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 29, 2019

This is the authors' response to the Anonymous referee (referee #2). We would like to thank the referee for the review and for several useful suggestions that we can use to improve the manuscript. His/her comments have also helped us to see that our main contribution and aim of the work have become unclear. In this response, we will go through the review and provide some suggestions on how we can clarify and edit the manuscript in order to address the referee's main concerns. First, we start with a general comment.

1 General comment

The main contribution of this paper is the demonstration that a model with two spatial fields gives benefits for safe use of (very) short records. In the case of annual runoff the two spatial fields are a year specific spatial field and one spatial field that is common for all years, i.e. what we refer to as the climatic field. The combination of these spatial fields enables utilization of short records in a new way. The main goal of the article is to demonstrate this through examples, and we use different time scales (annual and monthly runoff) to show how different hydrological spatial patterns affect the predictive performance of the framework.

We introduce two versions of our two spatial field model: The areal and the centroid model. We compare these methods with the gold standard for spatial interpolation of catchment based data in hydrology; Top-Kriging. The presentation might have been clearer if we instead of comparing to Top-Kriging, compared our model to our own model with only one spatial field (only the year specific field). However, in the paper we prioritized to compare with one of the most recognized methods available for interpolation of streamflow variables.

We see that our main objective, to investigate the implications of including two spatial fields in a geostatistical model for runoff interpolation, can be made clearer in the manuscript. The article should include enough information to verify that the areal model works as we claim, and describe when our two models should be used (areal if we model mass-conserved variables, centroid if we don't care about mass conservation or if we have a point referenced variable like e.g. precipitation). However, comparing the areal and centroid model should not be the main topic of the analysis. The areal model is already documented in (Roksvåg et al., 2019) (but here for a smaller case study of annual runoff where also precipitation observations are included).

2 Concerns specific for the areal model

We now go through the referee's comments more specifically. First, the referee has concerns regarding the areal model. He/she writes that *"my main concern is with the method that the authors are most excited about, the areal method. The manuscript leaves the reader with concerns about the benefits, performance and utility of this method"*.

As stated in our reply to the other referee Dr. Gregor Laaha, 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, we demonstrate benefit 1 through a real case example: Figure 7 shows that the posterior standard deviation of the areal model is different from the posterior standard deviation of the centroid model. Here, we also see that the areal model and Top-Kriging have a similar representation of uncertainty characterized by a larger posterior uncertainty for small catchments. Table 1 showing the coverage percentages, also gives an indication of the the areal model's benefits when it comes to the modeling of uncertainty.

However, we agree that benefit 2 is not demonstrated in this article, and as the referee writes *"the second property proposed by the authors is that the areal method conserves mass. This is only demonstrated hypothetically. As a valuable claim, I think it important to document explicitly."* The reason why it is not included is that we did not find a example where the areal model represented a clear benefit compared to the centroid model when it comes to posterior mean in our cross-validation experiment. As also stated in the response to Dr.Gregor Laaha, 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. As we think that the main point of this article should be to document the methods' ability to exploit short records, the concern of the referee regarding this might be resolved by referring to Roksvåg et al. (2019) and by writing the conceptual example in Section 3.2.6 with more mathematically notation. See our suggestions in Section 5 below.

Next, the referee writes: *"Section 3.2.4. shows how runoff would be accumulated across the drainage area, but I am concerned that this would not conserve mass because, as one example, it does not account for routing. That is, summing all the grid cells, so to speak, on a day does not produce the outlet runoff on that same day"*. The areal model is not suitable for modeling daily runoff for this particular reason. It should be used for variables that are approximately mass-conserved, like the annual runoff, since it does not account for routing.

3 About the computational feasibility of the methods

The referee has some concerns about the computational feasibility of the methods, in particular the areal method. We comment these concerns in this section.

The referee writes: *"On utility, the authors acknowledge that the areal model is computationally prohibitive for any real-world application. For example, see line 11 on page 19 and section 3.3, where the author points out that the spatial discretization of the areal method means that several substantial assumptions must be made to simplify the areal method for application. While there is certainly value in presenting a hypothetical model for discussion, this leaves the reader feeling like the areal method is only a hypothesis that cannot be tested"*.

The simplifications mentioned in Section 3.3 are mainly necessary because we are dealing with a full Bayesian model with two spatial fields that need to be estimated ($c(\boldsymbol{u})$ and $x_j(\boldsymbol{u})$), both with several target locations and $x_j(\boldsymbol{u})$ for several years, in addition to having 6 model parameters, all with a prior and posterior distributions. The computational challenges here are met by using the INLA and SPDE methdology, and are used not only for the areal model, but also for the centroid model. Thus, these approximations are not something that is introduced only for the areal model. This can be emphasized in line 5 in Section 3.3, as we see that it can be read this way.

A bit more regarding the INLA methodology: It represents an approximate alternative to MCMC. The approximation is in general accurate and fast, see Rue et al. (2009) for more. INLA has become quite common to use within different fields of science, and has made "unfeasible" Bayesian models computational feasible. See for example Khan and Warner (2018); Opitz et al. (2018); Yuan et al. (2017); Guillot et al. (2014); Ingebrigtsen et al. (2015) for other papers that use SPDE and/or INLA to fit complex Bayesian models.

The areal model is indeed slower than the centroid model due to more target locations, but is not computational infeasible as shown by fitting the areal model 20 × 4 × 3 times: For 20 cross-validation folds, 4 different datasets (annual, January, June, April) and for 3 settings UG, PG and PG-N.

Furthermore, the referee writes that *"the biggest evidence of the weakness of the areal method and the centroid method, and the biggest undercut to the authors' claims of advance, is that, when it comes to application, even these authors do not use their proposed methods. See sections 4.3 and 5.5. where the authors present a new, untested method to reproduce* *annual values across Southern Norway. The reader is left interested in the hypothetical method, but surprised that it is not used."*

It is not correct that we don't use our own method to produce the annual runoff map: We use the centroid method to produce the map in Figure 15 b. This map is compared to the results obtained by a simpler reference model (Figure 15c) where each year of data is treated separately from each other. While the cross-validation represents a simplification of a real world problem, Section 5.5 is included in order to show that the model actually is computationally feasible when considering 30 years of mean annual runoff (instead of 10) and catchments with different (and realistic) record lengths.

4 Choice of evaluation set-up and the length of short records

We here comment the referee's concerns about the choice of experimental set-up/evaluation. In particular the referee criticizes the choice of having a short-record of length 1 and writes: *"Indeed, there is a long history of such procedures, but I find it surprising that authors simulate a partially gauged site as one having on a single year of annual data. This is an extreme, and possibly unrealistic, case of partial gauging that will substantially affect the performance of the methods presented. While it is difficult to work with short records (e.g. 10 years of annual data), I would represent the ungauged case with three or more values."*

As stated above, the centroid and areal model are feasible for real case examples. However, the computational complexity is large when performing a cross-validation for 2 methods, 4 time-scales, 3 settings and 20 folds, i.e. we need to fit the models 480 times for our set-up. Hence, some choices had to be made in order to make a full cross-validation possible. Our choice was to fit 10 years of runoff and include a short-record of length 1. This might not be the most realistic hydrological dataset, but we think that the cross-validation still is valuable. We can look at it as mainly an experiment performed for providing an understanding of how the suggested model works. For example, the cross-validation shows how the increase in predictive performance when adding a short-record is related to the parameter values σ_c, σ_x, ρ_x and ρ_c (e.g. in Figure 11 and Figure 12). We are able to show how the posterior uncertainty is distributed (Figure 6, middle plot) for the methods, and how this uncertainty is affected by adding a short-record of length 1 (Figure 13).

Furthermore, it is not that unrealistic to have a short record of length 1. Considering the Norwegian data showed in Figure 15a, five of these catchments have only 1 annual observation, five catchments have 2 annual observations and 10 catchments have only 3 annual observations between 1981 and 2010. Our cross-validation shows how important a small bit of information like this can be if the weather patterns (thus also the model parameters) are similar to the Norwegian case ($\sigma_c >> \sigma_x$ and $\rho_c < \rho_x$).

Here, it is also important to note that we provide a model where it is relatively risk-free to include very short record lengths in our framework, which is also one of our main contributions. The model itself figures out if the study area is driven by a stable hydrological pattern that repeats itself every year ($\sigma_c > \sigma_x$) or more local year-dependent effects ($\sigma_c < \sigma_x$). If the latter is the case, the short record will not influence the results particularly. It will only have an influence for the particular year for which we have data. The monthly predictions demonstrate this point: In January the spatial patterns of runoff are not stable in Norway, and including a short record of length 1 (PG in Table 2) does not affect the model negatively compared to the UG case (Table 1). If, on the other hand, $\sigma_c > \sigma_x$, the suggested model uses information both in time and space, and a short record will on average influence the predictive performance positively.

This leads us to another point stated by the referee: *"I suggest dropping the sections on monthly analysis. The simulation of monthly streamflow tends to imply that one is producing monthly sequences line Jan-Feb-Mar, but this work is looking at Jan-Jan-Jan (for example). This is akin to only predicting a new statistic of streamflow and is not a novel advance of the method. While it could be expanded to provide a more robust analysis, removing it might help streamline the manuscript".* The monthly predictions are mainly included to illustrate the method for a different hydrological regime and a different set of parameters. The monthly predictions can illustrate that the model itself adjusts the two spatial effects $x_j(\boldsymbol{u})$ and $c(\boldsymbol{u})$ relatively to each other, and that the impact of including a short record when $\sigma_x > \sigma_c$ is small (however, it does not affect the model negatively either). We suggest rephrasing the monthly predictions part to emphasize why these can be informative about what we can expect from the methods for other parts of the world where other precipitation types dominate.

Furthermore, the referee writes: *"Given that the use of one point for partial gauges and two points for regression (line 12, page 24), it would seem wise to use a consistent number of points to represent partial gauging."* There is already an answer to this in our previous reply to Joris Beemster: 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.
Again; it is a main contribution to provide a model where also very short records can and should be included.

5 Suggestions to address the main concerns

In order to address the main concerns of the referee, the authors' have the following suggestions for clarifying and improve the manuscript:

- Keep the cross-validation set-up as it is and consider it as a demonstration of the model properties (and parameter values).
- Keep the monthly predictions, but rephrase this part and emphasize that these predictions are included to illustrate the model performance for a different set of parameters, i.e. these could represent the performance for a other part of the world with a different hydrological regime. The monthly predictions show that it is safe to include short records even if the annual patterns are more unstable than the Norwegian annual patterns since the results are not affected negatively for regions where $\sigma_x > \sigma_c$.
- Edit the conceptual example in 3.2.6 in order to explain the areal model in more detail. Here, we can write out parts of the model more mathematically to show 1) how we can predict larger values than any of the observed values and 2) how we go from areal observations to point observations and the other way around. Furthermore, we can refer to Roksvåg et al. (2019) in order to document a case for which the areal model makes a difference compared to the centroid model when it comes to runoff modeling.
- Include a more realistic case example e.g. based on the data in Figure 15a, where we predict the mean annual runoff for a 30 year period for some selected fully gauged catchments (which we first treat as ungauged for the evaluation, then partially gauged with different record lengths). In this example, we can use short records from neighbouring catchments as they are (the record lengths are 1-30 years, with average 15 years), and compare the performance of the areal, centroid and TK methods for predictions of mean annual runoff (for the 30 years period as a whole). By this we aim to show that the suggested models are feasible also for a realistic case, and demonstrate the differences between the methods for a bigger dataset. This experiment can replace Section 5.5, and will also provide an assessment of the methods' performance on mean annual runoff as requested by referee Dr.Gregor Laaha.

6 Minor comments

Finally, we comment two of the referee's minor comments.

"Page 11, line 2: Why would we expect the long-term spatial average runoff (c(u)) to have a zero mean". The effect $c(\boldsymbol{u})$ has indeed zero mean. The long-term spatial average runoff, however, has mean equal to β_c, i.e. we have a distinct component/parameter to model the mean.

Page 11, line 4: Why would expect a sequence of annual values to be independent? Is this true for monthly values? The sequence of annual/monthly values are not modeled as independent. They are dependent through components $c(\boldsymbol{u})$ and β_c that are common for all years. In this line, we talk about the year-specific spatial fields $x_j(u)$. These are assumed to be independent realizations, or replicates for $j = 1, ..., r$. However, the models are stationary in time. This can be supported by looking at time series of mean annual runoff: We don't see any e.g. increasing/decreasing trend or change in spatial pattern over time. Stationarity in time also makes sense when modeling monthly time series as $Jan - Jan - Jan$, and not $Jan - March - Feb$.

The other comments will be taken into account.

References

G. Guillot, R. Vitalis, A. le Rouzic, and M. Gautier. Detecting correlation between allele frequencies and environmental variables as a signature of selection. a fast computational approach for genome-wide studies. *Spatial Statistics*, 8:145 – 155, 2014. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2013.08.001.

R. Ingebrigtsen, F. Lindgren, I. Steinsland, and S. Martino. Estimation of a non-stationary model for annual precipitation in southern Norway using replicates of the spatial field. *Spatial Statistics*, 14:338 – 364, 2015.

D. Khan and M. Warner. A bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with r-inla. *Journal of data science: JDS*, 18:147–182, 01 2018.

T. Opitz, R. Huser, H. Bakka, and H. Rue. Inla goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes*, 21, 02 2018. doi: 10.1007/s10687-018-0324-x.

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.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:319–392, 2009.

Y. Yuan, F. Bachl, F. Lindgren, D. Borchers, J. Illian, S. Buckland, H. Rue, and T. Gerrodette. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *The Annals of Applied Statistics*, 11:2270–2297, 12 2017. doi: 10.1214/17-AOAS1078.