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

October 25, 2019

This is the author's response to the comments of referee Dr. Gregor Laaha 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"*. We first want to thank Dr. Gregor Laaha for his constructive suggestions and insightful comments. In this response we go through his comments on "1.Scope of the paper", "2.Geostatistical methods", "3.Evaluation method", "4.Discussion" and "5.Conclusions and Specific comments".

1 Scope of the paper

As the referee correctly state, we focus on filling gaps in annual flow index series in the article and don't provide an evaluation of the method's performance on predicting long-term average indices. Adding the latter to the final version is a good idea, and could replace Section 5.5. One suggestion is that we include an evaluation of the method's ability to estimate the mean annual runoff between 1996-2005 (for the 10 year period as a whole). For Top-Kriging, the approach has to be implemented with observation weights corresponding to the observation length as the referee suggests. This option will keep the presentation tidy and the results comparable to the other experiments that are done.

Another possible option is to design an entirely new experiment where we use the data in Figure 15a (all available data from 1981-2010), and do the evaluation for only some selected catchments and settings.

2 Geostatistical methods

About disaggregation, change of support and water-balance for the areal model:

- In our methodology the latent field (of point runoff) and the parameters (of the covariance model) are estimated simultaneously. A variogram is not explicitly estimated as in the Top-Kriging approach. Hence, we use a fully Bayesian approach as opposed to an empirical Bayesian or a frequentist approach. See Lindgren et al. (2011) for details about the inference procedure.
- We link the observed runoff (that represents catchment areas) to point runoff (averaged over the catchment areas) through the likelihood and Equation (11) for the areal model. In Equation (11), runoff is modeled as the average point runoff over a discretization of a catchment area $\mathcal{A}$. Furthermore, the likelihood and Equation (11) are used to put constraints over the gauged catchments. This means that the posterior mean runoff over the grid nodes that represent the area of a gauged catchment is constrained to the actual observed value with some (small) uncertainty through the likelihood. With one constraint for each gauged catchment, annual runoff is estimated (posterior mean) for each grid node in the discretization for the whole study area. See the illustration below for an example of how the posterior mean results can look like at point level. The posterior mean point runoff $q_j(\boldsymbol{s})$ is aggregated according to Equation (11) to produce figures like Figure 7, i.e catchment (areal) predictions. A figure

similar to this is behind all of the predictions presented in our article. See Moraga et al. (2017) for more technical details about the areal formulation.

- We should also mention that we have a quite strict prior on the uncertainty of the areal observations (page 12): This is to keep the esimated value in the gauged catchments as close as possible to the actual observed runoff (i.e we require the estimated runoff of a catchment to follow the observed runoff closely), and to avoid too much smoothing.
- Since two catchments that overlap share grid nodes, a catchment can not have a smaller posterior mean annual runoff runoff in m^3 /year than an overlapping sub-catchment for the areal model. (The posterior mean point runoff $q_j(\boldsymbol{u})$ from the illustration below is aggregated to catchment runoff for all nested catchments).

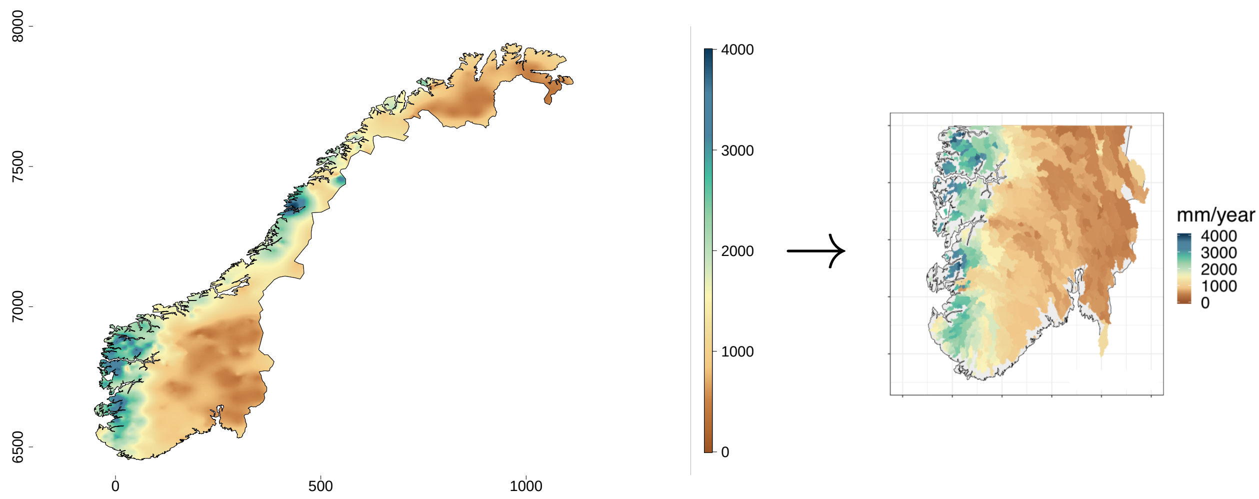

Illustration: *From posterior mean at point level $q_j(\boldsymbol{u})$ for a selected year j (left) to posterior mean at catchment level $Q_j(\mathcal{A})$ (right). If we use the areal model, we go from the left plot to the right plot by aggregating point runoff according to Equation (11) for all catchments. If we use the centroid model, we simply take the posterior mean at the catchment centroid in the left plot to produce the catchment runoff in the right plot.*

The referee writes that *"most methods appear sound, but I have concerns about the actual value of the areal model"*. The areal model has two main benefits compared to the centroid model: 1) It gives a better representation of the posterior uncertainty and 2) its ability to fulfill the water-balance and distribute the annual runoff correctly over sub-catchments. In our article, only benefit 1 is demonstrated through a real case example (Figure 7 showing the posterior standard deviation of A, and Table 1 showing the coverage of A compared to C). We did not find a clear example in our dataset where property 2 represented a large benefit over the centroid model, which is also the answer to the referee's later comment: *"It would be interesting to see how the proposed model performs in different estimation settings [...] This will enable the authors to show how well the areal model is able to incorporate the water-balance constraint in a useful way as stated in the introduction, and how far it is equivalent to Top-kriging in this respect"*.

Benefit 2 of the areal model is probably easier to demonstrate through a simple case example. One example from Voss in Norway is already available on page 15-16 in Roksvåg et al. (2019), accessible at `https://arxiv.org/pdf/1904.02519`. In Roksvåg et al. (2019) we use the same model as in this article, except that also point referenced precipitation data are used in the analysis. Here, the areal representation of nested catchments allowed us to correctly predict larger values in Catchment 3 than any of the observed values (P+A in Figure 5 in Roksvåg et al. (2019)), which was our statement in Section 3.2.6 (page 14-15) in the article under discussion. The centroid model would not be able to do this. We can search for a similar simple case example for the final manuscript. However, since we did not see any clear trends in our original dataset, we think that the presentation will be clearest if we stick to the overall global performance and use the Norwegian dataset as it is.

Further, the referee has some comments about the monthly runoff extension and choice of words. This can be rephrased as the referee proposes. The main reason for including the analysis of the monthly data is to investigate and explain the method's performance for a different "climate" and parameter set (σ_c,σ_x, ρ_c, ρ_x).

3 Evaluation method

As already mentioned, we did not find a clear trend describing when the areal method performed differently from the centroid method. We did not find a trend for when Top-Kriging performed better/worse either. This is why we focus on the *"global overall, performance"* of the method rather than *"more specific assessments"*. See for example the RMSE plot for UG in Figure 7: The three methods typically fail for the same catchments, but on average the areal and centroid method fail a bit more than Top-Kriging for the ungauged case (UG) in terms of RMSE and CRPS. As stated in the discussion, we investigated if one of the methods was better for e.g high elevated catchments, location (nested catchments) or for wetter/dryer catchments, but we did not find a correspondence.

Furthermore, we did not find a general rule for when a short record has a large impact on the final results. See Figure 13 of the RMSE: Many of the catchments already have a low RMSE when treating them as ungauged (UG). The catchments that have a large RMSE for UG, typically gets a considerably lower RMSE when adding one observation of annual runoff (PG). However, from this figure it is difficult to see what kind of catchments that generally benefit from adding a short record. In a preliminary simulation study, we saw that short records have a larger impact on locations that are not nested and far away from the other catchments in terms of the spatial range. This is as expected. For catchments that are nested and/or have many surrounding observations, the increased value of adding one observation was low for simulated data. However, the nature is more complicated and never fits perfectly to any statistical model. In our case study, we have several examples of catchments where we get a large increase in predictive performance when adding a short record, even if the target catchment is nested or has several neighboring catchments (RMSE in Figure 13).

As no prominent patterns were found explaining when one of the models (A, C or TK) or settings (UG vs. PG) were beneficial for this specific dataset, we think that the presentation of the method is clearest when we restrict the main results to the overall global performance.

The referee also writes *"It would be interesting to see the importance (magnitude) of each effect. This will be informative about whether the yearly deviation from the annual pattern is rather constant, or has a spatial structure"* and suggests that we make a table. The authors are not quite sure about what is meant here, as the spatial range of $x_j(\boldsymbol{u})$ is constant for each $j = 1, ..r$ within a cross-validation fold. In Equation (4), the $x_j(\boldsymbol{u})$'s for $j = 1, ..r$ are modeled as independent realizations of the same underlying model, i.e each year has the same underlying range ρ_x. This range is already reported in Table 4 for the different time-scales. Here, we also report the magnitudes of σ_c and σ_x which gives a lot of information about *"the importance (magnitude) of each effect"*. However, we can easily add *"additional plot of maps showing the spatial variability of the annual residual xj(u) as compared to the average spatial pattern c(u)"*, as the referee suggests, i.e a figure that shows the spatial fields $x_j(\boldsymbol{u})$ and $c(\boldsymbol{u})$ with the full spatial point pattern $x_j(\boldsymbol{u}) + c(\boldsymbol{u})$ (as in the above illustration) for one or a few selected years j. Even if the range ρ_x of $x_j(\boldsymbol{u})$ is constant for $j = 1, , .r$ within a fold, the picture produced for each j can be very different from one year to another. A figure like this gives a good visualization of the magnitude of each spatial field, and the authors agree that this will be informative for the reader.

4 Discussion

The referee has several good suggestions here for improving the discussion part. The referee e.g writes that *"The section should address how far the findings depend on the particular Norwegian setting and how far they can be generalized"*. This can indeed be discussed more. One of the reasons for including monthly data, was to investigate if the results can be generalized to other runoff regimes. We saw that for June, short records have a great value, while for January and April they have a smaller value because the spatial pattern of runoff is more year dependent due to meteorological processes like snow melting/storage that are less stable from one year to another. This will also be the case for other areas and countries *"where precipitation processes dominate that occur on a smaller space-time scale, such as convective events"*. However, using the suggested models will not affect the results negatively if this is the case, because the model adjusts σ_x relatively to σ_c: We showed that the predictions for January and April for ungauged catchments (UG, Table 1) are approximately equally as good as the predictions for partially gauged catchments (PG, Table 2).

In this context we can also add that the main author has done some quick experiments with a dataset consisting of 10 years of annual data from around 550 catchments in Austria. In a preliminary study we found that $\sigma_c > \sigma_x$ and $\rho_c < \rho_x$, i.e similar to the Norwegian annual data. This suggests that the framework could be useful in Austria too. The density of stream gauges is indeed larger in Austria than for the Norwegian dataset, but we saw in Figure 13 that even in areas with a higher gauging density, a short record can be valuable. Particularly in areas with rapid changes of runoff, and/or a prominent weather divide (like in e.g the Alps).

5 Conclusion and specific comments

The referee's suggestions will be taken into account in a final version of the manuscript.

References

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:423–498, 2011.

P. Moraga, S. M. Cramb, K. L. Mengersen, and M. Pagano. A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, 21:27 – 41, 2017.

T. Roksvåg, I. Steinsland, and K. Engeland. A knowledge based spatial model for utilising point and nested areal observations: A case study of annual runoff predictions in the Voss area. *arXiv:1904.02519*, 2019.