Dear HESS Editor Jim Freer,

Thank you again for your efforts with our manuscript. We have now addressed the points you raised and included most of the comments into the text. Below we provide a point-by-point reply to each of the comments as well as an explanation on how we included them into the text (blue text).

Additionally, we have identified and fixed other minor issues in the manuscript such as a missing affiliation for one of the co-authors, a broken cross-reference, or the missing "Acknowledgements" section.

With the aforementioned modifications we hope that this contribution meets the quality requirements to be published in Hydrology and Earth System Sciences.

Kind regards,

Marc Girons Lopez, Louise Crochemore and Ilias Pechlivanidis

Authors’ response to HESS Editor Jim Freer

Dear Authors,

Thank you for your changes to the manuscript and responses to 3 comprehensive assessments by the reviewers, all who were positive to your paper. I have no problem with the fact that these minor corrections have been well tackled and you have developed considerable changes to the manuscript, your due diligence is appreciated.

We would like to thank the editor for the positive feedback.

One point though, that I was surprised did not come out in the reviews, but I would like personally to tackle you on. Namely that there seems a general disconnect in the paper that I think needs to be better discussed. The issue, as I see it, is as follows (and I know some of these matters you could note for a number of other hydrological modelling papers):

The first two points raised by the editor are very interesting. Yet, to our knowledge, there is no study that robustly addresses them. Hence, we are limited on only providing our individual views on the points below, supported by a set of references.

1. KGE appears to be your benchmark for the basis of the forecast model. Clearly KGE has certain properties of evaluation to streamflow. You do not anywhere in the paper state why models based on maximizing KGE are valuable for your forecasting objectives - in fact this whole matter (objectives over experimental design) is hardly discussed anywhere. A model with different calibration statistics will clearly be different and thus be different as the underlying vehicle for the forecast analysis. Please can this be discussed in a meaningful way, so why these choices?

   This is a very good point. We start by saying that the S-HYPE model setup is the one used operationally by SMHI and not a version tuned to a specific model objective and/or set of user needs. It is also important to note that the model developments in the S-HYPE hydrological setup have been continuous since its first implementation in 2008. This operational setup has been used in early warning services, hydropower decision-making, water resources management etc. and hence its parameterization is not towards a (set of) characteristic(s) of the hydrograph (e.g. high flows for flood forecasting).

   Additionally, S-HYPE parameters have not been exclusively optimized using the KGE metric. The various model upgrades included model evaluations (calibration and verification periods) using, among other metrics, the NSE, KGE, relative bias, and multi-objective combinations between NSE and biases, whilst quite extensive effort was given into visual evaluation and manual fine-tuning of the model parameters. Here, we decided to select the KGE as a metric to communicate model performance, since after the Gupta et al. (2009) article, KGE has been considered as a benchmark in hydrological modelling. For S-HYPE the median KGE value for more than 530 stations is 0.79, which indicates a high model adequacy (to our knowledge, there is no other national hydrological setup with that high performance over that many stations). More specifically, this shows that different properties in the hydrograph (timing, volume and variability) are well represented. The stations that show poor KGE values (<0.2; again, this threshold is subjective and driven by our own experience), are usually subject to dam/reservoir regulations, for which characteristics such as timing and variability are almost impossible to capture.

We addressed this comment by adding these considerations to Section 2.2 of the manuscript.

2. Then we move to the statistics on the output where you have table 2 showing a whole bunch of hydrological signatures used to explore a 15 dimensional space of forecast skill. Again there is a big disconnect here to any intelligent discussion of the objectives of the forecasts and why these signatures, and these signatures alone, have value. Indeed if at all why they should be considered (as it seems they are) equally weighted in the forecast assessments (do you even have an appreciation of their individual sensitivity to how good or bad the forecasts are?). Again I really think it should be explained to the reader why this table and why these metrics and do they really all have value to the types of the forecasts you want?

   I am often frustrated at the 'many hydrological signatures' means a good jib has been one but without justification for their use. So we should justify this better, to explain the experimental designs we use. Again a different set of metrics or more or less of these signatures would generate potentially different results, and who is to say what best highlights 'what you need' for a given forecasting objective... And as I stated before why indeed a more comprehensive treatment of the model hindcast skill using various calibration metrics to test different periods and magnitudes of flow, were not used?

   This is a very good point, which, to our knowledge, has not been addressed by previous investigations. As McMillan et al. (2017) stated, there is still a lack of consensus on a comprehensive set of hydrologic signatures to be used by the hydrological community.

   Please further note that an investigation of the sensitivities to the selected hydrological signatures is not within the objectives of this article. Here, we aim to identify links between forecast skill and commonly used hydrological signatures. Our selection of signatures is driven by previous literature on hydrological classifications (Euser et al., 2013; Viglione et al., 2013), and by applications lead by scientists at SMHI oriented towards process understanding (Pechlivanidis and Arheimer 2015; Kuentz et al., 2017) and forecasting skill attribution (Pechlivanidis et al. 2020). We are also aware of other hydro-climatic clustering investigations (i.e. Knoben et al., 2018). Yet, to our knowledge, there is no investigation that guides the community towards a unique set of signatures for identifying hydrologic similarity (for example, Westerberg et al., (2016) only approached this topic from the view of uncertainties in hydrological signatures).

   Furthermore, it is important to note here that, in our paper, we firstly assessed the link between forecast skill and individual hydrological signatures (step 1). To further generalize the insights from step 1, we investigated the link between forecast skill and hydrological regimes, as these are defined by the clustering of the signatures. We assume that the selected signatures are robust enough to guide the analysis towards the correct identification of hydrologically similar river systems.

Regarding the last question, please see our response to point 1.

We addressed this comment by adding these considerations to Sections 2.4 and 4.3 of the manuscript.

3. Finally I also want to clarify your use of a scale for KGE in Figure 1. Can you explain why you have set the scale to zero? KGE does not have the same properties as NS Efficiency where 0 is the mean predictor, in fact the mean predictor is -0.41 Knoben, W. J. M., J. E. Freer, and R. A. Woods (2019), Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores, Hydrology and Earth System Sciences, 23(10), 4323-4331. for an explanation of this (please note I am not at all trying to get you to cite our paper). I just think you want to explain better why your figure has chosen certain limits.

   Please note that in Figure 1a the lower limit is not 0. Yet, all stations with KGE values lower than 0 are represented in grey colour. We took the decision to group all stations (in total 8 stations out of 539) with KGE < 0, since the paper's focus is not on digging deeper on historical model performance and its spatial variability. As we explain in the discussion, poor model performance is observed in river systems that are regulated, since regulation schemes are almost impossible to fully represent. In such cases, model performance is thus poor in terms of timing (correlation coefficient) and variability (alpha term in KGE), which negatively affect the KGE values. Consequently, we do not believe that increasing the scale would add any value to the paper. On the contrary, it may distract the reader from the main objective of Fig. 1a, which is to show that KGE is very high for most stations.

References

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. Hydrology and Earth System Sciences, 17(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

Kuentz, A., Arheimer, B., Hundecha, Y., & Wagener, T. (2017). Understanding hydrologic variability across Europe through catchment classification. Hydrology and Earth System Sciences, 21(6), 2863–2879. https://doi.org/10.5194/hess-21-2863-2017

Knoben, W. J. M., Woods, R. A., & Freer, J.E. (2018). A quantitative hydrological climate classification evaluated with independent streamflow data. Water Resources Research, 54, 5088–5109.https://doi.org/10.1029/2018WR022913

McMillan, H., Westerberg, I. and Branger, F.: Five guidelines for selecting hydrological signatures, Hydrological Processes, 31(26), 4757–4761, https://doi.org/10.1002/hyp.11300, 2017.

Pechlivanidis, I. G., & Arheimer, B. (2015). Large-scale hydrological modelling by using modified PUB recommendations: the India-HYPE case. Hydrology and Earth System Sciences, 19, 4559–4579. https://doi.org/10.5194/hess-19-4559-2015

Pechlivanidis, I. G., Crochemore, L., Rosberg, J., & Bosshard, T. (2020). What are the key drivers controlling the quality of seasonal streamflow forecasts? Water Resources Research, 56, e2019WR026987. https://doi.org/10.1029/2019wr026987

Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins – Part 3: Runoff signatures in Austria. Hydrology and Earth System Sciences, 17(6), 2263–2279. https://doi.org/10.5194/hess-17-2263-2013

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., & Freer, J. (2016). Uncertainty in hydrological signatures for gauged and ungauged catchments. Water Resources Research, 52, 1–19. https://doi.org/10.1002/2015WR017635