Interactive comment on “Topological and canonical kriging for design-flood prediction in ungauged catchments: an improvement over a traditional regional regression approach?” by S. A. Archfield et al.

E. Pebesma (Referee) – *referee comments in italics*

I enjoyed reading this manuscript that quantitatively compares three methods for predicting flood quantiles for ungauged catchments. The paper is well written, but could be improved in a number of areas, in particular by providing some more detail (or reference to where details can be found) of the methods used.

The authors claim that “To provide a more equally-weighted assessment of the fits between the empirical and predicted quantiles across all flood quantiles, the NSE value computed from the natural logarithms of the empirical and predicted flood quantiles (NSE-L)”, it was not clear to me in which respect this assessment would be more equally weighted. What is the operational meaning, or advantage, to work with logarithms of quantiles? Obviously, as NSE is a linear measure, everything will change after a non-linear transform of the data. Why is this an improvement?

Response:
The reviewer Edzer Pebesma raises a good point here, as this was not properly expressed in the original manuscript. When using the NSE index for comparing models, the emphasis is on the magnitude of the deviations between predictions and observations. As floods are to a large degree linearly dependent on the catchment size, the NSE index will then also emphasize on the fit to the largest catchments. For example, the results show that CK performs particularly poorly for the largest catchment of the study area, that is site 02352500, which confirms what other studies revealed (see e.g. Castiglioni et al., 2011). This single catchment is significantly larger than all other catchments in the study area (see Fig. 2) and the NSE index is significantly impacted by this single prediction.Therefore the comparison of NSE-L across methods dampens out the effect of this very prediction and provides an additional and more fair quantification of prediction performance for the whole set of catchment in the study area. This consideration is not clearly expressed in the manuscript and will be incorporated in the revised version. We also point out that we use several other goodness-of-fit metrics to assess the differences in methods and the conclusions drawn from the examination of the NSE-L values are consistent with the comparison of the other goodness-of-fit metrics.

In terms of prediction accuracy, “TK consistently outperformed GLS and CK”. This is not surprising, as TK is the only technique that exploits spatial correlation in the residuals for spatial prediction. The authors then continue to try to combine TK’s and CK’s strenghts, but I wondered why they did not combine TK with GLS’s strenghts. This method, essentially universal kriging (sometimes called external drift kriging or regression kriging) would further improve TK by including the physiographical variables in a regression setting, which may be easier to comprehend than working in the space of the first two canonical variables. If multiple collinearity would play a role, ridge regression alternatives might be considered.

Response:
Good point! We perfectly agree with EdzerPebesma that it would be very interesting to combine TK and GLS and we appreciate the suggestion. Nevertheless, we believe that this application is out of the scope of the present study and we will definitely try this approach out in future analyses. We disagree in part that the main outcome is not surprising, for a series of reasons:

- Castiglioni et al. (2011) showed that CK slightly outperformed TK when predicting low flows for a different study area

- CK will indirectly take into account some effects of spatial correlation through catchment descriptors such as lat. and long. and, to some extent MAP (mean annual precip.)
- The possible advantages associated with a combination of CK and TK were never investigated before, although Castiglioni et al. (2011) indicated the potential for such a combination.

Also, GLS is used as the benchmark approach in this study that focuses on the application of kriging techniques to the PUB problem, as GLS is the current method for flood regionalization. As a consequence of the very results of this study, coupling TK and GLS becomes an interesting research idea, which we will try to pursue. These considerations will be included in the discussion section of the revised manuscript.

Both TK and CK involve a number of details in the procedure that raise questions from a data analysis perspective.

1. The description of TK claims that the point variogram can be derived from area support (catchment summary) data. For me it is still an open question whether this is the case, and if yes, how; a reference to the paper that explains the details how the nugget effect of the point variogram is estimated would be helpful. (More in general, it might be worth thinking about what the point variogram exactly represents in this context, whether observation at a point support is possible at all, and how measurement error at the area support is modelled, and dealt with).

Response:
The nugget effect is evaluated as proposed in Skøien et al. (2012), which we cite in the text. However, we can definitely add additional text in Section 3.2.1 better detailing the calibration procedure.

2. For CK, the authors decide to work with two canonical directions. Why two? How much of the variability in the dependent do these two directions explain?Why not three or four?

Response:
Thank you for the suggestion. We will add in the revised manuscript an indication of the total variability of the original variables that can be explained by the first two canonical variables generated by CCA. We did not use the remaining canonical variables (i.e., 3^{rd}, 4^{th}, etc.) because the additional information that they can provide is not significant, and also they can be shown to be highly correlated to each other. Nevertheless, considering further canonical variables is not a limiting factor, as 3D kriging could also be used if deemed necessary. We will include a comment on this in the revised manuscript.

Then: how are the scores in this new space normalized: do they have unit variance, or variance proportional to the explanatory power of a dimension?

Response:
The scores have zero mean and variance proportional to the explanatory power. We will include this information in the revised manuscript.

Is it reasonable to assume isotropy (assuming the authors did this) in this new space?

Response:
We tested the hypothesis of possible anisotropy in the variogram, but this hypothesis did not improve the performance in cross-validation. We therefore adopted the assumption of isotropy. We can better detail this in the revised manuscript.

Are the physiographical properties of ungauged catchments averaged before a point CK is carried out for this average, or are CK predictions averaged over the physiographical space covered by the catchment (and does this matter)?

Response:
We had some difficulty interpreting this comment and we realize that we need to add more description of the computation of the physiographic properties and how they were used in the CK routine. We interpreted this comment as inquiring whether the physiographic properties of the ungauged basins were computed over the entire area of the contributing catchment. The physiographic properties (Table 1) were computed over the entire area of the contributing catchment to the basin and then used in the CK procedures. This is not clear from the manuscript and not clear in Table 1. We will add text to the manuscript in Section 2 to make this point as well as add this to the header for Table 1.

3. It was not clear to me what the authors meant by "distance in terms of residuals" (Fig 7). If they mean difference between residuals (absolute?), then I could not follow the conclusions drawn from this figure.

Response:
Your point is well point taken. We acknowledge that the description in the body of the manuscript is not clear and not detailed enough. Distance in terms of residuals means that two catchment have similar residuals and we want to assess whether these catchment are also close to each other in the geographical space (for coupling CK with TK) or in the catchment descriptors space (for modeling TK prediction residuals with CK). We will improve the clarity of the description with examples in the revised manuscript.

I am not sure whether the authors have in mind to publish their data and scripts in a way that allows us to reproduce the results published here, but if this is not the case, the authors may want to reveal some more detail, e.g. about how (and for CK in which space) variogram models were chosen, and fitted to sample variograms. Also, when combining methods, it would be useful to describe the statistical model adopted, rather than the algorithmical procedure followed. This would make it more transparent which assumptions were made at what point.

Response:
We will mention in the paper that we implemented CK using a publicly available software written in Matlab called "EasyKrig3.0"; also, we will explain that preliminary analyses showed that the Gaussian theoretical model resulted to provide the most stable and accurate representation of the empirical variograms among all the models we tested (i.e., linear, exponential, spherical, exponential-bessel, gaussian-bessel, etc.), which are available in EasyKrig.