Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-415-RC2, 2019

Hydrology and Earth System Sciences
Discussions

Interactive comment on "A geostatistical framework for estimating flow indices by exploiting short records and long-term spatial averages – Application to annual and monthly runoff" *by* Thea Roksvåg et al.

Anonymous Referee #2

Received and published: 15 November 2019

The authors have presented a Bayesian geostatistical approach to filling in gaps in the record of annual streamflow. Two variants of the method are presented: One that attempts to conserve mass using areal weighting and a second that uses the more traditional centroid-referenced approach. Bayesian methods are introduced to allow for parameter uncertainty in the underlying processes. The manuscript is well-prepared, and the methods are clearly articulated. The impact of this work could be deepened by providing a more thorough exploration of the purported properties of these estimators and further quantifying the variability of performance.

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.

On benefits of the areal method, the authors make several claims as to the superiority of the areal method (with some of these benefits extending to the centroid method, generally), but rarely is evidence provided to document these benefits. Around line 13 of page 14, the authors suggest that the areal and centroid methods are uniquely designed to take advantage of short and long-term records. This particular benefit is true of many kriging methods, as shown by various authors' suggestions of weight top-kriging and kriging estimates differently (see, e.g., Skoien, 2006; Farmer, 2016). In both reports, the authors show how building different variograms can increase the importance of longer records to build regionalize geostatistical approaches. This pervious work does not invalidate what is presented here, as this manuscript includes some unique Bayesian work, but does demonstrate that further evidence of this claim can be provided through experimentation. 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. 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 runoff at the outlet cell is rarely a product of the contemporary runoff at a cell at the top of the drainage area. I'd like to see some further evidence of how this mass conservation is validated, especially in a sparse network.

On performance, the results of Tables 1, 2 and 3 do not convince the reader that the areal method is a meaningful improvement over the previous methods. Here, we consider only averaged performance across all basins. Even in this case, the top-kriging is superior in the ungauged cases, while linear regression and the centroid method are superior in the annual cases of partially gauged and partially gauged networks,

respectively. There is some benefit in the monthly cases. As a first order, I think it inappropriate to compare means and make definitive statements. As these RMSE and SRPS means are simplifications of data that is available, I strongly advise the use of significance testing to understand if the differences between these methods are meaningful. Given the variability Figures 8, 9 and 10, the differences may not be significant.

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 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.

In addition to this main concern, I will now move on to some other major concerns. Addressing these will, I hope, improve the manuscript.

I find the authors' simulation of short records somewhat concerning. At the bottom of page 1, the authors discuss the PUB initiative and its relevance to short records. First, I think it important to explicitly state that PUB is taken to apply to any ungauged point in space and time – that is, it included the completely ungauged and partially gauged cases. Line 23 claims that a few years of data could be useful for estimation. 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 (page 17, line

24). 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. Indeed, on line 7 of page 18, linear regression is performed with only two data points. This is upsettingly problematic as linear regression is meaningless for two points – it's just a line connecting the points. A minimum of three points would be required for any meaningful regression. (Or, are the regression built across the entire region simultaneously resulting in a single regression for all sites? Even that is suspect.) 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. (An additional analysis that may be beyond the scope of this work could consider the sensitivity of these methods to partial record length.)

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.

I also suggest dropping the simulation of the mean annual runoff map for southern Norway. This provides no additional methodological advance and substantially undercuts this work. In tables 1, 2 and 3, the annual values are shown to be best reproduced by various methods (TK, LR and Centroid), none of which are used in this application. The narrative of the manuscript might be improved by removing this section.

Finally, I'd love to seem some additional analysis on the regional variability of performance of these methods. What seems to drive the varying levels of performance across Norway?

Some more minor comments:

Page 4, line 33: This figure does not show that runoff is lowest or highest at any single site, it only shows the relative distributions. Please correct this statement.

Page 6: This discussion focusses on whether or note the lines look parallel. This is highly subjective and should be quantified in some way. For example, if I changed the vertical axis to run from 0 to 100,000, all the lines would “look parallel”. Please provide some quantification of correlation.

Page 11, line 2: Why would we expect the long-term spatial average runoff (c(u)) to have a zero mean?

Page 11, line 4: Why would expect a sequence of annual values to be independent? Is this true for monthly values?

Page 22: The coloring of figures like Figure 7 make it difficult to see the variability of performance. The results appear highly skewed, resulting in almost all RMSEs being brown; an alternative scale might distinguish performance better.

Throughout: The numbering of figures and tables is inconsistent. The figures and tables should be numbered according to the order of presentation in the prose.

Finally, thanks for a great read. I look forward towards revision and future discussion. Great work!

REFERENCES:

Farmer, W. H.: Ordinary kriging as a tool to estimate historical daily streamflow records, Hydrol. Earth Syst. Sci., 20, 2721–2735, https://doi.org/10.5194/hess-20-2721-2016, 2016.

J. O. Skøien, R. Merz, G. Blöschl. Top-kriging - geostatistics on stream networks. Hydrology and Earth System Sciences Discussions, European Geosciences Union, 2006, 10 (2), pp.277-287. ffhal-00304844f.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-415, 2019.