Review of manuscript: *"A geostatistical framework for estimating flow indices by exploiting short records and long-term spatial averages – Application to annual and monthly runoff"* by Roksvåg et al. (2019a)

Joris Beemster[1]

[1]Wageningen University and Research, Wageningen, The Netherlands

This review was prepared as part of graduate program course work at Wageningen University, and has been produced under supervision of dr Ryan Teuling. The review has been posted because of its good quality, and likely usefulness to the authors and editor. This review was not solicited by the journal.

Recommendation: Accept manuscript after minor revisions.

Introduction

The manuscript presents a new framework for the spatial interpolation of hydrological data, that includes a mix of fully and partially gauged catchments. It tackles the problem of combining short and long discharge records, when predicting streamflow indices. A key aspect of the novel method is that it models several years of runoff simultaneously using two Gaussian random fields (GRFs). One that is common for all years, representing the climatology and catchment characteristics, and one that is year specific, representing the deviation from the climatology. Based on this framework, the authors propose two models. One that satisfies the catchment mass balance (the areal model) and one that does not (the centroid model).

The models are tested on a dataset of 195 fully gauged Norwegian catchments and the results are compared to those produced by Top-Kriging (Skøien et al., 2006) and linear regression. The performance is evaluated by means of three tests, all based on cross-validation. In the first test, the models predict the runoff of ungauged catchments. During the second test, the ability to predict the runoff of a partially gauged catchment is evaluated. Lastly, the models predict the runoff of an ungauged catchment, of which the three nearest neighbors are partially gauged.

Top-Kriging outperforms the newly proposed method in the ungauged case. In the case of partially gauged catchments, the areal model generally performed best, indicating that the method is capable of exploiting short records, which seems to be the main advantage over existing methods. The presented method also outperforms Top-Kriging, when interpolating ungauged catchments with partially gauged neighbors, but differences are smaller.

As the presented methodology is capable of transferring information contained in short discharge records across different years, it significantly increases the value of large amounts of short records. Furthermore, it provides an easily interpretive way to indicate the relative importance of the climatic signal and catchment characteristics and the interannual variation. The novel method fits the scope of *Hydrology and Earth System*

Sciences well and the paper is generally understandable. The main weaknesses of the paper are the minimal explanation of mass conservation in the areal model, the missing discussion of the wider applicability of the method and the limited study area description. Therefor, I recommend to accept the manuscript after some minor revisions. My major, minor and textual suggestions are outlined below.

General Comments

The first main shortcoming of the manuscript is that it remains unclear how catchment discharge is disaggregated to point discharge. I assume that in catchments that do not contain any subcatchments, it is defined as the average point discharge and in catchments with nested gauged nested catchments, the nodes in the ungauged part are chosen such that the water balance is preserved. It is a simple and clever way to ensure mass conservation, but it should be explained more clearly in the methodology. If, in the results section, examples would be given of a system of nested catchments instead of only providing the global picture, the way nested catchments are incorporated into the method would become more transparent and advantages and disadvantages could be highlighted more clearly. An alternative would be to mention more clearly that the author has also applied the method to a more limited dataset, of only 5 catchments in the Voss area (Roksvåg et al., 2019b).

My second concern relates to the discussion. Although it thoroughly reflects on the strengths and weaknesses of the novel method applied to annual and monthly discharge of southern Norway, it does not address the applicability of the method to other flow indices and different geographical settings. A nice way to bridge between the two aspects would be to move the comparison with other studies (5.4) to the discussion. (More generally, I would recommend to reconsider which sections/sentences should be in the result and which in the discussion). This should be followed by a discussion whether or not the method is expected to perform particularly well for the Norwegian setting and how the spatial interpolation of annual and monthly average runoff compares to the interpolation of other flow indices.

Related to my previous comment, to give the reader a sense of how well this method would work in different geographical settings as well as explain the results, it would be helpful to further elaborate on the study area. There are over 300 large dams in Norway (Icold, 1998). The presence of large reservoirs in a catchment will undoubtedly have an impact on the performance of the method. In my opinion, it would strengthen the description of the study area, if an assessment of the degree to which the catchments are influenced by human activity was made. 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. Adding a couple of sentences mentioning the amount of nested catchments, as well as, a reference to figure 1b, will also improve the study area description. Lastly, 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.

Specific technical comments:

- p1, line 17: Please provide more than one reference if you state that "Average annual flow is *often* used..."
- p3, lines 19-20: "A similar model has already shown promising results". Please mention how it differs.
- p9, lines 29-30: "However, it has been shown that Top-Kriging also performs well for variables that are not mass conserved, like e.g the specific 100-year flood". Please support this statement by one or more citations. Skøien et al. (2006) indeed show that Top-Kriging works well for the 100-year flood. Top-Kriging was also successfully applied to low flow data (the daily discharge that is exceeded 95% of the time) (Laaha et al., 2014) and to streamflow temperatures of Austrian rivers. (Laaha et al., 2013).

- p10, lines 30-31: "Likewise is c(u) a spatial effect that models the long-term spatial average of runoff, or the spatial variability caused by climatic conditions in Norway". The spatial variability is not only caused by climatic conditions, but also by the catchment characteristics.

- p12, lines 10-12: Petersen-Øverleir (2004) indeed shows that heteroscedasticity is a widespread problem of Norwegian gauging stations. However, he also shows that differences between gauging stations are large and that there is at least one example where the uncertainty decreases with increasing runoff. Please motivate why this value for the scaling factor was chosen.

- p12, lines 22-23: Please provide a source for the following statement: "This corresponds well to what we know about the measurement uncertainty for runoff in the study area."

- Inference (3.3): In this section several simplifications are mentioned aimed at reducing the computational complexity. Could you please comment on the effect these simplifications have on the expected outcome?

- p17, line 31: 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.

- Evaluation scores (4.2): The performance of the model is mainly evaluated in terms of RMSE and CRPS, two evaluation scores that are scale-dependent. In my opinion, adding a scale independent evaluation score, such as the Nash-Sutcliffe or the Kling-Gupta efficiency, would make the comparison between averaged annual and monthly runoff more straightforward. Furthermore, this would enable the evaluation in terms of the correlation, the conditional bias and the unconditional bias (Gupta et al., 2009).

- Figures 8-10: The units of the y-axis are not mentioned. Please add them. In my opinion these figures could be left out of the paper, because they are made redundant by table 1-3 and figure 7.

- p24, line 11-13: "However, recall that a short-record of length 2 from the target catchment is needed in order to use this method, while our areal model performs approximately equally well with a short-record of length 1 (and observations from other neighboring catchments)." 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.

- Figures 11-12: It would be useful to add the scatter plots for the **PG-N** case, for completeness.

- p27 line 18-20: "The results from UG and PG-N in Table 1 and 3, indicate that exploiting long-term spatial averages in the interpolation scheme as in our methods, can be just as important as finding a process-based way of determining the Kriging weights which is the idea behind Top-Kriging." This sentence belongs to the discussion section.

- Mean annual runoff map for southern Norway (5.5). In my opinion, this section does not have much added value. Instead, as mentioned before, it would be interesting to highlight the results of specific catchments. Also, if you choose to keep this section, it would benefit from adding a map produced using the Top-Kriging method, for completeness.

Minor textual suggestions

- p1, line 17: "a predictor" instead of "an predictor"

- p9, line 30: "e.g." instead of "e.g"

- p11, lines 19-20: Consider to change "This alternative does not require preservation of the mass balance" to "This alternative does not assure preservation of the mass balance".

- p14, line 1: "is an indicator" instead of "is a indicator"

- p27 line 18: Please change "... annual scale in Table 3" to "annual scale (Table 3)"

References

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, 377(1-2):80–91.

Icold (1998). World register of dams.

Laaha, G., Skøien, J., and Blöschl, G. (2014). Spatial prediction on river networks: comparison of top-kriging with regional regression. *Hydrological Processes*, 28(2):315–324.

Laaha, G., Skøien, J. O., Nobilis, F., and Blöschl, G. (2013). Spatial prediction of stream temperatures using top-kriging with an external drift. *Environmental Modeling & Assessment*, 18(6):671–683.

Petersen-Øverleir, A. (2004). Accounting for heteroscedasticity in rating curve estimates. *Journal of Hydrology*, 292(1-4):173–181.

Roksvåg, T., Steinsland, I., and Engeland, K. (2019a). A geostatistical framework for estimating flow indices by exploiting short records and long-term spatial averages – application to annual and monthly runoff. *Hydrology and Earth System Sciences Discussions*, 2019:1–35.

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

Skøien, J. O., Merz, R., and Blöschl, G. (2006). Top-kriging - geostatistics on stream networks. *Hydrology and Earth System Sciences*, 10(2):277–287.