Title: "Evaluation of global hydrological models for climate change impact assessment: How well can they translate interannual climate variability into streamflow variability?" by Thedini Asali Peiris and Petra Döll, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/egusphere-2026-3412, 2026

Response to the peer reviewers' comments

Our response is written *in italics*. Changes in the revised text are shown **in bold**.

Comments from Reviewer 2 (Anonymous Referee #2)

The manuscript proposes evaluating global hydrological models (GHMs) for climate change suitability by comparing simulated and observed annual streamflow anomalies rather than raw magnitudes, arguing this better isolates climate-driven variability from structural bias. Three GHMs are tested against 589 gauges for two sets of climate forcing. The topic is an important one given how widely GHMs are relied upon for water resource and climate change decision-making across jurisdictions, often in contexts where basin-specific calibration is impossible. Any methodological advance that helps end-users judge which models are fit for purpose in climate impact work has potential for practical value. The paper is well written throughout: the structure is logical, the figures are high quality and the manuscript is commendably concise relative to its scope. However, I feel some changes in framing and coverage of literature are warranted, and in addition I suggest an additional analysis that would add value, below.

AC8: The authors thank the reviewer for the careful review and constructive feedback. We particularly appreciate the reviewer's recognition of the potential value of this work in providing end-users with additional information for assessing the suitability of GHMs for climate change impact applications.

Title and framing

The title promises an analysis of how models translate interannual climate variability into streamflow variability, which primes the reader to expect engagement with modes of climate variability (ENSO, IOD, etc.) and how their signals propagate into streamflow. That is not what the paper delivers. What is actually assessed is model performance on annual streamflow anomalies (residuals) rather than an explicit climate-to-streamflow translation, and rather than daily/monthly values. This is a legitimate and different critique of the models, and the title should be revised to reflect it accurately rather than imply a climate-driver analysis that isn't attempted.

AC9: We agreed with the reviewer and revised the title to:

Evaluation of global hydrological models for climate change impact assessment: How well can they reproduce observed interannual streamflow variability?

Literature review gaps

Despite the introduction's length, three adjacent bodies of literature that bear directly on the paper's stated motivation are absent. I feel all three should be covered in the introduction, and the third could also be brought within scope as per my comments later.

1. Drought propagation: if GHMs are to be trusted for climate projections, their ability to propagate anomalies through the hydrological system with realistic lags and persistence is directly relevant, yet this literature is not engaged with (e.g. van Loon and others).

AC10.1: We thank the reviewer for highlighting the relevance of drought propagation studies. We agree that hydrological processes can influence the duration and persistence of streamflow deficits. We have therefore added the following text to the Introduction:

"Hydrological processes can also influence the duration and persistence of streamflow deficits, according to the drought propagation studies (Van Loon et al., 2012). Consecutive years with below-mean streamflow may therefore represent prolonged periods of reduced water availability, even when individual years do not meet a specific drought threshold."

As this study focuses on evaluating the ability of GHMs to reproduce observed interannual streamflow variability rather than specifically assessing drought-related applications, we have not included a detailed review of drought propagation studies.

Nevertheless, following the reviewer's suggestion, we have extended our analysis to evaluate the maximum duration of consecutive positive and negative annual streamflow anomalies. We have added the following sentence to the study objectives in the Introduction:

"In addition, we evaluate the ability of GHMs to reproduce the maximum duration of consecutive years with positive and negative relative annual streamflow anomalies."

The methodology and results are presented in the new Sections 3.4 and 4.4, respectively, and their relevance and limitations are discussed in Section 5.1.

2. Shifts in the rainfall-runoff relationship: work such as Saft et al. (https://doi.org/10.1002/2015GL067326) has shown hydrological models struggle to represent non-stationary rainfall-runoff relationships, and others such as Avanzi (https://doi.org/10.5194/hess-24-4317-2020) and Massari (https://doi.org/10.5194/hess-26-1527-2022) have shown non-stationarity occurs across continents, so I think they warrant a mention here.

AC10.2: We thank the reviewer for highlighting these relevant studies. We agree that changes in rainfall–runoff relationships under changing climatic conditions are important when interpreting historical GHM performance. This concern has also been raised by Reviewer 1 and has been addressed in our response to Reviewer 1, AC3.

We have therefore strengthened Section 5.1, which discusses the interpretation and the limitations of the anomaly-based evaluation, to acknowledge the potential non-stationarity of hydrological responses and its implications for the transferability of historical model performance. We have included these studies in the Discussion, where they directly support the interpretation of the limitations of historical model evaluation for climate change impact assessments. The revised text states:

"Hydrological responses can change during prolonged climatic shifts; for example, prolonged dry conditions have been shown to alter rainfall-runoff relationships and water-balance partitioning (Saft et al., 2016; Avanzi et al., 2020; Massari et al., 2022). Such changes can also reduce the transferability of hydrological model performance across climatic conditions (Coron et al., 2012)."

3. Runs of dry years: because large reservoirs can buffer a single dry year, multi-year dry sequences may matter most for water security. The study of the tendency of flow sequences to bunch into sequences above or below the mean has a long history (e.g. Hurst's work on the Nile) but the current annual, single-year framing de-emphasises this.

AC10.3: *We thank the reviewer for highlighting the importance of consecutive years with reduced streamflow for water resources assessments. While this analysis is related to the broader question of persistence in hydrological time series highlighted by the reviewer, it does not assess long-term dependence using the rescaled-range approach associated with Hurst's work.*

The methodology and results are presented in the new Sections 3.4 and 4.4, respectively, and their relevance and limitations are discussed in Sections 5.1.

Reframing away from NSE improvement

Rather than emphasising the raw magnitude of NSE improvement under the anomaly configurations, the paper would be more valuable if it characterised specific cases (locations) where a model previously considered unsuitable (due to magnitude bias) might now be reconsidered as suitable for projections because the percentage changes it produces are directionally and proportionally correct. This is a more actionable framing for the paper's stated audience (those assembling model ensembles for impact studies).

AC11.1: The manuscript already uses the illustrative station presented in Figure 2 for this purpose, and the corresponding location is annotated in Figure 3, A1 and A2, together with the respective NSE values, to facilitate comparison of NSE values across the different streamflow representations. However, we agree that the relevance of this example for model evaluation was not sufficiently elaborated in the text. We have therefore added the following text to Section 4.1 and improved the caption of Figure A1 to clearly identify the illustrative station.

"As demonstrated by H08, a model showing poor performance in reproducing absolute annual streamflow may still reproduce certain aspects of proportional interannual streamflow variability, while the contrasting behaviour of WaterGAP indicates that good performance for absolute streamflow does not necessarily translate into equally good performance for relative anomalies."

We have retained one illustrative station because the purpose of this example is to demonstrate how Q-based and anomaly-based evaluations can provide different information about model performance. Additional examples could illustrate similar contrasts, but explaining the location-specific and model-specific causes of these differences would require further investigation beyond the scope of this study.

Relatedly, the NSE comparison between the Q and Q^*/Q^{**} configurations is not strictly apples-for-apples once you are calculating NSE over a sequence of anomalies rather than the original series. This should be acknowledged and the "NSE improvement" narrative (e.g. in the labelling of Figure 3) de-emphasised.

AC11.2: Thank you for pointing this out. We agree that NSE values calculated for Q, Q^* and Q^{**} represent different aspects of model performance and that higher anomaly-based NSE values should not be interpreted as improvements in overall model performance.

We have clarified this distinction in Section 4.1:

"At this station, the differences in NSE among the models for Q and Q^{**} therefore reflect the different properties evaluated by the two configurations rather than differences in overall model performance."

Furthermore, we have removed references to "improved NSE" and "improved performance", replacing them with terminology such as "higher NSE" or "increased NSE values", where appropriate. Accordingly, we have revised the text in Section 4.2 and the labels and captions of Figures 3 and A2 to describe differences in NSE values without implying an improvement in overall model performance.

Separately, since the NSE reported in Table 3 (for Q_{90}/Q_{10}) is a single value computed across all 589 stations rather than per-station, this should probably be labelled a "spatial NSE" (or similar) to distinguish it from the per-station, temporally-computed NSE in Table 2. Readers will likely get confused otherwise.

AC11.3: Thank you for this suggestion. We agree that the distinction between the NSE values reported in Tables 2 and 3 should be made clearer. We have therefore changed the column heading in Table 3 from "NSE" to "Spatial NSE" and added the following clarification to its caption:

"Spatial NSE is calculated across the 589 analysis locations using extreme anomaly values derived from their respective 30-year periods."

We have also updated Section 3.3 to use the term "spatial NSE" consistently. This distinguishes the NSE calculated across analysis locations using extreme anomaly values in Table 3 from the station-specific temporal NSE reported in Table 2.

Suggested additional analysis

In line with the above comments, the method analyses the statistics of anomalies but discards information about their sequence. Streamflow anomaly sequences tend to show more persistence than the climate sequences that generate them, and this is an important behaviour for models to replicate. It would be easy enough to analyse the length of runs above and below the mean here, and I strongly recommend the authors consider this.

AC12: We thank the reviewer for this suggestion. We agree that the persistence of annual streamflow anomalies is an important aspect of GHM performance, particularly for climate change impact assessments concerned with consecutive years of reduced water availability.

We have therefore extended our analysis to evaluate the maximum duration of consecutive positive and negative relative annual streamflow anomalies (Q^{**}). We selected the maximum run duration to specifically assess the ability of GHMs to reproduce the duration of prolonged periods of above- or below-mean streamflow.

The results show that H08 and WaterGAP generally reproduce the observed maximum run durations more closely than MIROC-INTEG-LAND, which tends to simulate longer durations, particularly for negative anomalies. However, this analysis compares the maximum run durations without requiring the observed and simulated runs to occur during the same years.

We have added the methodology in Section 3.4, the results and a new figure (Figure 5) in Section 4.4, and discussed the findings and their relevance to climate change impact assessments along

with the limitations in Section 5.1. Following that the main findings are included in the Conclusions.

Discussion

The paper references existing model-weighting schemes for multi-model ensembles (e.g., Krysanova et al., 2020) but does not extend its own findings into guidance. Given the paper's argument that relative changes could be a new focus of climate-impact evaluation, it would strengthen the discussion considerably if the authors offered concrete guidance on how ensemble weighting schemes should be constructed based on relative changes.

AC13: We thank the reviewer for this suggestion. We agree that the implications of anomaly-based evaluation for selecting and weighting GHMs within multi-model ensembles should be discussed more explicitly.

We have therefore revised Section 5.2 to provide guidance on how the evaluation results could inform model selection and weighting, depending on the quantity of interest in a climate change impact assessment. When relative hydrological changes are of interest, models should not be excluded or strongly down-weighted solely because of poor performance in reproducing absolute streamflow magnitude. Instead, anomaly-based performance should be considered alongside Q-based performance, with their relative importance depending on whether the assessment focuses on absolute streamflow conditions or relative hydrological changes.

Specifically, we have added the following clarification in Section 5.2:

"However, when the quantity of interest is relative hydrological change, poor Q-based performance alone may not provide sufficient grounds for excluding or strongly down-weighting a GHM. A model that poorly reproduces streamflow magnitude may nevertheless reproduce observed proportional interannual streamflow variability. Anomaly-based performance could therefore be considered alongside Q-based performance when characterising individual ensemble members for climate change impact assessments concerned with relative hydrological changes. Conversely, good anomaly-based performance should not compensate for poor Q-based performance when accurate streamflow magnitude is required. The relative importance of the two evaluation dimensions should therefore depend on the quantity of interest in the climate change impact assessment."

We have not proposed a specific numerical weighting scheme, as this would require additional methodological choices and evaluation beyond the scope of the present study. Furthermore, as discussed in Section 5.1, historical anomaly performance alone cannot establish the reliability of future hydrological projections, particularly when model process representations respond differently to future climate conditions, such as active vegetation response to the increasing atmospheric CO_2 and other climatic changes.

Also, line 294: among what? Consider expanding this sentence.

AC14: Thank you for pointing out this incomplete sentence. We have corrected it in Section 4.2 as follows:
"Correlation (r) remains unchanged among Q, Q^* and Q^{**}."