Comments on "Technical note: Extending sea level time series for extremes analysis with machine learning and neighbouring station data"

Anonymous referee #2

September 8, 2023

Note: quotes from the pre-print are in blue.

1 General comments

The preprint discusses the artificial extension of a short (10 years) tide gauge record in the city of Halmstad, using longer records from neighbouring stations, in order to better estimate return levels of extreme sea levels in Halmstad. To do so, a statistical relationship between the return levels of Halmstad and its neighbouring stations is found in the overlapping 10-year period through linear regression and quantile regression forest.

I think the paper addresses a complicated, yet very interesting issue. The extension of a short record to better estimate extreme return levels is a delicate task, both from the statistical and physical/operational point of view. I think the paper lacks discussion on the limitations of the study, although many important issues are already mentioned in the text. Also, I believe that the explanation of the statistical method is somewhat confusing, and could be made much clearer for the reader. Suggestions in this direction are given in the specific comments.

For these reasons, my opinion leans toward a major review of the paper before it can be accepted.

2 Specific comments

(a) Lines 32-33: This technical note evaluates a machine learning (ML) method for extending the sea level time series obtained by a tide gauge of interest using a longer time series at a neighbouring tide gauge in the context of analysing sea level extremes.

This is particularly challenging from a statistical point of view. You should discuss more the hypothesis behind this work. You should always keep in mind that your objective is to have a more robust estimation of ESL based on your extrapolation method.

Which ESL can your algorithm reproduce ? Only the following ones:
- ESL that were observed during the short period (since LR is mostly driven by "average" values, and QRF cannot reproduce out-of-sample values)
- ESL that are associated with a trace on other tide gauges

Therefore, your algorithm cannot reproduce:
- ESL that were never observed (and it is very likely that they happened during the period that you try to simulate)
- ESL that leave no trace on neighbouring tide gauges.

The problem of "unobserved extremes" can be tested, as these might have left a footprint on the neighbouring tide gauges that is stronger than what was observed during the 2010-2020. If there are observed ESL of neighbouring stations outside the 2010-2020 period that are stronger than the ones observed in the 2010-2020 period, then it is likely that Halmstad also encountered extremes stronger than the ones of 2010-2020.

The issue of ESL leaving no trace on the neighbouring stations could also be tested, or at least discussed: did you see any ESL in the 2010-2020 period that the QRF and/or LR failed to reproduce ? What would they be linked to ? I suppose it would be very local events ?

These points are crucial and were not tested, I think you should test them or at least mention them clearly.

On the same topic: your confidence intervals on RLs will decrease as you use the extended time-window with sea-levels predicted from neighbouring stations (e.g., Figure 4). However, is this reduction justified, based on the previous observations ? It seems to me that the added error inherent to LR and QRF is not clearly included in your evaluation of RLs: it seems like the data you add is given the same value as the real observations. Please correct me if I am wrong on this point.

(b) *(This comment is related to the previous one.)* End of Table 2's caption:
Because of the short length of the testing period, we do not calculate the bias on the annual maxima.
Also lines 176-178:
As shown above, using a QRF or LR method, we can in principle reconstruct Halmstad sea levels back until 1891 for the period before observations became available in 2009 with reasonable confidence, using the station Hornbaek as a predictor, since this has the longest observed time series.

Since you cannot calculate the bias on annual maxima, what are your guarantees that your method will be able to reproduce extremes in the past ? I suppose it would be available from the bias on daily maxima ? However, in Table 2, the line with the largest computed biases on annual maxima is Ringhals-Hornbaek, and this line shows pretty low values of biases on daily maxima, suggesting a weak link between the biases on daily maxima and on annual maxima. Similarly, high values of r are not incompatible with high values of annual maxima, as confirmed by the Ringhals-Viken example. However, since we do not know if a few tens of centimeter is a large value, we are not able to assess, as a reader, if the method can be trusted.
You should explain why you have confidence in extrapolating towards extremes in the past. This demands statistical tests that you could try to do on your short sample (12 years ?). For instance, you could train/validate your model on 10 years and then test on the 2 remaining years, and repeat this operation by switching which years are used for training/validation and which years are used for testing. This is classical in machine learning. This would allow to evaluate, at least a little bit, how your model behaves on extremes for the Halmstad station. Lines 186-188: "When comparing the QRF method to LR, slightly better RMSE and r values are found for the LR, but when looking at higher percentile levels, the QRF results in higher corresponding values than for LR in all sets (not shown)."
This piece of information is probably more interesting than the r values showed in Table 2. I think you should focus on that.

(c) I believe that the expression "machine learning" is used in a confusing way throughout the manuscript. The QRF method is called "machine learning" in opposition to the LR, but in fact the LR is also a machine learning technique, only perhaps simpler and much more common than QRF. I therefore suggest that the expression "machine learning" be used, but only in the introduction and conclusion, and it should refer to *both* methods, LR and QRF. I recommend that the acronym "ML" be withdrawn from the manuscript and replaced by "QRF" to avoid ambiguity.

(d) Lines 85-87: Would the results change if the validation period was different (i.e., two years in the beginning instead of the end, or even two random years picked in the 10 years). Testing the sensibility to this choice would make the results even more robust.

(e) Lines 91-92: Based on each x/y predictor-reconstruction station pair, a linear equation is found using the least squares method as means of determining the best fit coefficients.

It could be nice to see the coefficients and try to interpret their meaning : I assume they would be positive (high sea level at one station means high sea-level at another station) and would reflect both
- the ratio of intensity of sea-level variations between stations
- the correlation between stations.

Even if you choose not to show the values of the coefficients in the final version of the article, I would like to see them as a sanity check, and I believe a little comment on our interpretation of the coefficients would be nice in the manuscript, even if you do not show them.

(f) Lines 92-93: It feels like the predictor for the linear regression and QRF could have been chosen to be a little bit more complex, revealing other time-space relationships between the stations. Since the sea level is sensitive to meteorological conditions, which are advected by the winds, the sea level at one station at time t might be better predicted if using time-lagged sea-level from another station, e.g. the sea-level at times $t\pm$ a few days (this could be tested very easily from your algorithm, at least for a couple of stations). Even time-delayed embeddings could be used, with sea-level at times $t - m$days, ..., $t - 1$day, t, $t + 1$day, ..., $t + n$days where m and n would have to be optimized. Although this might be out of the scope of the study, it should be tested or at least mentioned. But maybe you have reasons to believe that this would not be useful/necessary ?

(g) I believe section 2.2.2 and section 2.2.4 need to be clarified. Although I am not an expert of QRF, it seems to me that your formulation is misleading.

Lines 96-97 The QRF method yields a mean and a standard deviation for each predicted value (Breiman, 2001; Meinshausen, 2006. You should add equation to this sentence to make it clear. QRF estimates quantiles of a predictand, not average and standard deviation, therefore your formulation is confusing. What do you call "predicted value" ? You need to be specific and to use precise vocabulary. What you should do is to describe the method in detail, to help unfamiliar readers. I suggest to extend this paragraph and to add equations.

Lines 97-98 The QRF model is implemented using the MATLAB function TreeBagger where the regression method is based on a number of trees and minimum leaf size hyper-parameters. Reading the MATLAB documentation indicates that TreeBagger cannot be used alone, another function must be used to perform the regression. Did you use "quantilePredict" ? "predict" ? "fitrensemble" ? Something else ? Please specify this to allow reproducibility and help understanding.

Line 98 These parameters are here set to 500 and 1, respectively. Consulting the documentation on TreeBagger MATLAB method, it seems that 1 is the default parameter for classification trees. Some justification for the choice of 500 would be nice here.

Line 219 QRF method with random sampling to evaluate return levels (RLs) "QRF method with random sampling" is not a known terminology, it is one you designed for the purpose of this study. Therefore, it must be made very clear in the manuscript. For instance, you could say "in the following, we denote "QRF method with random sampling" the following methodology:..." and then describe your methodology. The description must be very clear and thorough, including every step of the calculation, to avoid misunderstanding and allow reproductibility of your results.

Also, since this method is compared with another way of estimating return levels (Figure 4) you should explain this other method of estimating RLs (simply named "QRF" in Figure 4) in this section as well.

Lines 120-122 Based on the QRF daily means and standard deviations, we calculate the corresponding annual maxima from the reproduced time series and their associated standard deviations. This isn't clear to me. I suppose "the corresponding annual maxima" are the annual maxima of average QRF predictions ? But how do you incorporate the standard deviations in the maxima ? You should write equations for this.

Lines 122-123 We assume that a Gaussian distribution describes the probability for each predicted annual maximum. It seems that this is a hypothesis that you could (and should) test for.

Line 123 10 000 sets of 30-year maxima You have Gaussian distributions for annual maxima, and you use it to draw 30-year maxima ? Why 30-year maxima ? What does this mean ?

Line 125 This yields an ensemble of randomly drawn RL curves. Why would you trust this method rather than simply using the QRF-mean daily maxima ? If one method is better from a statistical viewpoint, then there is no point in doing both (and showing both in the report). This adds confusion. I suggest you consider keeping only one, either "QRF" or "QRF random sampling". If not, this choice should be motivated.

Due to this confusions, Figure 4 appears unclear to me, while it is the most important figure of the pre-print. Maybe it is due to my lack of knowledge of QRF methods, but I doubt that this is the only reason. Anyway, this technical note should be accessible for readers unfamiliar with QRF.

(h) Figure 3 Since you show only one example, I think it would be better to show one with Halmstad as predictand, as this is the main objective of your study. This would also allow you to illustrate the points mentioned here in comment (a).

3 Technical corrections

1. I think using only "ESL" and not "ESLs" would be enough, and clearer. However this decision is yours to make.

2. Figure 1: Some of the fonts are too small to be read (the latitudes/longitudes, as well as the city names on the left panel). Either enlarge the font or supress the text.

3. Line 71: from which the annual (daily) maximum → you could remove "(daily)".

4. Lines 76-77: Conversely, long-term linear trends (i.e., sea level rise) were estimated for all time series and found to range between 0.34 and 1.47 cm per decade..
   Could you indicate all values of computed linear trends, along with the corresponding city ? Since there are only 4 stations it would not be excessively lengthy. Also, I think the way these linear trends are estimated should be explained in a bit more details. There are different ways of estimating linear trends for sea-level, corresponding to different hypothesis. In particular, for the Hornbaek station, is the linear trend computed using the whole time series (i.e. before 1900) ? Is this relevant or should the rise start later ? Does it make any difference for the estimated time-series ? Although it might not make a huge difference, it seems important to indicate this, since you are estimating long return periods and since Hornbaek is the longest time-series in your dataset, and therefore the time-series which contains a large part of the information on which your ML techniques rely.
   "It is worth noting that since we use observed tide gauge data, long-term trends, that is, climate change induced sea level rise are implicitly considered, although site-specific changes in the relative sea levels due to, e.g., human activities may introduce biases."
   Same here. These points are crucial to your study and need to be debated more.

5. Lines 83-84: The proposed approach for extending short sea level time series uses one neighbouring station as predictor (station x) of past sea level data at the station of interest (station y). The way x and y are defined is not fully clear. I assume that you are using daily maxima, with long-term linear trend removed. Please recall this here.

6. Lines 92-93: the sea level at station y is predicted from the sea level at station x. Perhaps you should make it clearer that it is the sea level at station y, time t, that is predicted from the sea level at station x, time t. See comment above for the same lines 92-93.

7. Table 3: You provide "uncertainties" from the paper by Andersson (2001), however, I cannot find an explanation of what these uncertainties are, more precisely. This would help to compare it with your "95th percentile ensemble spread".

8. Also in Table 3: I think you should be able to give uncertainties associated with the QRF-based RLs from every station, and therefore add a line of the type "uncertainties" below each line. This is a type of output available from a QRF model I believe.

9. Also in Table 3: Are the RLs computed by Anderson based only on Viken as predictor, or are the winds also used as mentioned in the text ? If this is the case, you should specify it in the table with something like "Viken + wind". If not, you should specify it in the text to avoid confusion. Also, I think a bit more description of Andersson's method would be helpful here, since you use it many times for comparison.

10. Lines 193-194:
"Here, observed values are added to the extended time series to get the longest time series possible before a GEV fit is applied."
You should be more specific. Which observations are added ? How many years/months ? How much does that strengthen your model ?

11. Lines 194-195:
"Even so, RLs are still lower than the ones displayed by Andersson (2001) when based on the ML mean outputs (fig. 4-a; Table 3)." How could you explain this systematic bias ? What does it reflect ? Also, I think you should mention that the estimated RLs in Table 3 are all in the uncertainty range of Andersson's study (2001), which is a good sign, except for the 200-year RL with Viken as predictor.

12. Line 130 RMSE RMSE is not shown → perhaps add it to the table, in [cm]. It would give an idea of the relative importance of the biases, which is not clear here: a few centimeters seems to be small, but if we don't know the amplitude of typical variations of sea-level there is no way to really know (and this information is station-dependant). OR you could show the relative biases in the Table (for instance: baises normalized by RMS sea-level decadal variations around the mean)

13. Line 131 *perc95-bias* I do not see this in Table 2 ?

14. Lines 131-133 For the annual maxima, the 95 th , 97 th , and 99 th percentiles sets, marginally higher r and lower RMSE values are found for the LR in nearly all cases, with a maximum difference of 6 cm for the RMSE and 0.10 for the r value Not shown. Also, how can that be understood together with the fact that the biases are somewhat smaller when using the QRF ? It seems counterintuitive, this should be explained.

15. Lines 133-134 Overall, RMSE values are between 10 and 40 cm, and r values are between 0.6 and 0.9 in most cases. what does that indicate ?

16. Line 135 a slight underestimation of the observed extreme values for both the LR and QRF how do we know that -30cm is "slight" ?

17. Line 138 is observed in nearly all cases add "not shown".

18. Lines 140-144 In those two cases, the QRF does not correctly reproduce the extreme range, as they are out-of-sample while the predicted values are bounded, since the ML can only reproduce in-sample events. Compared to an LR, it is clear that the inherently non-linear QRF is better able to account for the few moderate extremes that occur during the 8-year training period, whereas they are likely to be suppressed in a linearized model. To me, this is a very interesting point here. You seem to conclude that the QRF is better than the LR, since the latter smooths out the extremes, however you also point out that the QRF is not able to produce out-of-sample predictions. I would recommend a more nuanced conclusion.

19. Figure 3

    - Standard deviation is also available when performing LR (it is assumed to be always the same, this is called homoscedasticity), it is given by the RMSE, therefore you should also plot the error bars for the LR.

    - It seems like you are not using QRF but simply RF, since you estimate a mean and standard deviation, am I mistaken ?

    - I see coloured stars close to the diagonal y=x but I do not see how these could be 1st and 99th percentiles ? This part isn't clear.

    - The figure is quite fuzzy in the dense area of average sea levels between 0 and 50cm. I recommend that you do not show all error bars, perhaps only for extremes (i.e., above/below certain quantiles) as this is the main objective of your study.

20. Line 152 the model accuracy clearly decreases It does not seem "clear" to me. For instance, the best r values (0.91, LR) are found for the pair Viken-Ringhals and Ringhlas-Viken, which seem to be pretty far apart on the map. Only a more systematic study of the statistics (r, biases, etc.) with respect to the distance between the cities would reveal undoubtedly this distance-dependency (which is probably true).

21. Line 153 around 0.7 I don't see this value in Table 2 ?

22. Line 153 (9 km apart) perhaps show distances somewhere in the Table to improve readability ?

23. Line 153 r coefficients around 0.3 I don't see the value "0.3" in the table ?

24. Lines 155-156 (annual maxima or 95 th, 97 th and 99 th percentiles). add the mention "not shown"

25. Lines 156-157 imilar results are found when comparing the sea level time series for Hornbaek and Ringhals, based on Viken data, and when comparing predictions for Viken and Hornbaek sea levels based on Ringhals data. add "not shown"

26. Lines 159-160 This can probably be explained on physical grounds however, this is beyond the current technical note. There is nothing we can do with this information, I think it should not be specified, or you should say more.

27. Lines 160-161 In general, the QRF method seems to be more accurate than the LR when predicting local sea levels from stations located further away e.g., between Ringhals and Viken / Hornbaek as compared to Viken and Hornbaek Should we see this in the annual maxima ? It is not true when we look at the predictions of Ringhals based on Viken, or Hornbaek based on Viken (200-2010, 5cm stronger bias for the annual maxima using QRF). **More generally, you should always mention which numbers in the Table support your claim otherwise it can be questioned.**

28. Table 2 About the colors, is it the right choice ? You do not indicate why the sign is so important that you highlight it in colour. Red highlighting connotes "danger", should we fear an overestimation of sea-level ?

29. Table 2 You highlight in bold when QRF is better than LR in bias on annual maxima by 5cm. To be fair, you should also highlight (differently) when LR is better than QRF by 5cm, this is the case for the couple Viken-Hornbaek 2000-2010.

30. Line 183 the setup period replace by "the train and validation period"

31. Line 184-185 When analysing the error metrics over the testing period, the model based on Viken station presents the best results. What makes you say that ? The biases are smaller with the Ringhals station as predictor.

32. Line 189 (not shown However this is probably the most important piece of information !

33. Lines 193-194 Here, observed values are added to the extended time series to get the longest time series possible before a GEV fit is applied. What are these added values ? You have to be more specific for readers to know what you have done.

34. Table 3

    - Station x you could add "predictor".

    - 5th line, for the Andersson (2021) study, you write simply "Viken", but I understand that winds are also used to make this estimate ? Therefore you should perhaps write "Viken+winds" in the first column, 5th line of the Table.

35. Line 203-204 The inferred RLs are slightly higher than the RLs derived directly from observations are these observation-based RLs shown anywhere in the paper ?

36. Figure 4's caption (end) Black error bars show RLs and 95 th percentile CI calculated from Andersson (2021). → this should also be in the legend on the right.

    - also, what is "MLE" in the legend ?

    - also, all elements of the legend should be in the same box, here it seems like the upper box is for the upper panel and the lower box for the lower panel, which is not the case

37. Lines 232-233 This limitation is a known issue when applying ML-based prediction models (Tyralis et al., 2019; Hengl et al., 2018); Wrong use of "ML" : many machine learning algorithm can produce values outside of the observed range. The two cited paper are about random forests, not ML in general.

38. Line 239 but it may also confuse the interpretations at times, could you be more specific ?

39. Line 241 but this is certainly an active research area I think you should replace "active" by "promising"

40. Line 250 The best reconstructions are generally achieved for stations spatially closer maybe this would change if you allow to use time-delays in the definition of x. See comment (f) above.

41. Line 252 We tested the QRF method with random sampling Replace by "We tested another method that we named 'QRF with random sapling'."

42. Line 281 That doi seems to point to another article which is not the one you mention in the text.