Response to the comments on the manuscript (HESSD-2019-415) "A geostatistical framework for estimating flow indices by exploiting short records and long-term spatial averages - Application to annual and monthly runoff"

T. Roksvåg, I.Steinsland and K.Engeland

November 14, 2019

This is the authors' answer to the interactive comment posted by Joris Beemster. We are very thankful for Joris Beemster contribution. His review will for sure contribute to improving the paper.

1 General comments

First, Joris Beemster comments that *"it remains unclear how catchment discharge is disaggregated to point discharge"*. This comment is similar to the comment by referee Dr. Gregor Laaha. The authors conclude that this part of the methodology should be described in more detail.

Joris Beemster next comments that there should be more discussion regarding *"the applicability of the method to other flow indices and different geographical settings"*. We like his suggestions about moving the comparison with other studies (5.4) to the discussion part. We will also add some more discussion on how we expect the methodology to perform in other areas and for other indices.

"Furthermore, it would be valuable to mention the amount of nested catchments and degree of nestedness". Currently, figure 1b gives an indication, but no reference to this figure is made in the study area description". Regarding this comment, we can add a reference to Figure 1b, and include a sentence about how many of the catchments that are nested and not nested. We can also add that only catchments where human activity has a minimal effect on the mean annual runoff are included in the analysis.

Joris Beemster lastly comments that *"it is unclear to me why records from all over Norway are used (figure 1a), but all maps in the results section only show the results for southern Norway. It seems more consistent to limit the analysis to southern Norway or to present the results for the entire country to increase transparency"*. The analysis and cross-validation is done for the whole country. However, we only show figures of southern Norway in the results section. This choice is made to make the presentation clearer and the figures as clean as possible. In Figure 1 we show the whole study area, and here it is a bit difficult to see the smaller catchments. We wanted to avoid this in the results section, and solved it by only plotting a "zoomed-in" version of Norway. Furthermore, each plot is included to illustrate a property of the model and these properties can be illustrated without including the whole study area. For example the point of Figure 7 is to show that the three methods all are able to reproduce the spatial pattern in Norway, and to show that they typically fail for the same catchments. As stated in the caption, *the methods give similar results for northern Norway*, i.e including northern Norway does not add any new information regarding the difference in performance of the three methods. The purpose of Figure 13 is to show that we are able to decrease the RMSE for some of the "problematic" catchments by including a short record of length 1. Again, this can be showed without including northern Norway, and is a choice made for making the illustrations clearer.

2 Specific comments

We will now reply to some of the specific comments by Joris Beemster. The specific comments that are not commented in this reply are mostly related to the references, choice of words and adding more explanations, and will be taken into account in a final version of the manuscript.

First, Joris Beemster writes that *"it would be interesting if the case of ungauged neighbors is also evaluated. Likely, large improvements will be seen in the PG-N, relative to the UG-N case for the new methods that are less apparent for Top-Kriging"*. This is a good suggestion. However, we think that we should prioritize making an assessment of the methods' predictive performance on mean annual runoff (for a period as a whole, not for individual years) as suggested by referee Dr. Gregor Laaha. As it is now, PG-N is included to show the difference between the three methods when having a sparser dataset with neighboring catchments that have a lot of missing data. This is informative also without adding UG-N.

Next, Joris Beemster writes that *"if you would also test the areal and centroid method for partially gauged catchments with a record length of two, the comparison with linear regression would be more straightforward"*. We chose to use a record length equal to one for our methods to emphasize that 1) our methods are able to exploit a short record of length one. Linear regression requires a record of length two or more. 2) We can provide predictions approximately equally as good or better than linear regression by using a shorter record length.

Regarding evaluation scores: We have chosen RMSE and CRPS which works well when the main interest is in comparing the predictive performance (and predictive uncertainty) between methods rather than between catchments or time scales. Furthermore, in Figure 14 we include the ANE and r^2 for the areal model for all settings. These two scores are scale independent and make it possible to compare the predictive performance on the annual scale to the predictive performance on the monthly scale.

Regarding Figures 8-10: These give some additional information compared to Table 1-3 because they show how large spread there is in the RMSE for each method and setting. The mean RMSE over all catchments can be equal for two methods (considering only Table 1-3), but the variability can be larger for one method compared to the other (which can be learned from Figures 8-10).

Regarding Figures 11-12: We chose to not include a scatter plot for PG-N here. The purpose of these scatter plots is to show the increase in the predictability of annual runoff when including a short record of length one, and that the impact on April/January is less apparent. The dataset used for PG-N is a lot sparser, and not directly comparable to the UG and PG case.

Similar to referee Dr. Gregor Laaha, Joris Beemster comments that the mean annual runoff map in section 5.5 does not have much added value. As stated in our previous reply, this section can be replaced with an assessment on the methods' predictive performance for mean annual runoff (for a longer period as a whole, not only for individual years as in the rest of the paper).