Responses to the comments of Referee 2

Seuri, Basilio Kuosmanen
Magnus, Hieronymus

September 2025

1 Responses

First and foremost, we are sincerely grateful for the time and thought Referee 2 has taken in to reviewing and commenting on the manuscript. The comments where both constructive and insightful, all of which we believe will led to the reassurance of the robustness of our results, and an much improved presentation of the overall science. Our responses to the comments made by Referee 2 are presented below.

1.1 Overall assessment

I believe that the present paper is unsuitable for publication. There may be something useful here, but in its present state referees/readers spend too much time attempting to see past presentational problems (errors of grammar, expression, syntax, inconsistent notation, typos, poor graphics, incorrect references, . . .) that undermine any understanding of the appropriateness of the methods and reliability of the substantive results.

A general comment: while it is good that the authors provide extensive code to document what they did, I do not think it reasonable to expect reviewers to have to parse code (possibly in a language they do not know) when refereeing a paper. A well-prepared and readable (and not too long!) supplement in English is also needed. I did not read all the code, but there seem to be many points treated in it that are not mentioned in the paper itself, and which might, if properly explained, reassure the reader that all is well (despite concerns generated by the paper).

1.1.1 Summary of contribution

From what I can understand (but I may be wrong for the reasons mentioned above), the paper applies standard methods from the statistics of extremes (or extreme-value theory, EVT) and Bayesian hierarchical modelling (BHM) to data series of varying lengths on extreme sea levels in parts of the Baltic and North Seas.

It is claimed that hierarchical models that take account of spatial relationships between the tide gauge sites perform better than other such models, in terms of better predicting the results from individual maximum likelihood fits to the tide series maxima (which are treated as ground truth). This is hardly surprising, even if the extent of the improvement is marked (but see below), and the application to this particular region may be novel.

Comment 1. *At line 519 the authors claim that what is novel is the use of coordinate-dependent random functions and random coefficients using priors based on kernel density estimates taken from the same data. If this statement is true, then the random functions and coefficients might not be identifiable (but so far as I can see only the coefficients, and not the functions themselves, are random), and the gain in precision would be illusory because it would partly stem both from (improperly) using the data twice (or three times if the comparison against the "baseline" maximum likelihood fit is included).*

Response 1. We thank the referee for highlighting these concerns, and we shall conduct a comprehensive revision of the paper to improve clarity and presentation. We stress that the we use Gaussian processes (see Section 2.2), which are stochastic processes, and thus random functions by definition. The mean functions of the Gaussian processes are indeed modeled as a linear combination of an intercept and two random coefficients, one for each coordinate axis. We refer to the below discussion for a detailed response pertaining the remarks regarding both the priors and **"improperly using the data twice or three times"**.

Comment 2. *On page 3 we are told that $\zeta(s)$ is a weakly stationary (not weak stationary!) series corresponding to the observed sea level. However the sea levels are not weakly stationary, as they are subject to tidal effects, seasonality and the lunar cycle — indeed, the authors say this at line 75. The block maxima used may be stationary (though we are told in the initial sentences on page 1 that the mean sea level is rising), but the underlying data certainly are not. It would suffice here to simply say that the maxima are treated as realisations of*

(conditionally) independent GEV random variables, not invoke clearly incorrect reasoning to justify this. It would be necessary to check this assumption using QQ plots or other suitable methods (not density plots, see below) for the assessment of fit.

Response 2. We are in agreement regarding the misuse of the term weak stationary. Per the recommendations of the referee, we suggest removing the term, and adding the sentences : "We assume that the block maxima at each location is a realizations of (conditionally) independent GEV random variables". From there we check the previously stated assumption using QQ, PP and return levels plots, for each station in the training data. We suggest that the actual plots be contained in the supplementary material, and simply comment on the results of the assumption validation in Section 3.

Comment 3. *The fitted models treat the annual maxima as independent, conditional on the parameter surfaces. However this is untrue, because there may be common causes for annual maxima (e.g., particular tidal or meteorological conditions). This dependence is mentioned (line 90) but the opposite statement is also made (lines 217, 223). It is not clear whether this matters in terms of point estimation (it appears that the goal is improved estimation of return levels at individual locations, as well as the possibility of prediction at ungauged locations) but the authors do not seem to realise that not taking account of this dependence will mean that confidence intervals for the return levels are too short, since the equivalent amount of 'independent information' in the data may be smaller than is assumed by their models. I say 'may be' because I could find no attempt to check this in the paper or supplement.*

Response 3. We apricate the concerns raised by the referee in this comment. We would first like to note that the fitted latent spatial process models, **Latent** and **Hilbert**, treat **new** annual maxima as conditionally independent given the parameter surfaces. This implies that the posterior predictive distributions are pairwise independent, for any pair of distinct stations (locations). Moreover, we would like to highlight that the two models capture dependency only at the level of the parameter surfaces, via the (Hilbert space approximated) Gaussian processes. Furthermore, we assume that this dependency is somewhat relates to the distance between any two distinct pairs of stations (gauged or ungauged).

We agree with the referee regarding the existence of common causes for annual maxima in the region of interest (see line 90). Nevertheless, we recall that independence and conditional independence are fundamentally different concepts,

by definition. We emphasize that the conditional independence stated in line 217 follows directly from the definition of a latent spatial process model: **once the parameter surfaces are known, knowledge of $Z(s_j)$ provides no additional information about $Z(s_i)$ for any $s_i \neq s_j \in S$.** This restriction is inherent to the latent spatial process model design and is explicitly acknowledged in lines 223–224.

Concerning the use of confidence intervals in our models, only the Baseline model employs confidence intervals derived from bootstrapping (see our response above). The Bayesian models instead use credible intervals (see Section 2.1.2). All Bayesian models in the paper are designed to estimate the marginal distributions of the block maxima at each station, not their joint distribution across stations. Since extremal dependence is not identifiable from the marginal distributions, there is no requirement to account for cross-station dependence when estimating the marginal return levels and their associated uncertainty. The reported HDIs therefore provide a valid uncertainty quantification for the station-level marginal return levels.

Additionally, in the reference "Statistical Modeling of Spatial Extremes," both advantages and disadvantages of latent spatial process models are discussed: the ability to estimate marginal properties of extremal distributions fall under the former, whereas the concerns and limitations raised by the referee and ourselves fall under the latter.

A. C. Davison. S. A. Padoan. M. Ribatet. "Statistical Modeling of Spatial Extremes." Statist. Sci. 27 (2) 161 - 186, May 2012.
https://doi.org/10.1214/11-STS376

Comment 4. *The parameters of the priors used in the hierarchical specification seem to be estimated from the individual maximum likelihood estimates. It is not clear from the text how this is done (Figure 2 seems to have something to do with it) but this amounts to using the data twice. The authors claim that this is an empirical Bayes approach, but such an approach would not generally specify both the mean and the variance of the prior using the original data, as is done in this paper. Moreover the assigned priors seem to treat the parameters as independent a priori, when they have been estimated from common data using maximum likelihood, again giving the impression of higher information content than the data actually contain.*

Response 4. We suggest adding the additional error term $\varepsilon_{corr} \sim \mathcal{N}(0, \sigma^2)$ to each of the three GEV parameters this would make each of the three parameters

correlated via ε_{corr}, thus reflecting the dependency between the MLE parameters. Regarding the priors we suggest preforming a prior sensitivity analysis using the method outline in "Detecting and diagnosing prior and likelihood sensitivity with power-scaling" with respect to both the priors of the mean vectors and the error terms of the covariance functions. The results of these prior diagnostics can effortlessly be presented as tables in the supplementary material, and fittingly stated in the beginning of the results section. If the priors are (effectively) non-informative, then the posterior is primarily determined by the data, thus both the mean and the HDIs of the posteriors are comparable to the CI of the MLE, since the CI of the MLE is also data dependent.

Kallioinen, N., Paananen, T., Bürkner, PC. et al. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. Stat Comput 34, 57 (2024). https://doi.org/10.1007/s11222-023-10366-5

Comment 5. *The overall effect of (b) and (c) is to give tighter confidence sets than would be justified by a more appropriate analysis. Since this is a key selling point of the results, it is difficult to regard them as entirely reliable. Why not simply use a BHM with vague priors, not those taken from the data?*

Response 5. We thank the referee for raising this important concern. As detailed in our responses above, we have clarified the role of the priors in our Bayesian hierarchical models, emphasized that the Bayesian analyses use credible intervals rather than confidence intervals, and explained that the models are designed to approximate the marginal distributions, with dependence modeled only in the parameter space. For good measure, we have suggested (see our response above) adding a sensitivity analysis of the priors to reassure the reader of the robustness of our results.

Comment 6. *The handling of missing data is unclear. According to line 241 missing data are added — does this mean that known values are deleted (and if so why?), or that missing ones are imputed? What imputation technique, if any, was used? We are told that different series are of different lengths, but this is a common problem. What is unusual is to have around 70% of the maxima missing (line 247), presumably because some series are much longer than others.*

Response 6. We suggest elaborating on the handling of missing data in the revised version. We assume that the missing data is missing at random, which we motivate by the fact that missing values are probably dependent on the station and the years, and not the extreme events themselves. The missing data is imputed using

Bayesian imputation, which is preformed automatically with in the PyMC python package.

Comment 7. *The representativeness of the test/train split is unclear. It appears from Figure 1 that long stretches of coastline have no test stations. Why?*

Response 7. We use all stations in the GESLA dataset that have at least 20 years of data. We know that more sources exist, but trying to gather them is a lot of work and beyond the scope of what we can hope to achieve here.

Comment 8. *Line 273 mentions a bootstrap. There are many possible bootstraps: what was used, precisely? Parametric? Nonparametric? Treating all observations as independent? Treating years as independent but allowing for spatial dependence? Resampling blocks of years?*

Response 8. Here we use a simple nonparametric bootstrapping, and resample the observed data with replacement, this is done for each station independently. For each station, we generate 10,000 synthetic datasets using the collection of annual maxima's and resampling with replacement. From there we fit the resampled data to a GEV and obtained the MLEs parameters. The approach view the annual maxima's at each station as independent observations, and does not account for spatial dependency. No additional parametric assumptions are made beyond the GEV likelihood.

Comment 9. *The authors choose to use kernel density estimates (KDEs) to compare datasets. This is statistically poor, both because it does not show the individual data and because KDEs have poor tail behaviour and are unreliable unless based on sample sizes of hundreds of independent data. In particular, the authors claim (lines 276–277) that the training data adequately match the regional block matrix, but this claim is unjustified without some indication of the uncertainty of the estimates. QQ plots or two-sample tests would be more appropriate, even if themselves inadequate due to dependence in the data.*

Response 9. The comparison between datasets was made with respect to the MLEs of the GEV parameters rather than the raw annual maxima. However, because the number of annual maxima for stations not included in the training data is small, such comparisons are unreliable. Therefore, we propose to remove these comparison plots and instead show only the KDEs of the MLEs for stations in the training data.

Comment 10. *Problems with the Hilbert specification: at line 317 it is not clear what x_1, x_2, x_3 represent.*

It appears (line 346) that 2640 basis functions are being used to model variation in three parameters at 3083 elements of a matrix (lines 246–247), which seems excessive.

There also seems to be no attempt to assess the sensitivity of the results to the choice of priors and/or their parameters, and the many other apparently arbitrary choices made when fitting the models; this holds true also for the other models. Hence the results cannot be regarded as robust.

Response 10. The Hilbert model uses Hilbert space approximated Gaussian processes instead of Gaussian processes. The approximation is made in the 3-dim Cartesian coordinate system, where $(x_1, x_2, x_3) = \Psi(\text{lat}, \text{lon})$ is obtained from latitude and longitude via the standard geographic-to-Cartesian transformation, which we denote by Ψ. We use this transformation because the tide gauge data are provided only in latitude/longitude, and without details regarding the local coordinate systems.

The basis function are not used as model parameters. Instead the basis functions are used to approximate the Covariance matrix over the entire coordinate grid, the number of basis function is related to size of the coordinate grid (domain), the length scale of the covariance function, and how accurate this approximation should be at the boundary of the domain. Although the 2640 basis functions seems excessive, they are justified by the need to cover the coordinate grid.

We apply posterior predictive checks following the recommendations outlined in Bayesian Workflow and Visualization in Bayesian workflow. Per a previous comment we suggest adding a prior sensitivity analysis.

Comment 11. *The authors claim to be able to estimate 10,000-year return levels accurately. However the prior handling of the data involves fitting a 20-year centred moving average to remove trend. So, looking ahead to the estimation of such return levels for (say) the year 2040: how can the results be used, since the 20-year centred moving average for 2040 is (as yet) unavailable?*

Response 11. We appreciate the concern raised by the referee. We don't claim to be able to estimate 10000 year return levels accurately. This is simply because there is no accurate information to test it against. However, we can say that our method gives similar results as MLE, with the addition of tighter uncertainty intervals. For future projections we would use the industry standard and consider a

2040 return level as a sum of a projected mean sea level for that year and return level curve based on the calculations done here. Mean sea level projections are widely available e.g. from the IPCC for multiple different future emission scenarios with a one degree horizontal resolution. For Sweden they are available with even higher resolution through SMHI.

1.1.2 Other issues

Comment 12. *Grammar and syntax are poor, with many typos and incomplete sentences.*

Response 12. We will make a comprehensive revision of the language.

Comment 13. *Authors should distinguish passive and active citation of references.*

Response 13. We will distinguish passive and active citation of references in the revised version.

Comment 14. *Literature review is deficient. The earliest paper on spatial modelling of annual maximum sea levels I am aware of is Coles and Tawn (1990, Phil. Trans. R. Soc. Lond. A, Statistics of Coastal Flood Prevention), and there are many related papers since, e.g. Coles and Tawn (2005). See also relevant chapters in the Handbook of Ecological and Environmental Statistics (link). Several references are wrong or incomplete (e.g. Coles (2001) has one author; Davison et al. (2012) in Statistical Science lacks details; Beirlant et al. 2004; Dudley 2002; Robert is single-author, etc.).*

Response 14. We thank the referee for these hints, and we will add these citations in the revised version.

Various statements are false or misleading. For example:

Comment 15. *Line 109: Consistency and asymptotic normality of MLE for shape parameter holds when the* true parameter $> -1/2$*, not the estimate.*

Response 15. We agree with the referee and we propose changing from estimated to true parameter in the revised version. Additionally, we would could add the caveat: **Getting a best MLE estimate $\hat{\xi}_{ML}$ of the true parameter $< -\frac{1}{2}$, means that if the estimate is accurate the model might not be consistent. So in practice one would be skeptical of such locations regardless.** This caveat is further supported by the diagnostics (PP, QQ, and return level plots) from those stations look poor.

and

Comment 16. *Line 314: Common and Separate models do allow Bayesian interpolation to new locations (though perhaps not useful)*

Response 16. We are not sure what the referee means here. We recall that Common and Separate are not *spatial*-models, so **spatial-interpolation** is not applicable for any of the two models. We suggest that we change "Bayesian interpolation" to "Bayesian spatial interpolation" in the revised version.

1.1.3 Citation

doi:10.5194/egusphere-2025-1257-RC