thus, smaller trends will only be detected if they are amplified, due to chance, by sampling variability. In the following we use two experiments based on simulation to illustrated the issues outlined above and show how our BHM is much less affected by these issues. As a measure of model skill, we use fractional differences (FDs), which are defined as $\mathrm{FD} = |(x_{\mathrm{true}} - \hat{x})/x_{\mathrm{true}}|$ where x_{true} and $\hat{x}$ are the true and estimated values of the μ trend, respectively. Small FDs indicate high skill (a value of zero denotes a perfect match).

In the first experiment, we generate a synthetic tide gauge data set by sampling a simulated max-stable process (with a prescribed μ trend pattern) at the times and locations of the real tide gauge observations. We then estimate the trend in μ at each site by fitting a single-site GEV model with a time-varying μ to the synthetic tide gauge data. We only consider tide gauge records that have at least 40 valid annual maxima in the period 1960-2018. For comparison, we also fit our BHM to the same synthetic data set. Note that by analysing only one realization of a max-stable process (as opposed to averaging over multiple realizations), we are able to quantify the effects of sampling variability on the trend estimates. We find that the trend estimates from the BHM are much closer to the true value than estimates from the single-site GEV model at most stations (Fig. S1a,b), as indicated by the much lower FDs (median value of 0.3 for the BHM compared to 1.0 for the single-site GEV model). Comparing the trend values at sites where the posterior mean is at least 1 standard deviation (SD) away from zero (Fig. S1c) reveals the issue of trend inflation affecting the estimates from the single-site GEV model. Note how the single site model systematically overestimates the magnitude of the trends at most stations. In contrast, estimates from the BHM are generally close to the true values, indicating that the BHM is able to separate the effect of sampling variation from the true long-term trend. Another point worth emphasizing is the fact that the SDs associated with the trend estimates are, on average, more than two times larger for the single-site GEV model, which highlights the higher ability of the BHM to detect trends. Finally, it is important to recognize that this comparison is just for one realization of a max-stable process. Results for other realizations may show smaller or larger differences between the two models, depending on how sampling variability projects onto the trends, but the issues associated with the single-site GEV model will still persist.

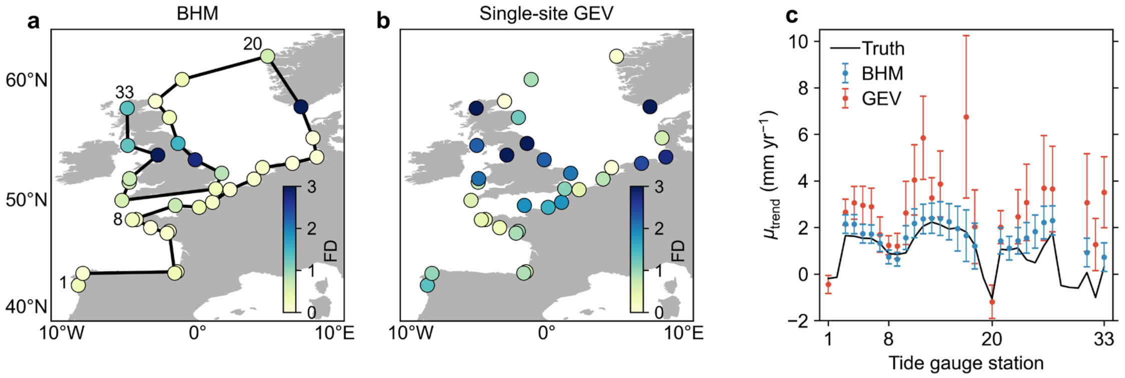

Figure S1. **Low ability of single-site GEV models to estimate trends**. Fractional differences (FDs) between the true trend values in the GEV location parameter (μ) and the trend values estimated by (**a**) the BHM and (**b**) a single-site GEV model. Small FDs indicate high skill. **c**, Direct comparison of the trends along with the associated uncertainties (1 SD). Only trends that are at least 1 SD away from zero are shown in **c**. The trends in **c** are plotted following the coastline in the order indicated by the black line in **a**, starting at the station denoted by the number 1. Numbers along the *x*-axis refer to the identification number shown in **a**.

The second experiment aims to show how the low statistical power of single-site GEV models can lead to the detection of spurious trends. To this end, we generate another synthetic tide gauge data set by sampling a max-stable process, but this time we assume a stationary process (i.e., μ is constant in time). As in the first experiment, we only consider tide gauge records with at least 40 values in the period 1960-2018. We then fit both the single-site GEV model and our BHM to the synthetic data and compare the trend estimates (Fig. S2). The trend estimates from the single-site model show a large scatter around zero, with values ranging from -3.5 to 2.6 mm yr^{-1}. In contrast, the BHM estimates lie within a narrow range (-0.1 to 0.4 mm yr^{-1}) centered on zero. There are 7 stations where estimates from the single-site GEV model are more than 1 SD away from zero, illustrating how sampling variability can cause low-powered methods to detect trends when in fact they are absent. BHM estimates fall well within 1 SD of zero at all stations, indicating a significantly reduced chance of finding spurious trends.

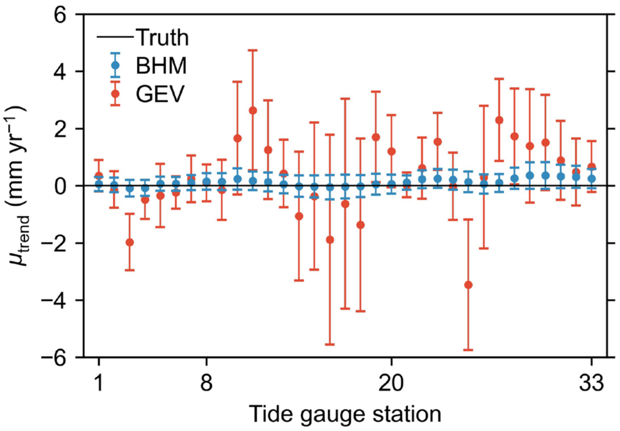

Figure S2. **Spurious trends in single-site GEV models**. Trends in the GEV location parameter (μ) as estimated from synthetic tide gauges by the BHM (blue) and a single-site GEV model (red). The synthetic data were generated under a stationary max-stable process, and thus the true value of the trends is zero. The error bars represent posterior SDs. The trends are plotted following the coastline in the order indicated by the black line in Fig. S1a, starting at the station denoted by the number 1. Numbers along the *x*-axis refer to the identification number shown in Fig. S1a.

S2. Sensitivity to prior distributions

To assess the sensitivity of our results to prior choices, we compare trend estimates based on different priors for the standard deviations ($\gamma_{\mu}, \gamma_{\mu_{00}}$) and length scales ($\rho_{\mu}, \rho_{\mu_{00}}$) of the Gaussian processes used to model the temporal evolution of μ (see ref. 13 for model equations). We compare the following combinations of priors for the standard deviations/length scales: 1) half-$\mathcal{N}(0,1)$/half-$\mathcal{N}(0,0.5)$; 2) half-$\mathcal{N}(0,4)$/half-$\mathcal{N}(0,2)$; and 3) half-t(4)/Inv-$gamma(6,4)$, where half-t(ν) denotes the Student's t-distribution with ν degrees of freedom. The first combination corresponds to the actual priors used in this study. A half-normal distribution is one of the recommended priors for scale parameters in hierarchical models[47] as it enables us to constrain the value of the parameter from above while allowing it to be arbitrarily close to zero. The assignment of priors to the length scales needs careful consideration because the likelihood for such parameters can become non-identified if the priors are too diffuse. In this regard, we should note that there is no information in the observational data to characterize length scales above the maximum distance between tide gauge stations. The priors should encode this information. In the following, square brackets denote 5-95% credible intervals (CIs). Spatially averaged μ trends in R1 (regions defined as in Fig.1) for the three combinations of priors are, respectively: 1.1 mm yr^{-1} [0.3, 1.8], 1.0 mm yr^{-1} [0.2, 1.8], and 1.0 mm yr^{-1} [0.3, 1.9]. In R2, the trends are: -1.2 mm yr^{-1} [-1.9, -0.6], -1.2 mm yr^{-1} [-1.9, -0.6], and -1.2 mm yr^{-1} [-1.9, -0.6]. Hence, the trend estimates as well as the associated CIs are virtually the same for all prior combinations, indicating low sensitivity to the choice of the prior for these parameters.

We have also tested the sensitivity of the attribution analysis to the prior for the amplitude of the fingerprint of external forcing (β_{ext}). We have compared estimates based on the following priors: 1) $gamma(1.5,0.5)$; 2) half-$\mathcal{N}(0,3)$; and 3) $\mathcal{TN}(1,2,0,\infty)$, where $\mathcal{TN}(a,b,l,u)$ is a truncated normal distribution with location a, scale b, and lower and upper limits l and u. The spatially averaged contribution from anthropogenic forcing to μ trends in R1 for the three different priors is, respectively: 0.2 mm yr^{-1} [-0.3, 0.9], 0.2 mm yr^{-1} [-0.3, 0.8], and 0.2 mm yr^{-1} [-0.3, 0.9]. In R2, the contributions are: 0.6 mm yr^{-1} [0.0, 1.7], 0.6 mm yr^{-1} [0.0, 1.6], and 0.6 mm yr^{-1} [0.0, 1.6]. Hence, the posterior means are the same for the three priors and the CIs are very similar, indicating low sensitivity to the choice of reasonable priors on β_{ext}.

S3. Robustness of the BHM trend estimates

In ref. 13, we validated the BHM extensively, but most of the emphasis was on the time-mean properties of the extremes and we did not explicitly assess the ability of the model to estimate trends in the GEV location parameter. Hence, as an additional assessment of the model, here we present the results of an experiment that evaluates the skill of the model to infer the pattern of μ trends from a sparse tide gauge data set. In this experiment, we first simulate a total of 15 spatiotemporal processes under the same model as the one used to fit the observational data (i.e., a non-stationary max-stable process) and sample each process at exactly the same times and locations as the real tide gauge record. All of the 15 realizations are based on the same model parameters (set equal to those inferred from the observations) and contain the same trend pattern (in μ), and thus they can be viewed as random samples from the same distribution. Next, we fit the BHM to each one of the 15 synthetic tide gauge data sets, average the 15 estimated trend patterns (this is to minimize the influence of sampling variability), and compare the resulting pattern with the true trend pattern. Note that this experiment assumes a perfectly adequate model, and thus any differences between the true and estimated trends are due to the sparseness of the tide gauge data (aside from Monte Carlo error).

We find that the BHM is capable of characterizing the trend pattern with high accuracy (Fig. S3), despite the sparseness of the tide gauge record. Both the spatial structure and magnitude of the true and estimated trends are remarkably similar, and FDs are low at most locations with a median value of 0.21 over all data sites. Such FD values imply that estimates of the μ trend at individual locations are accurate, on average, to within ~21% of the true value.

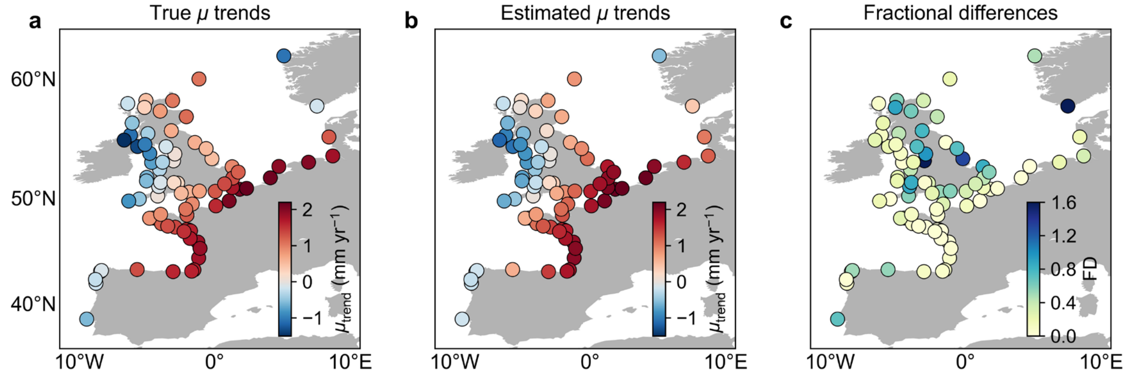

Figure S3. **Validation of the BHM using simulated data**. True (**a**) and estimated (**b**) μ trends at tide gauge locations. **c**, Fractional differences (FDs) between the true and estimated μ trends (small FDs indicate good agreement). The estimated trends shown here represent the mean trend over the 15 realizations of simulated surge annual maxima used in this validation experiment.

S4. Resolvability of the anthropogenic signal

Here, we present the results of an experiment that quantifies the ability of the BHM to identify the fingerprint of anthropogenic forcing under different levels of forcing. In this experiment, we generate data under a max-stable process with a time-varying μ parameter that evolves as the sum of a trend pattern related to anthropogenic forcing (i.e., the fingerprint) and a pattern associated with internal climate variability. To ensure that the results of this experiment are transferable to the real world, we set the trend pattern related to internal variability equal to the observational pattern shown in Fig. 3b. For the fingerprint of external forcing we choose three different levels of intensity where the observational fingerprint (Fig. 3a) is scaled by a factor of 2 (scenario1), 1 (scenario2), and 1/2 (scenario3). We sample the three simulated max-stable processes at the same times and locations as the real tide gauge record and fit BHM2 to the synthetic tide gauge data in order to estimate the anthropogenic contribution. For each scenario, we repeat this procedure for seven different realizations of the max-stable process to take account of the effect of sampling variation. If the BHM is able to identify the fingerprint, the posterior mean for the anthropogenic contribution averaged over the seven realizations should be close to the true contribution. The assessment is conducted in terms of trends spatially averaged over R1 and R2, just like we do in Fig. 3.

We find that the spatially averaged contribution from anthropogenic forcing to μ trends in R1 for the three scenarios (1,2, and 3) is, respectively: 0.35 mm yr^{-1} [0.01, 0.97], 0.17 mm yr^{-1} [-0.20, 0.74], and 0.06 mm yr^{-1} [-0.61, 0.74]. The estimated contribution in R2 is: 1.23 mm yr^{-1} [0.21, 2.78], 0.69 mm yr^{-1} [0.08, 1.81], and 0.45 mm yr^{-1} [-0.09, 1.47]. Both the posterior means and CIs have been computed as the average over the seven realizations. The true contribution in R1 for the three scenarios is, respectively, 0.38, 0.19, and 0.09 mm yr^{-1}, whereas in R2 it is 1.24, 0.62, and 0.31 mm yr^{-1}. Hence, in scenario1 we are able to estimate the contribution form anthropogenic forcing with good accuracy. In scenario2, posterior means are also close to the true trend values, but there is a slight tendency to overestimate. In scenario3, the overestimation is more significant, indicating that the anthropogenic fingerprint may not be resolvable at this level of forcing. We note that the CIs are fairly wide in all three scenarios, even though the posterior means in scenarios 1 and 2 are close to the true values.

The results of this experiment suggest that, assuming that the anthropogenic fingerprint is well simulated by the surge model (the validation presented in the next section suggests that this is the case), our best estimate (i.e., the posterior mean) of the anthropogenic contribution to the historical trends in surge extremes is reliable. However, the results above also suggest that the size of the anthropogenic signal in the observations (scenario 2) is close to the limit of what can be resolved using our BHM, largely because of the relatively large contribution from internal variability. This limit to resolvability manifests as an overestimation of the anthropogenic contribution when the contribution is small, such as in scenario 3.

S5. Validation of the storm surge model

In order to assess the performance of the surge model used to generate the ensemble of simulations, we fit the BHM to the annual maxima simulated with the surge model based on the ERA5 predictors and compare estimates of trends in the GEV location parameter μ with estimates from tide gauge observations (Fig. S4). Estimates are computed for the ERA5 period 1979-2018. In comparing the two estimates, it is important to note that we only require the surge model to capture the shape of the spatial pattern (i.e., the second order statistics) not the absolute values. This is because when quantifying the contribution from external forcing to the μ trends, we only assume that the spatial structure of the simulated response pattern is similar to that of the true pattern, while the amplitude and mean of the pattern are inferred from the observations. We find that the spatial structure of the pattern of μ trends based on data from the storm surge model is overall very similar to the one inferred from the tide gauge observations (only a few sites in Scotland show a discrepancy), with a spatial correlation of 0.85 (Fig. S4a,b). The estimates based on the surge model data, however, tend to underestimate the trends (median factor of 1.8). Plotting the trend estimates on top of one another with the stations sorted following the coastline (Fig. S4c) further illustrates the good match between the two spatial patterns. The good agreement in terms of the spatial structure of the μ trends gives us confidence that we can use the storm surge model to estimate the fingerprint of external forcing on surge extremes.

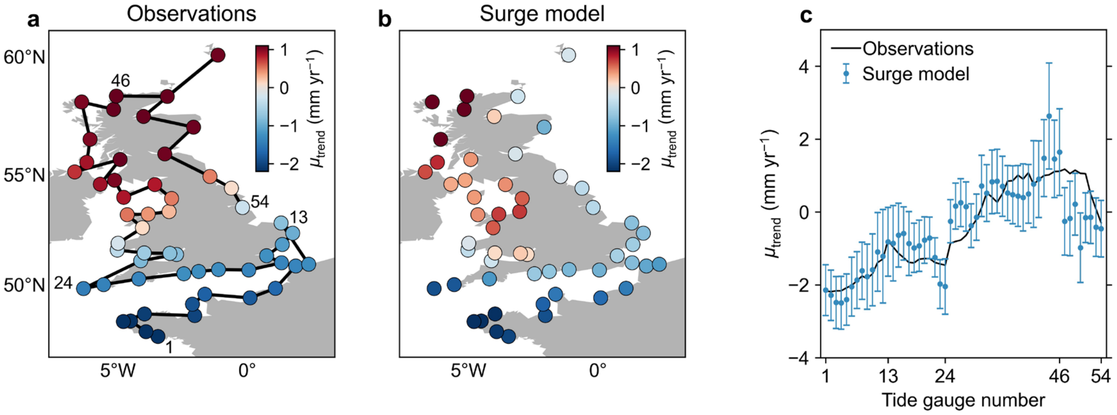

Figure S4. **Validation of the storm surge model**. Estimates of trends in the GEV location parameter μ derived by fitting the BHM to annual maxima from (**a**) tide gauge observations and (**b**) a storm surge model for the period 1979-2018. The pattern of μ trends based on the data from the storm surge model has been scaled by a factor of 1.8

and its mean set equal to the mean of the observational pattern (i.e., multiplicative and additive biases have been removed) in order to emphasize the spatial structure of the pattern. **c**, Direct comparison of the trends shown in **a** and **b**. The uncertainties (1 SD) associated with the trend estimates based on data from the storm surge model are also shown. The trends in **c** are plotted following the coastline in the order indicated by the black line in **a**, starting at the station denoted by the number 1. Numbers along the *x*-axis refer to the identification number shown in **a**.