Response to the reviewers

---

1. Reviewer #1

We would like to thank the reviewer for their time and effort in carefully reviewing our manuscript. We very much appreciate each of the comments we received. Please find our responses below.

1.1. General comment

First, one of the objectives of the study is "prediction of the median flood at ungauged locations" (Line 125-126). I am not quite sure if this is addressed adequately. One of the catchment descriptors is the mean annual runoff of the catchment of the SeNorge 2.0 dataset, which is based on observational data. In my opinion prediction at ungauged locations means, that there is no information about the runoff at this stations, so also no information about the mean annual runoff can be included in a prediction model. I understand that the information maybe necessary for the comparison with the RFFA_2018 model, but I think it can also be beneficial to show that the GAM model performs as good without the mean annual runoff (and leave the mean annual runoff as predictor for the RFFA_2018 model for simplicity).

Reply: The use of mean annual runoff ("Q_N") as a predictor is a modeling choice specific to Norway. The mean annual runoff variable is computed using a gridded hydrological model covering all Norway, making it accessible at both gauged and ungauged locations. The output from the gridded hydrological model is bias-corrected using observed mean annual runoff at selected locations. The reason we use mean annual runoff, instead of alternatives like mean annual precipitation or extreme precipitation, is because in Norway, we possess a more comprehensive spatial coverage of runoff observations–especially at high altitudes–compared to precipitation. The flood frequency guidelines for Norway therefore recommend using modeled runoff [Engeland et al., 2020, Sælthun et al., 1997].

Of course, the uncertainty and bias in Q_N might vary a lot. In some locations, you might use a Q_N that is close to the observed one, whereas at other locations, Q_N might be interpolated. In the first case the estimated error metrics may be a little optimistic. However, for the model-to-model comparison that is a focus of the study, the inclusion of Q_N is sufficient (and necessary) given that regional models in Norway have included Q_N for the past 20 years.

We apologize that this wasn't clear in the original version of the manuscript and will update the discussion to acknowledge this point. Since the difference between the modelled runoff and meterological data in Norway is a question of data quality, it is difficult to construct a fair comparison between models with and without Q_N. We experimented with replacing mean annual runoff with a precipitation variable and found reduced performance with the precipitation variable.

1.2. General comment

My second point concerning the prediction in ungauged locations, is the variable selection. If I understood it correctly, the variable selection is performed on the full dataset (with a cross validation scheme), and the selected variables are then used as input for the validation study (again with a cross-validation scheme). If this is correct, the variables are selected on the full dataset, not on a subset, so the prediction error is somehow biased, as the variable selection already included information about the full dataset. If my

understanding of the validation scheme is wrong, I would suggest to make this clearer in the methods section.

Reply: Thank you for this thoughtful feedback. The variables are, in fact, selected on subsets of data within the cross validation scheme. We thank the reviewer for an opportunity to clarify a key point of the pre-selection and will update the manuscript to explain how the cross-validation is used in the predictor selection / validation schemes. To briefly summarize:

We employ the same 10-fold cross-validation framework for both predictor selection and model validation. First, machine learning-based pre-selection (Section 4.1) is applied to the same 10 folds used in the model validation procedure, with results depicted in Fig. 4. Next, expert judgment interprets Fig. 4 to identify the "potential predictor set," listed in the manuscript as the seven catchment descriptors on lines 279-283. This potential predictor set undergoes formal statistical treatment (selection via shrinkage estimation) within the same 10 folds used in the model validation procedure. Notably, each fold has the potential to yield different predictor selections from the identified set. It was not clear in the original manuscript that we were using the cross validation folds in the shrinkage estimation step of the predictor selection. In our study, shrinkage estimation identified the same predictors for all 10 folds. This, combined with ambiguous language in the original manuscript, made the use of the cross-validation scheme unclear. We will add clarification to Line 289.

1.3. General comment

The manuscript is a bit too long. Some parts of the methods are in the Appendix, which is fine, but makes it hard to read at some point. Two examples: (i) The description in the Introduction between line 29-46 may be shortened, (ii) or the very detailed description of the response variable (Sect 2./2.1) can may be written more concise.

Reply: We agree with the reviewer and will make all efforts to shorten the manuscript in a revised version. We thank the reviewer for the concrete suggestions on how we might do that.

Additional minor comments

The additional minor comments provided by reviewer 1 are specific and well thought out. We will incorporate each one into a revised version of the paper. The comment pertaining to line 384 is more substantial than the others so we address it below:

1.4. Line 384

"XGBoost is assessed only on the MAE; optimal predictors for the other four error metrics are not accessible for XGBoost when the data are assumed log normal." Why is XGBoost then used as for the variable selection procedure - and the MAPE as error metric, if this will produce unreliable result? Additionally, was it considered to alter the loss function in XGBoost for better comparison in terms of the chosen error metrics? I think this should be clarified in the main manuscript.

Reply: It is common practice to evaluate predictive performance of competing forecasting methods by assessing their accuracy under a variety of error metrics. As reviewed in [Gneiting, 2011], the various error metrics we use require different point predictions from a predictive distribution to minimize the expected error of a method. However, with forecasting methods such as XGBoost that only output a single best prediction and not the full

predictive distribution, it is not possible to obtain all of the point predictions needed for these metrics in a single run. Furthermore, since changing the loss function might alter the model estimates, treating multiple runs under different loss functions as a single forecast is not appropriate.

However, XGBoost has proven to be a powerful prediction method when a large number of potential— and potentially co-linear— predictors are available. Furthermore, many prediction settings don't require uncertainty assessments, and for such settings, XGBoost has proven to be a popular and powerful method. While we believe that uncertainty assessments are highly relevant in our context, we face the former issue of having many co-linear predictors to choose from, and we find that XGBoost performs well in this aspect of our problem, specifically in selecting a small subset of important predictors. For the sake of completeness, we thus also decided to include the model comparison in the prediction setting to the extent possible, since it also outputs predictions. We see that these nuances have not been sufficiently discussed in the current version of the manuscript, and we will update it accordingly.

The reference to MAPE on line 652 was an error carried over from an earlier version of the manuscript and we thank the reviewer for catching it. The automatic stopping condition relies on MAE, not MAPE

References

[Engeland et al., 2020] Engeland, K., Glad, P., Hamududu, B. H., Li, H., Reitan, T., and Stenius, S. M. (2020). Lokal og regional flomfrekvensanalyse. Technical report, NVE.

[Gneiting, 2011] Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

[Sælthun et al., 1997] Sælthun, N. R., Tveito, O., Bønsnes, T., and Roald, L. (1997). Regional flomfrekvens-analyse for norske vassdrag. *NVE rapport*, 14:1997.