

We thank the reviewer for their interest and their constructive feedback. Reviewer comments are reproduced below in grey indented text and responses follow in black.

The authors touch a very important and increasingly spotted (luckily) topic: should we blindly trust our traditional performance metrics for hydrological modeling? Aside other very interesting insights, they discuss a sad (although needed) truth: high NSEs (or even KGEs) do not necessarily mean that the simulations are adequate. It urges in some aspect our need to improve, as modelers, our optimization metrics. The paper is definitely a fit for HESS and **should be published**, but as should be expected, some concerns should be clarified/corrected/improve before, aside many suggestions.

1. I believe that the methodology used for the time-series decomposition needs to be better explained (with more details) and if needed, authors could make use of Appendix/Supporting information. This is a crucial part and needs to be ensured to be easy to follow by readers.

We will expand the description of the time series decomposition and include a flowchart of the steps in the SI, with some more example time series.

2. Also on that, I feel that authors could justify better the choice of the decomposition. Was it motivated by previous work? Are there more references? This needs to be made clear in the text.

We don't believe the exact decomposition that we used has been used before. However, others have used similar methods, and we include some references here. We will also include, in the SI, examples of decomposition using two other methods that we assessed: classical decomposition and Seasonality-Trend decomposition using LOESS (STL). These methods result in similar geographic patterns but their mathematical properties are a bit messier.

3. The authors called the seasonal component the long-term seasonality of the basins. Our rivers are under changes and the seasonality is consequently changing in many of our rivers. I think this could fit a bit better in the text. I understand the choice (L85-89), and also I believe that much of the change is captured in the irregular, but the text would benefit for a bit of clarification in the choices.

In our decomposition, any variance that is not captured in the (static) seasonal component will be captured in either the irregular or interannual components. A different mathematical definition of the seasonal component could allow the seasonality to vary with time, (for example, the STL decomposition) but this introduces new complications such as needing to

define parameters that control the rate of change of the seasonal component. No doubt this is possible, but we found our chosen method easy to explain succinctly and we believe this provides a good first step for this sort of analysis. We will discuss this point in the expanded description of the time series decomposition.

4. Simulations: If I understood correct, the authors used simulated data from several models (and in one case they simulated themselves). Did the authors check for the different periods of calibration/evaluation/tests for all the models? Or for overlapping period? Did the authors used only what was classified as test? my main concern, is that during the model comparison, the authors might be using simulated streamflow from test for some models and for "calibration" for other models. Or even, single-basin versus regional simulations. I see no problem in using different settings, but this needs to be extensively reported and discussed in the results. For example, I have the feeling that for the PREVAH-CH simulations, the authors might have used all the simulation (including calibration) and not only evaluation (I might be wrong). My suggestion is to review these aspects, and incorporate such information in the manuscript.

The reviewer is correct that the models represent a mix of calibration and test/validation periods. Where the simulations were available separately for the test period, we evaluated the models only on this test period. Similarly, where simulation outputs were available for ‘pseudo-ungauged’ basins (basins not used in calibrating/training the model) we used these simulations. For seven models (including the PREVAH-CH model) at least some of the test data were used in the calibration of the model.

We will clarify these points in the manuscript and include, in the SI, a table of the calibration/validation/test periods used for each model, and specify which period was used for each. Overall, we don’t believe this methodological variability affects our conclusions, which rely on within-model comparisons.

5. Regarding Figure 3 (along also L275 onwards) models that performed better for highly seasonal catchments were the ones with the lowest performances overall, or is it my impression? I think you should discuss better this, maybe showing the median performances? A box plot in appendix? Something to clarify if these models being better in seasonal are actually just the case that they had overall poor performance? Also touching point 4, how were these simulations obtained by the original authors? did they report them as the evaluation phase? or are they actually for the calibration period? This would be worthy clarifying for the readers.

Yes, the two models where the seasonal catchments perform better do seem to have lower performance overall. We allude to this at line 281:

On the other hand, the better representation of interannual and irregular variance in highly seasonal catchments in these two models may also be related to the poor performance in less seasonal catchments, as suggested by the low overall NSEs for less seasonal catchments.

However, these models (CE-COSERO and WNA-VIC-GI) do not use the same set of catchments as any other models within our sample, so to compare NSEs (including the component NSEs) across models is not a strictly valid exercise. We think it is best to remove the discussion of the two 'exceptions' (L276-284) to avoid confusion.

6 L328-332: Needs to be rephrased (maybe) after reviewing points 4 and 5.

We will also remove L328-332.