Editor Decision

HESS-2019-342 **A predictive model for spatio-temporal variability in stream water quality**

Christian Stamm, 22.10.2019

Dear Dr. Guo,

Thank you for responding to the four reviews of your manuscript. I appreciate the serious manner of how you have provided answers to comments and suggest that you revise the manuscript accordingly.

Nevertheless, a few aspect deserve more attention than what you have proposed. I list them below and recommend that you pay them due attention during the revision of the manuscript as well.

**Transformation of the data:**

Several reviewer comments questioned aspects of how using the transformed concentration values (R2: comment 12, R3: comment 2.1, 32, Rev. 3, comment 3.2). Your arguments to avoid comparing observations and simulations in the original space is not fully convincing: you argue that it was best to evaluate model performance in the transformed space because (e.g., response to Rev. 2, comments 2.1, 43) it was most informative and because absolute errors were less important in practice.

First, you don't provide an argument WHY it should be most informative in the transformed space. Actually, inspection of Fig. S13, reveals obvious model biases (even for the site-specific average concentrations, if I interpret the figure caption correctly). A careful look at Fig. 2 and 3 show similar deficiencies (e.g. systematic underestimation of high concentrations for TSS, TP, FRP, and NOx. However, these deviations are much less conspicuous than in the transformed space. This holds especially true because some of the chosen transformations are very non-linear making it very difficult to have a sense for the actual meaning of the transformed values. Additionally, inspection of Fig. S13 for EC suggests that there might be two populations of catchments: one population is very well represented by the model (close to the 1:1 line), while the second is definitely off. This can hardly be seen with the transformed data. Do the catchment being off share some commonality?

Second, the relevance of absolute errors is probably very context-specific. In some situations, practitioners do care about high concentrations and model uncertainty was important to them.

It is important to note that a systematic model deficiency (e.g., under or overestimation in a certain concentration range) is not alleviated by transforming the data. However, it allows for better fulfilling distributional assumption for making statistical inference. Therefore, the transformations make sense. However, to proper and transparently present the model performance and the effects of transformations, more information needs to be provided (as also suggested by the reviewers):

- Provide information on how you have determined the optimal log-sinh and Box-Cox parameters (L. 161, 164). What was the optimality criteria and how did you assess optimality (manual calibration, visual inspection of quantile plots etc.)?
- Provide information on the $\lambda$ distribution per site and constituent in Tab. S4. You may also consider to plot the respective distributions in the SI.
- Complement Fig. 2 – 5 with the regression lines between observations and simulations and provide the slope estimates (including uncertainty).
- Clarify whether Fig. 3 and Fig. S13 correspond to the same data.
- Include one figure in the main text comparing observations and model results in back-transformed form. This could be Fig. S13 or a time series that you have mentioned several times (e.g., response 2.1 to Rev. 4).

**Further editor comments:**

- L. 27: You focus here on improving the model fit for low concentrations. However, Fig. 2 and 4 suggest that the model is deficient in the low and the high concentration ranges. These systematic deviations should be addressed. If my interpretation was wrong, please provide a convincing argument why to put emphasis on the low concentrations. The argument mentioned above about the practical relevance that was less for high concentrations is not convincing. This very much depends on the actual context and some practitioners may be much more interested in high concentrations. Note that L. 154 – 155 would support this view as well.
- The data presented in the main text (e.g., Fig. 3 – 5) refer to site-specific mean concentrations across space. Of course, Fig. 4 and 5 represent such mean concentrations for different periods. But there is no information on how well temporal dynamics are captured at shorter time scales. Strengthening this temporal aspect as you mention several times is important.
- In this context, I am not fully convinced of your argument not to discuss in some more details how the model simulates the drought effects (see Fig. R3). If you consider the results solid in Fig. 4 and 5 enough to be presented in the manuscript you have also to demonstrate what makes the difference in the parameters for different periods. This is simply reporting your findings. It is subsequently fair enough to critically mention that an over-interpretation isn't warranted because of model deficiencies.

Sincerely

Christian Stamm

Editor HESS