Reviewer 1

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 decommission 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 have expanded the description of the time series decomposition and now include a flowchart of the steps in the SI, along with a more in-depth description.

> *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. We include at L106:

This particular decomposition method is a novel contribution. However, deseasonalizing data to obtain anomalies is a standard technique (Wilks, 2006), and spectral analysis (including using Fourier transforms) has been applied to streamflow to characterize hydrologic regimes (Brown et al., 2023; Smith et al., 1998).

> *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. We note this at L115:

Other decomposition methods are possible, including using wavelet transforms instead of Fourier analysis, or allowing the seasonal component to vary with time.

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 six models (including the PREVAH-CH model) at least some of the test data were used in the calibration of the model. In reviewing this point, we found that one model (CE-COSERO) in fact did use a split-sample methodology. We have updated our analysis to use only the test period for this model. This does not significantly change our results or conclusions.

We have made two changes: first, we simplified Table 2 in the main text to only include model name, type of model, region, and reference. Second, we included an expanded table (Table

S1) in the SI, which includes the type of calibration (global or basin), the type of training/testing split, the number of catchments, the percent of highly seasonal catchments, and anthropogenic impacts.

Overall, we don't believe that the mixing of testing and training periods should affect our analysis, since we focus on within-model comparisons of seasonal and non-seasonal catchments.

Regarding single-basin vs regional simulations, at line 329 we discuss this:

Where models are optimized simultaneously across many catchments, optimization algorithms may therefore greedily simulate the seasonal variance in seasonal catchments and neglect interannual and irregular variance. When seasonal variance is well-simulated, algorithms will prioritize improvements to modelling the irregular and interannual components in less-seasonal catchments where these improvements result in the largest increase to the average NSE. Learned human biases for what a 'good' simulation looks like could also bias modellers to neglect catchments where errors appear small compared to the seasonal pattern, since expert opinions and current quantitative metrics have been shown to be mostly consistent (Gauch et al., 2023). However, six of the models that we analysed were calibrated individually to each catchment, so this explanation is not sufficient on its own.

> *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 performed better (the 'exceptions') do seem to have lower performance overall. We alluded to this at line 281 in our initial submission:

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-GL) 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. In addition, for this revision

we evaluated CE-COSERO only on the subset of years use as ‘test’ in training the model, and this eliminated the ‘exception’; now, the highly seasonal catchments have significantly worse $NSE_{interannual}$ than the non-seasonal catchments. The statistics in Figure 4 did not change very much.

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 have also removed L328-332.

Reviewer 2

We thank the reviewer for their detailed review, and for their comments which will improve the clarity of the manuscript. Reviewer comments are reproduced below in grey indented text and responses follow in black.

> *This Technical Note (TN) uses available streamflow simulations from several datasets to show that high NSE and KGE values for seasonal catchments do not necessarily translate in good model performances of interannual variability of streamflow. This is a relevant topic in the scope of HESS. I agree with all points raised in RC1 and provide a few comments below.*
>
> *TITLE – the title could be more straightforward regarding highly seasonal streamflow regimes rather than specifying tropical, alpine, and polar catchments.*

We have revised the title to: "Technical Note: High Nash Sutcliffe Efficiencies conceal poor simulations of interannual variance in seasonal regimes".

> *ABSTRACT – the short summary reads better than the abstract. The introduction of the abstract is too long. There should be only one opening sentence (e.g., "…common metrics used to evaluate hydrological models…") followed by a sentence clarifying the scientific gap (e.g., "however, simulating interannual variability might be a problem…"). It should be made clear that the paper is mostly based on simulation available in the literature (i.e., the sentences "we show that hydrologic models…" and "we analyse 18 regional and global hydrologic models…" are quite ambiguous regarding the nature of this technical note).*

We have rewritten the abstract to be clearer about the scientific gap and the models that we analysed. It is reproduced below:

In highly seasonal regimes hydrologic models generally achieve high scores on common performance metrics such as the Nash-Sutcliffe Efficiency (NSE) and the Kling-Gupta Efficiency (KGE). However, variance in streamflow time series is composed of seasonal, interannual, and irregular variance, and the NSE and KGE do not differentiate between these components. Differences in performance on these three components have not been evaluated across a broad spectrum of hydrologic models and regions. We analyse open-access simulations from 18 regional and global hydrologic models. We find that in highly seasonal catchments models consistently achieve the highest NSE and KGE but are regularly worse at simulating interannual variability than in less seasonal catchments with lower NSE and KGE scores. The NSE of the interannual component is lower in highly seasonal catchments, and simulated year-to-year changes in ecologically relevant

hydrologic signatures are less accurate. This suggests that these hydrologic models may struggle to predict long-term responses to climate change, especially in highly seasonal tropical, alpine, and polar regions, which are some of the most vulnerable to climate change. We encourage hydrologic modellers to explicitly evaluate skill at simulating interannual variability, rather than relying only on aggregate measures such as the NSE and KGE.

> *L13 – is "irregular variance" the best term here?*

The decomposition of time series into trend/interannual, seasonal, and irregular components is standard terminology (Persons, 1919). Other terms have also been used for the irregular component (including remainder, noise, residual, and error) but we feel that 'irregular' is the most appropriate term in a hydrological context. At L40 we have added these other terms in parentheses.

> *L20 – how were "ecologically relevant" signatures determined?*

These are based on the indicators of hydrologic alteration (IHA) in addition to some other signatures that have been used in the (eco)hydrology literature – see Table 1 for references. We will add the following underlined text at L136:

"<u>The ecologically-relevant signatures that we consider are</u> the 32 indicators of hydrologic alteration proposed by Richter et al. (1996) in addition to…"

> *L21-23 – It would be nice to finalize the abstract with the important technical implications for hydrologic modeling (so what should we do now?) rather than a general comment about climate change and vulnerable regions (not really the core topic of this TN).*

We have now included a recommendation about model evaluation. See abstract above.

> *INTRODUCTION – The story is not clear. First 12 lines about streamflow and climate change. But this TN is about performance metrics. It seems that the most important paragraph starts at L50. This paragraph should be developed further to clarify the relevance of this TN.*

We have revised the introduction significantly to clarify the story. We now introduce the main problem at the end of the first paragraph:

However, there is a disconnect between how hydrologic change is assessed (using hydrologic signatures) and how hydrologic models are typically trained and evaluated (using aggregate performance metrics), which may lead to inaccurate predictions of future hydrologic change.

L37 & L56 – What about each of these references is interesting? Expand on it or cut it out.

L37: The references highlight different types of non-stationarity and how hydrologic signatures are used to detect and/or measure it. We have clarified the relevance of each reference to different types of hydrologic signature.

L56: These were simply examples of papers that compare their model performance to the climatological benchmark. We have removed the sentence.

METHODS –

Section 2.1 describes several data selection choices. Perhaps, moving Section 3 before Section 2 would be better.

Thanks for the suggestion. We prefer to describe the methods first, otherwise it is unclear for what the data will be used. The data selection choice (filtering to gauges with a minimum of 10 years of data with no missing days) is an important part of the methods, without which it is not possible to calculate the Fast Fourier Transform.

L74-77 – this is not completely clear.

We have expanded the description of the decomposition, also in response to reviewer 1. There is now also a section in the SI with a flowchart outlining the process.

Section 2.2 – this was done by running a model or using available simulations? The language is ambiguous here.

The climatological benchmark 'model' is not really a model in the way that most hydrologists think of models. It is simply the mean flow for each calendar day of the year. This is briefly explained in the introduction but we have revised to remind the reader of this fact in Section 2.2:

To answer our second question, we calculated the NSE_{cb} (the NSE for a climatological benchmark model defined as the interannual mean flow for each calendar day) for 20,338 catchments

L91 – Isn't the NSE using the average streamflow as a benchmark?

Yes, this is the benchmark inherent in the NSE. However, it may be an unreasonably naïve benchmark for seasonal catchments. To quote Schaefli & Gupta (2007):

"The use of the mean observed value as a reference can be a very poor predictor (e.g. for strongly seasonal time series), or a relatively good predictor (e.g. for time series that are essentially fluctuations around a relatively constant mean value)."

For this reason, the climatological benchmark performance NSE_{cb} is sometimes used as an alternative benchmark. We have changed the section title to "Climatological benchmark performance."

> *Section 2.3 – again using "Modelling" in the title is a bit ambiguous as to the methods.*

We have revised to *Representation of interannual and seasonal variability in hydrologic models*

> *L110 – Where do I_o and I_s come from?*

We have clarified:

Where I_o and I_s are the observed and simulated interannual components as derived from time series decomposition (Section 2.1).

> *L128 – Why is this interesting? It should be clear in the intro why changes in hydrological signatures should be evaluated. There are several important references missing here.*

We have revised the introduction substantially.

> *DATA – This data and simulation use should be clarified in the abstract and introduction sections.*

We have revised both to clarify the use of 'open-access' datasets and simulation outputs, which should clarify that we did not perform most of the simulations ourselves.

> *RESULTS AND DISCUSSION*
>
> *Section 4.1 – What is a highly seasonal catchment? What are the signature values that were used to classify the catchments?*

In section 2.3.1 we define 'highly seasonal' to mean a seasonal variance fraction greater than 0.5. We now remind readers of this in paragraph 3 of Section 4.1:

Highly seasonal catchments (those with a seasonal variance fraction above 0.5)

> *L194-205 – A lot of climatological explanation here, but nothing about important hydrological catchment characteristics. What is the area of the chosen catchments? What is average annual rainfall? What is ET? Why were these three catchments selected?*

These catchments are selected to illustrate the diverse drivers of interannual variability around the world. They are not exhaustive and we did not perform any formal analysis of how

high interannual variability is controlled by catchment characteristics. We would prefer not to overload the reader with a long list of catchment characteristics and encourage the reader to look at the references, which provide much more detail about the locations in question.

We have added the following : “The preceding examples are selected to illustrate the diverse drivers of interannual variability around the world. In-depth analysis of their hydrologic conditions is considered out of scope for this work.”

> *Section 4.3 The discussion here is not linear and difficult to follow. This section could be reduced considerably and the paragraphs should be grouped around main messages.*

We will have revised and condensed this section, and moved some text to the SI. We also added subheadings to signpost the topics.

> *L275 – Is this hypothesis exhaustive? Could you think about any other case where that would happen or any exception to this?*

We have removed this line and paragraph in the revised text.

We do not claim to have proven that hydrological models will always simulate interannual variability more poorly in seasonal catchments. In fact, we hope this is not the case! We have included discussion of the possible reasons for this behaviour, and possible solutions, in Section 4.3.3.

> *CONCLUSIONS AND RECOMMENDATIONS – This section is a bit convoluted and repetitive. The conclusions should strictly address the knowledge and recommendations without repeating the results section (e.g., “higher than 0.8. We observe, in Figure 3…”).*

We have revised and condensed.

> *L378-383 – a bit of repetition of the introduction.*

We are a bit confused by this comment as L378-383 (reproduced below) are stating and interpreting results of the current study, and we don’t feel that any of the statements from L378-383 repeat information in the introduction.

L378: This is most evident in tropical, alpine, and polar regions, where most of the variance in streamflow is seasonal. Poor interannual performance in these regions (and in some cases almost complete failure to simulate year-to-year variability) raises concerns about the ability of these models to accurately simulate nonstationary hydrologic processes and responses to climate change. This is especially worrying because these regions may be some of the most vulnerable to climate change (Flores et al., 2024; Pepin et al., 2022;

Rantanen et al., 2022) and are historically less-well studied regarding hydrologic extremes (Stein et al., 2024).

L387 – "Lastly…" Why is that? How much is enough?

As discussed from L347-L367 these regions are underrepresented in available datasets. Observational data are required to assess how well models perform in these regions, and to infer how well they might predict long-term changes. We have revised:

Lastly, we stress the need to collect and publish observations from more tropical, alpine, and polar regions, which are underrepresented in global datasets.