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 1 (Anonymous Referee #1)

Summary

This paper argues that, because simulations from Global Hydrologic Models tend to be biased and sub-yearly streamflow is hard to model accurately for a variety of reasons, assessing relative annual streamflow anomalies is a valuable addition to model evaluation practices. The paper then applies an example of this to simulations from 3 GHMs, using a set of some 500 gauges that have at least 30 years of data available. Practically, simulations are first de-biased ($Q > Q^{*}$) and then rescaled by their own simulated mean ($Q^{*} > Q^{**}$). The authors then discuss the performance of the three GHMs on Q, Q^{*}, and Q^{**}.

First of all, the authors thank the reviewer for the constructive feedback on the manuscript.

Comments

I do not believe this paper achieves its stated goal. To the best of my understanding, the authors simply state that assessment of model performance on relative streamflow anomalies is helpful, then do such an assessment, and then conclude that this assessment of model performance is more robust than assessing model performance on annual streamflow directly. What's missing is the evidence that this is in fact more robust, or even helpful.

AC1: We thank the reviewer for this important comment. We agree that, in some places, the manuscript overstated the evidence for describing anomaly-based evaluation as "more robust" than evaluation of original annual streamflow time series (Q: hereafter: Q-based evaluation). This claim was not sufficiently supported by our analysis. We have therefore removed it and revised the manuscript to clarify that anomaly-based evaluation is intended to complement, rather than replace or be considered superior to, Q-based evaluation of GHM performance.

However, we do not agree that the anomaly-based assessment is therefore uninformative. Q-based and anomaly-based evaluation assess different aspects of model performance. Q-based evaluation assesses agreement between the observed and simulated annual streamflow time series and is influenced by both differences in streamflow magnitude and the representation of interannual variability. In contrast, Q^{*} removes differences in the respective mean streamflow from the comparison while retaining the absolute magnitude of interannual deviations, whereas Q^{**} additionally normalises these deviations by the respective mean of each time series. Anomaly-based evaluation therefore provides a more explicit assessment of the ability of GHMs to reproduce observed interannual streamflow variability, including its proportional magnitude. This is particularly relevant for climate change impact applications in which hydrological changes are expressed relative to a reference period.

To clarify these distinctions, we introduced the term "Q-based evaluation" in Section 3.1 and revised the Abstract, Introduction, Discussion (Sections 5.1 and 5.2), and Conclusions. We also removed statements suggesting that anomaly-based evaluation is inherently more robust or that higher anomaly-based performance metric values demonstrate improved overall model performance.

Section 5.1 now states:

"Among the various methods available for evaluating GHMs, anomaly-based evaluation provides complementary information that is particularly relevant for climate change impact assessments in which relative hydrological changes are of interest."

It also clarifies the relationship between historical interannual variability and future hydrological changes:

"Although these changes are evaluated over different timescales, both are expressed as proportional changes relative to a reference value. Therefore, evaluating the ability of GHMs to reproduce observed relative annual streamflow anomalies provides information relevant to assessing their ability to represent proportional hydrological responses, which are of interest in climate change impact assessments."

Section 5.2 explains how this additional information may be used in practice:

"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."

Finally, Section 6 (Conclusions) reinforces the complementary interpretation:

"These findings demonstrate that 1) agreement in streamflow magnitude and agreement in interannual variability represent complementary aspects of historical GHM performance, and 2) GHMs differ in modeling the hydrological response to annual climate variability."

The revised Discussion and Conclusions also explicitly acknowledge that reproducing historical interannual streamflow variability does not guarantee accurate future hydrological projections, particularly because hydrological process responses may change under future climatic conditions.

Using the single gauge in Figure 2 as an example, the simulations from WaterGAP are much closer to the observed annual values than those of the other two models. However, when relative anomalies Q^{**} are used, WaterGAP emerges as the worst of the three, purely because the anomalies are calculated with respect to each model's own long-term mean flow. Because WaterGAP's mean flow is much closer to observations, its relative anomalies are larger than those of H08, despite being about an order of magnitude smaller in absolute terms. It's entirely unclear to me what I can learn from the Q^{**} assessment that isn't misleading about the model's actual capabilities. If the authors wish to argue that the Q^{**} analysis is more robust/meaningful than Q-based analysis, then they must show evidence that this is the case. Right now, I think Figure 2 mostly suggests that the Q^{**} approach isn't helpful.

AC2: We thank the reviewer for highlighting this point. We understand the reviewer's concern that, in the example shown in Fig. 2, WaterGAP reproduces the observed annual streamflow much more closely than H08 and MIROC-INTEG-LAND, whereas this ranking changes when the models are evaluated using relative anomalies (Q^{**}). We agree that this result could be misleading if the Q^{**} performance were interpreted as an indication of overall model

performance. This was not our intended interpretation, and we have revised the manuscript to make this distinction explicit. We also agree with the reviewer that our previous framing of anomaly-based evaluation as generally "more robust" or meaningful than Q-based evaluation was not sufficiently supported, and this claim has been removed.

The change in model ranking between Q and Q^{**} reflects the different properties evaluated by the two configurations. Q-based evaluation assesses agreement between the observed and simulated annual streamflow time series, whereas Q^{**} assesses agreement in the direction and proportional magnitude of interannual deviations relative to the respective mean of each time series. Normalising the anomalies by the respective mean streamflow changes the scaling of simulated variability, which can lead to different model rankings. Consequently, a higher NSE for Q^{**} does not imply that a model reproduces the original annual streamflow time series better.

The information provided by Q^{**} is whether a model reproduces observed proportional interannual streamflow variability, even when its simulated mean streamflow is biased. This is relevant to climate change impact assessments in which hydrological changes are expressed relative to a reference period. However, historical annual anomalies and future changes relative to a reference period are not equivalent, and successful reproduction of historical anomalies does not establish the reliability of future projections. We therefore consider anomaly-based evaluation a complementary diagnostic rather than a replacement for Q-based evaluation.

We have revised Section 4.1 to state this distinction explicitly:

"Under 20CRv3-ERA5, their NSE values for Q^{**} (0.27 for both models) are higher than that of WaterGAP (-0.06), despite WaterGAP showing much closer agreement with the observed mean streamflow and substantially higher NSE for Q. For Q^{**}, the comparison concerns proportional deviations relative to the respective mean of each time series and should therefore not be interpreted as indicating that H08 or MIROC-INTEG-LAND reproduces the original annual streamflow time series better than WaterGAP. In contrast to Q^{*}, relative anomalies modify the scaling of variability, resulting in changes in γ, while r remains unchanged. 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."

We have also clarified the purpose of Fig. 2. The station was randomly selected from the 589 analysis locations and is used to illustrate how the evaluation configurations affect performance metrics at an individual location, rather than to establish general model rankings. Section 4.1 now states:

"The example is intended to illustrate how the different evaluation configurations affect the performance metrics at an individual location; global performance patterns across all 589 analysis locations are presented in Section 4.2."

Finally, we have revised the Discussion to clarify the interpretation and limitations of relative anomaly-based evaluation. Section 5.1 states:

"In global hydrological modelling contexts, where calibration is challenging, relative anomaly-based evaluation can therefore identify GHMs that reproduce observed proportional interannual streamflow variability even when their simulated streamflow magnitude is biased."

Thus, we retain anomaly-based evaluation as a complementary evaluation of proportional interannual streamflow variability, while no longer presenting it as inherently more robust or superior to Q-based evaluation.

Stepping back a bit, there are parallels between the analysis done here and analyses done with global climate models. Because the climate models have biases, changes are often represented as percentage increase/decrease, rather than absolute values. The authors apply this here, but a fundamental difference between climate models and hydrology models is that the latter represents systems with considerable memory that is strongly controlled by the interaction between forcing and landscape. Evidence from the Australian Millennium Drought shows that under prolonged drying catchments can fundamentally change their runoff behaviour, and the idea of stationarity of hydrologic systems has been questioned for several decades. I therefore think that the assumption that relative changes predicted by GHMs are "safe" is at the very least unintuitive. Over the course of a future simulation model errors may become progressively larger as catchments move into consistently drier/wetter states. This stationarity of errors assumption needs to be at least explicitly stated, but evidence that this doesn't play a role (or one small enough to be negligible) would be good to add.

AC3: The authors thank the reviewer for raising this important point. We understand the reviewer's concern to be that, unlike climate variables, streamflow reflects the integrated response of climatic forcing and catchment characteristics and states, including water storage. Consequently, expressing streamflow changes in relative terms does not imply that hydrological response or model errors will remain stationary under changing climatic conditions. We agree that our previous wording did not sufficiently acknowledge this limitation.

We have therefore removed the statement in the Introduction suggesting that biases in simulated streamflow can be expected to remain similar between reference and future climatic conditions. We have also strengthened the Discussion to address the transferability of historical model performance under changing climatic conditions.

In Section 5.1, we now explicitly acknowledge that distinction raised by the reviewer:

"Historical model performance also cannot establish whether model errors will remain stationary under future climatic conditions. 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). These findings highlight the importance of considering changes in hydrological processes and catchment responses when interpreting historical model performance for climate change impact assessments."

We have also clarified in Section 5.2 that model selection for climate change impact assessments should consider hydrological processes that may become more influential under future climatic conditions, particularly vegetation responses to increasing atmospheric CO_2 concentrations and changing climate.

Our evaluation does not require an assumption of stationary model errors because it assesses model performance under observed historical variability. However, using this performance to infer the reliability of future hydrological projections involves uncertainty about the transferability of model behaviour. The present analysis cannot establish whether model errors will remain stationary or increase under future climatic conditions. We therefore do not interpret

good historical relative anomaly performance as evidence that future relative hydrological changes will necessarily be simulated accurately.

I also think the paper often conflates "improved model performance" and "higher metric values". The $Q > Q^{**}$ transformations may increase the scores found on NSE, r and gamma, but this does not mean model performance improves - it just means the scores are higher. The language used to describe the results needs to be tightened throughout the manuscript to be clear about what these increases in values actually represent, which is mainly that if you ignore specific errors you get higher scores on your performance metrics. This does not mean that the models themselves got any better.

AC4: We agree with the reviewer that an increase in a performance metric after transforming the streamflow time series should not be interpreted as an improvement in the model itself. It was not our intention to suggest that particular model errors should be disregarded in order to obtain higher performance metric values. We acknowledge that referring to these differences as "improved model performance" was inappropriate.

We have therefore revised the language throughout the manuscript, particularly in the Results and Discussion, to distinguish changes in metric values from improvement in overall model performance. We now describe the actual differences in performance metrics, such as increases in NSE or higher anomaly-based NSE values, rather than referring to improved model performance.

Section 4.2 explicitly states:

"Please note that higher NSE values for anomalies should not be interpreted as evidence of better overall model performance."

The Conclusions also clarify that agreement in streamflow magnitude and agreement in interannual variability represent complementary aspects of historical GHM performance. Thus, the revised manuscript distinguishes between changes in evaluation metrics resulting from the transformation of streamflow time series and the ability of GHMs to reproduce different characteristics of observed streamflow.

Finally I'm uncertain about the approach of doing the analysis on the annual scale. The main argument seems to be that on annual timescales any interannual differences are mainly driven by differences in climate, and inaccuracies on the shorter timescales (due to incorrect/incomplete process representations) can be ignored. I have two comments here:

1. What I'm missing, particularly in the later analysis that focuses on the driest/wettest anomalies, is some indication of how much of that pattern is already visible in the forcing data. In other words, if the driest and wettest Q years already coincide with the driest and wettest years in the forcing data, what are we really measuring at this timescale? If the model doesn't get in the way too much?

AC5.1: We thank the reviewer for raising this point. We agree that agreement between observed and simulated extreme streamflow-anomaly years may partly reflect climatic variability already represented in the forcing data. However, simulated streamflow anomalies also depend on how each GHM represents the hydrological response to that forcing.

In this study, we evaluated three GHMs, each driven by two different historical climate forcings (20CRv3–ERA5 and 20CRv3–W5E5). This allows us to compare the differences in streamflow-anomaly performance across GHMs and climate forcings. The results for extreme dry and wet anomaly years (Table 3) show that differences among GHMs are generally more pronounced than differences between the two climate forcings. These findings indicate that the choice of GHM influences the simulated streamflow-anomaly characteristics and their agreement with observations, even when the models are driven by the same forcing.

The revised Introduction clarifies the rationale for evaluating annual streamflow anomalies, recognising that observed streamflow reflects climatic variability together with catchment characteristics, hydrological conditions and human influences.

Nevertheless, our analysis does not directly compare streamflow anomalies with anomalies in the climate-forcing variables. It therefore cannot quantify how much of the agreement with observed streamflow anomalies originates from the forcing data and how much from the hydrological response represented by the models. Furthermore, both climate datasets are reanalysis-based, and the present evaluation does not independently establish how accurately they reproduce the actual climatic variability experienced by each catchment.

Section 5.1 further clarifies that disagreement between observed and simulated streamflow anomalies may arise from uncertainties in the climate forcing as well as from the representation of hydrological processes in the GHMs.

We therefore interpret the extreme-anomaly analysis as an evaluation of the combined effects of the applied climate forcing and the hydrological response represented by each GHM, rather than as an attribution of their individual contributions.

2. To the best of my understanding, the models are still run at the daily timestep. Does aggregating daily simulations that allegedly have substantial errors to annual simulations really let us ignore these errors? The underlying assumption seems to be that the errors to some extent cancel out over a year, but I doubt this holds.

AC5.2: We agree that aggregating simulated streamflow to the annual scale does not imply that errors at shorter timescales cancel out or can be ignored. Our original wording could have given this impression. The purpose of using annual streamflow is to reduce the influence of differences in the seasonal timing of human water use and management on the model–observation comparison, rather than to assume that errors in the underlying simulations disappear through aggregation. We have revised the Introduction to clarify this point:

"In contrast, aggregation to annual streamflow reduces the influence of differences in the seasonal timing of human water use and management on the model–observation comparison. This aggregation does not imply that errors at shorter timescales cancel out; errors that affect annual streamflow remain reflected in the evaluation."

We have also added literature supporting the challenges associated with representing human water use and management in GHMs (Döll et al., 2016; Pokhrel et al., 2016; Veldkamp et al., 2018).

I have a number of further comments in the PDF that I won't repeat here. These mainly concern clarity, logic and some grammatical/editorial things.

AC6: We thank the reviewer for the additional comments provided in the annotated manuscript. We have carefully considered these comments and revised the manuscript accordingly, with all revisions indicated using track changes. For completeness, we have also provided a point-by-point response to each annotated comment in Appendix A of this response document, referring to the line numbers of the originally submitted manuscript.

Overall, I think this paper needs a substantial amount of work, mainly aimed at clarifying the assumptions that underpin the method, and the addition of evidence that shows that the method actually leads to more robust results. I have particularly strong concerns about the suggestion throughout the paper that ignoring magnitude errors is a way to perform more robust model assessment. Honest assessments of our models' capabilities and the sort of information they can and cannot provide is critical. If getting "good performance" requires the amount of aggregation, de-biasing and rescaling that is shown here, perhaps more effort needs to be spent on making the models themselves better instead.

AC7: We thank the reviewer for this overall assessment. As detailed in our responses above, we have substantially revised the manuscript to clarify the assumptions, objectives, scope, and interpretation of the proposed evaluation, as well as the language used throughout.

In particular, we have removed claims that anomaly-based evaluation provides a more robust assessment than Q-based evaluation. Instead, we now present anomaly-based evaluation as a complementary approach that provides information on the ability of GHMs to reproduce observed interannual streamflow variability, including proportional deviations from the respective mean streamflow. The relevance of this information depends on the quantity of interest in the intended climate change impact assessment.

We have also clarified that removing mean differences or normalising streamflow anomalies does not improve the models themselves or their overall ability to reproduce observed streamflow. Higher anomaly-based performance metric values should therefore not be interpreted as improvements in overall model performance. Our intention is not to disregard model errors, but to distinguish different aspects of historical GHM performance and their relevance to different applications.

We agree that improving the representation of hydrological processes in GHMs remains essential. However, the purpose of the present study is to evaluate the capabilities and limitations of existing GHMs, rather than to improve their structure or parameterisation.

To reflect these clarifications consistently, we have revised the manuscript title and substantially rewritten the Abstract and Conclusions. The Introduction and Discussion have also been revised to clarify the underlying assumptions, interpretation, and limitations of anomaly-based evaluation, including the uncertainty associated with transferring historical model performance to future climatic conditions.

Appendix A:

Reviewer 1 annotated comments			
No	**Location in the submitted manuscript**	**Comment from the reviewer**	**Response from the Authors (AC)**
1	Line 74 (p. 3)	It's not particularly clear to me how the preceding studies show this part - the text only focuses on what the models don't do.	*We thank the reviewer for pointing this out. We agree that the preceding paragraph emphasised the limitations of GHMs without sufficiently highlighting the findings supporting their ability to reproduce hydrological variability. In particular, Heinicke et al. (2024) reported that GHMs can reproduce variability in daily and maximum annual discharge despite substantial biases in simulated discharge magnitude. We have therefore revised the discussion of this study to reflect both findings and clarified the concluding sentence accordingly.*
2	Lines 77-78 (p. 3)	I find it difficult to believe that Krysanova et al. is the only example that can be found. Perhaps the search is overly limited? Any assessment of GHMs' ability to reproduce historical discharge by extension tells us something about their suitability for climate change impact. See e.g. Tiwari et al (2025; https://doi.org/10.1038/s44221-025-00435-6).	We thank the reviewer for this comment and for suggesting Tiwari et al. (2025). We agree that historical GHM performance evaluations provide valuable information for assessing their suitability for climate change impact assessments. Our intention was to distinguish general historical performance evaluations from approaches explicitly developed to inform GHM selection or weighting for climate change impact assessments. We have therefore revised the introductory statement to clarify this distinction and included Tiwari et al. (2025) in the preceding literature review. Specifically, we have added: **"Tiwari et al. (2025) further demonstrated that good GHM performance in reproducing river discharge does not necessarily translate into good performance in reproducing other hydrological variables, such as terrestrial water storage."**
3	Lines 80-81 (p. 3)	What are these three metrics and why are they sensitive to bias?	We thank the reviewer for this comment. The three metrics are NSE for monthly streamflow, PBIAS, and NSE for long-term mean monthly streamflow. PBIAS directly quantifies systematic bias, while NSE is sensitive to differences between simulated and observed streamflow, including systematic bias. We have revised the sentence to explicitly identify these metrics.
4	Line 85 (p. 3)	I don't fully understand how these are different, apart from how the change is reported. E.g. a future +100% change compared to a baseline of 500 m^3/s is	We thank the reviewer for this comment. We agree that absolute and relative changes are mathematically related. However, evaluating proportional changes is particularly relevant for GHMs with biases in simulated baseline streamflow. We have clarified this distinction in the revised Introduction, emphasizing that absolute and relative changes provide complementary information.

		effectively the same as reporting a future value of 1000 m^3/s.	
5	Lines 96-97 (p. 4)	This argument seems to say that GHMs are particularly suited for assessing relative change because they have biases when it comes to absolute change. I don't think that logically follows - why couldn't a model be bad at both absolute and relative change, making it suitable for neither?	We thank the reviewer for this comment. We agree that biases in absolute streamflow do not necessarily imply good performance in reproducing relative changes. We have therefore revised the statement to clarify that these capabilities should be evaluated separately and that anomaly-based evaluation complements conventional streamflow evaluation.
6	Lines 100-102 (p. 4)	Why can this similarity be expected? The only condition under which this holds is if catchment behaviour is stationary in time, and climate change only affects flow timing, but not the relative importance of different hydrologic processes. Recent(-ish) work on the Australian Millennium Drought, for example, shows that basins can experience considerable and long-lasting changes to their runoff behaviour after prolonged drying, and this suggests that in the context of predicting long-term changes, model biases may not be stationary. See e.g. Fowler et al. (2022; https://doi.org/10.5194/hess-26-6073-2022) and references therein. More generally, the idea that hydrologic behaviour can be assumed to be stationary has been questioned for years (e.g., Milly, 2008; https://doi.org/10.1126/science.1151915).. If the authors wish to assume that catchment behaviour and thus model errors are stationary, they should provide evidence that this is reasonable to assume.	Thank you for this comment. We have addressed this comment in our response to major comment AC3.
7	Lines 111-113 (p. 4)	Similar to the above, this assumption seems key to the later method. Please provide evidence that this is a reasonable assumption to make.	This comment has been addressed in our response to major comment AC5.2

8	Line 129 (p. 5)	There is nothing inherent to split-sample tests that says that the samples must be back-to-back.	We thank the reviewer for this clarification. We agree that split-sample tests do not necessarily require consecutive periods. We have replaced "**as required by**" with "**as often used in**" to avoid this implication.
9	Line 143 (p. 5)	The abstract and most of the introduction made me believe that this paper introduces a new method for GHM evaluation, yet the stated goal is to evaluate three GHMs (and not, develop a new GHM evaluation method). Please clarify what the goal of the paper actually is.	This concern has been addressed in our responses to the major comments, particularly AC1 and AC4. We have clarified the objective and scope of the study in the revised Introduction, emphasizing that anomaly-based evaluation provides complementary information on historical GHM performance, as demonstrated through the evaluation of three GHMs.
10	Line 144 (p. 5)	Given the existence of multiple large-sample datasets with several hundreds of basins each, this number seems low. How representative can these findings be considered to be?	We thank the reviewer for this comment. As explained in Section 2.3, our selection was based on the dataset of 1,509 gauging stations co-registered to the DDM30 drainage direction map by Müller Schmied and Schiebener (2022). The requirement for continuous 30-year streamflow records, necessary for evaluating interannual variability, reduced the sample to 589 stations. This is comparable to the sample size used by Heinicke et al. (2024). The geographical distribution of the selected stations is presented in Figure 1. While the stations cover different regions worldwide, we acknowledge that they do not provide uniform global representation, and our findings should be interpreted accordingly.
11	Line 196 (p. 7)	Can it be clarified what this is?	We thank the reviewer for this comment. The term "inter-station catchment area" follows the terminology used in previous WaterGAP publications (e.g., Müller Schmied et al., 2024). It refers to the additional contributing catchment area between gauging stations along a river network. The minimum threshold of 30,000 km² was applied to reduce spatial redundancy among the selected stations.
12	Lines 200-201 (p. 7)	Selected from what? The 1509 stations previously selected by Muller-Schmied? Given that those authors apparently were looking for short records of mostly complete daily records, why is that a good starting point if the authors of this current paper are after long-term monthly records? There are many hundreds of basins out there in various data sets like CAMELS-x and derivative initiatives, as well as in the aggregate CARAVAN data if scouring individual data sets is considered too much effort.	We thank the reviewer for pointing us to additional observational datasets, such as CAMELS and Caravan. The observational dataset used in our study combines streamflow records from GRDC, GSIM and ADHI, as compiled by Müller Schmied and Schiebener (2022). A deciding factor in using this dataset was that the 1,509 gauging stations had already been carefully co-registered to the DDM30 drainage direction map, allowing consistent comparison with the GHM simulations. From these 1,509 stations, we selected 589 that satisfied our requirement for continuous 30-year streamflow records, which is essential for evaluating interannual variability.
13	Line 203 (p. 7)	This is often the case in these sort of plots, but there's a lot of white in Africa, the middle east and large parts of Asia. I think this needs to at least be acknowledged here.	We thank the reviewer for the comment. We agree that the geographical distribution of the selected stations is uneven, particularly in Africa, the Middle East and parts of Asia. We have added a sentence in Section 2.3 acknowledging this limitation.

14	Equation (2), line 220 (p. 8)	Why is this normalized by mean modeled streamflow instead of observed? The previous text grammatically at least suggests that it would be normalized by the observed mean ("Relative anomalies normalise Q_i^* by the long-term mean [of Q_i]"). I think normalizing by modeled mean may be potentially misleading - see comment on Fig. 2.	We thank the reviewer for this comment. Relative anomalies are calculated separately for observed and simulated streamflow using their respective long-term means. This normalization is intentional, as our objective is to evaluate the ability of GHMs to reproduce proportional interannual streamflow deviations independently of biases in mean streamflow. Normalizing simulated anomalies by the observed mean would retain the influence of these biases and therefore address a different evaluation objective. We agree that good performance for relative anomalies does not necessarily indicate good performance in reproducing absolute streamflow. We have clarified the normalization in Section 3.1 and revised the interpretation of anomaly-based NSE accordingly.
15	Lines 221-222 (p. 8)	As stated previously, I don't support this statement. All Eq. 2 is is Eq. 1 normalized by a local constant. Why are percentages comparable if magnitudes are not? A 10% error might be well within the measurement uncertainty if observations are low, but could equally represent a substantial amount of water if observations are high.	We thank the reviewer for this comment. We agree that expressing streamflow anomalies as percentages does not make their absolute hydrological significance or observational uncertainty comparable across catchments. Our intention was to indicate that relative anomalies allow proportional interannual variability to be compared across catchments with different mean streamflow magnitudes. We have revised the statement to clarify this distinction.
16	Line 225 (p. 8)	What makes these metrics complementary? The correlation at least appears in both NSE and KGE, so it's a bit tricky to call that one complimentary at least. Specifically for KGE, if the bias is zero by definition as the table says, KGE does not add any information over what's already contained in r and gamma. Why add it at all?	We thank the reviewer for this comment. We agree that the metrics are not mathematically independent. However, KGE serves as a composite performance metric, summarizing the combined effects of correlation and variability, while r and gamma allow these components to be examined individually. Thus, although KGE does not provide additional independent information for anomaly time series, it provides a complementary summary of model performance. The mathematical relationships between these metrics are presented in Table 1.
17	Line 228 (p. 8)	I'm missing some information about whether this uses calendar or water years. Water years seems more appropriate to account for snow effects, but introduces an additional challenge in determining when the new water year starts locally. The common (US-specific?) Oct-Sep choice does not necessarily apply everywhere globally.	We thank the reviewer for this comment. Our analysis uses calendar years (January–December). We agree that hydrological years may be more appropriate for certain catchments. However, the appropriate starting month varies geographically. We therefore used calendar years to ensure a consistent temporal definition across all analysis locations. We have clarified this in Section 3.3.

18	Lines 229-232 (p. 8)	How robust is this approach? Generally, flow percentile are derived from much higher temporal resolution data, and using just 30 data points seems rather low. Are there longer records that could support some rudimentary sensitivity analysis on this?	We thank the reviewer for this comment. The Q10 and Q90 thresholds are calculated from annual streamflow values because our objective is to identify relatively dry and wet years in the context of interannual variability, rather than extreme flow conditions at shorter timescales. We acknowledge that estimating percentiles from 30 annual observations introduces sampling uncertainty. However, the thresholds are used to identify relatively dry and wet years within the selected 30-year periods, rather than to estimate rare hydrological extremes.
19	Lines 256-257 (p. 10)	I have some reservations about the framing/phrasing of model performance in this section, as exemplified by this sentence. This sentence states that model performance "improved" but what happens is that specific errors get removed (bias) or rescaled (variability). This does not improve model performance, it just increases the metric scores by hiding (for want of a better term) what the model doesn't get right. This i a very different thing, and the text currently conflates "model is better" with "errors aren't quantified". Please adjust the text to be more precise about what these increased metric scores actually represent.	This comment has been addressed in our response to major comment AC4. We have revised the Results and Discussion to describe differences in performance metric values rather than improvement in model performance.

20	Lines 261-262 (p. 10)	I strongly disagree with this conclusion. First, what so far has been presented is a quantification of model performance in terms of NSE, r and gamma. There has been no discussion or reasoning of why these numbers "do not imply poor representation of interannual variability". By any sort of measure an NSE of -200 would indicate poor interannual variability, and if a de-biased, rescaled NSE of 0.27 somehow illustrates the opposite, then some reasoning for stating this needs to be presented. Second, I am unconvinced that the applied transformations to the observations and simulations are meaningful in quantifying the models' performance on interannual variability. Why does it matter that a model like H08 produces smaller variability than WaterGAP relative to their own simulations? If one were to look at the bottom row of figure 2, one would probably conclude that H08 is a better model than WaterGAP (higher NSE, slightly worse correlation, slightly better gamma). The top row however shows that the interannual variability of H08 is terrible, and that WaterGap's is much closer to reality. To me this suggests that the applied transformations are misleading about model performance. If these transformations (somehow) add meaningful information, it must be clarified what this is, in a way that is consistent with the models' performance without the transformation. Right now, the foundation/justification for these transformations is missing, and conclusions like the highlighted one are unsupported.	We thank the reviewer for this comment. This concern has also been addressed in our response to major comment AC4. We agree that higher NSE values for transformed streamflow series do not indicate improved overall model performance. We have revised Section 4.1 to clarify that NSE(Q) evaluates agreement in absolute annual streamflow, whereas NSE(Q^{**}) evaluates agreement in proportional interannual deviations relative to the respective mean streamflow. Consequently, differences between NSE(Q) and NSE(Q^{**}) reflect the different properties being evaluated and should not be interpreted as improvements in model performance. We have also clarified that the station example illustrates differences between H08 and WaterGAP in reproducing absolute streamflow and proportional interannual deviations, rather than establishing that either model performs better overall.

21	Line 264 (p. 10)	This number is already rather low, and the table reduces this even further by showing only P5, P50, P95 (i.e., 3 basins). Why not show the actual distributions in boxplots in a figure instead?	We thank the reviewer for this suggestion. Table 2 presents the 5th percentile (P5), median (P50), and 95th percentile (P95) of the performance metrics across all 589 analysis locations, rather than results for three individual basins. These statistics provide a concise numerical summary of the distributions. In addition, Figure 4 presents the distributions of the performance metrics as boxplots, providing a more detailed view of their variability across the analysis locations.
22	Figure 2 (p. 11)	I initially didn't understand why WaterGAP has the largest relative anomaly and had to go back to the equations, only to realize that the normalization is based on the modelled mean flow (instead of observed mean, as I assumed). I think this is potentially misleading. I expect that everyone looking at Fig 2a (top left) would agree that H08 provides the worst simulations for this basin. However, Fig 2c (bottom left) that shows the relative anomalies would suggest WaterGAP produces larger errors than H08 (relative anomaly is further away from observed anomaly then H08's). This is a consequence of normalizing by the simulated mean: WaterGAP's simulated mean is much closer to observed mean than H08's. I don't think the interpretation of the relative anomalies is at all intuitive here, and would ask the authors to comment on this.	We thank the reviewer for this observation. We agree that Figure 2c may lead to a misleading interpretation if relative anomalies are considered without reference to the original streamflow series. Normalization by the respective observed and simulated mean streamflow is intentional, as Q^{**} evaluates proportional interannual deviations relative to the mean of each time series. For example, an absolute anomaly of 100 m^3/s corresponds to a relative anomaly of 10% when the mean streamflow is 1,000 m^3/s, but 50% when the mean is 200 m^3/s. Consequently, WaterGAP may reproduce absolute annual streamflow more accurately than H08 while showing larger discrepancies in proportional anomalies. These results reflect different properties of simulated streamflow rather than differences in overall model performance. A model that reproduces the direction and proportional magnitude of observed annual streamflow anomalies demonstrates an ability to represent proportional interannual streamflow variability under historical climatic conditions, even when its simulated mean streamflow is biased. This capability is relevant when evaluating GHMs for climate change impact assessments concerned with relative hydrological changes. However, as with conventional streamflow evaluation, good historical performance does not guarantee accurate simulation of future hydrological changes, particularly where hydrological responses may change under future climatic conditions.
23	Figure 2 caption (p. 11)	Please add sub-figure labels (a, b, c, etc)	Thank you for the comment. We have now revised the figure 2 accordingly.
24	Table 2 (p. 12)	Why are P50 values in bold?	We thank the reviewer for this comment. The P50 values are presented in bold to highlight the median performance across the 589 analysis locations and to facilitate comparison between GHMs and climate forcings.

25	Line 284 (p. 12)	Again, performance is not improved - metric scores simply go up because specific errors are progressively de-emphasized. This two things are not the same. Please clarify this throughout the paper.	*Thank you for this comment. This concern has also been addressed in our response to Reviewer 1's main Comment 4 (AC4). We have revised Section 4.2 and the captions of Figures 3 and A2 to distinguish higher NSE values from improvements in model performance.* *The revised figure captions state:* **"Colours indicate NSE values for anomaly-based streamflow time series: darker colours denote higher anomaly-based NSE relative to NSE for Q, and lighter colours denote lower or equal anomaly-based NSE relative to NSE for Q. Median NSE and the fraction of basins where anomaly-based NSE is higher than NSE for Q ...."**
26	Line 291 (p. 13)	Of course it is. Is this really worth mentioning?	We agree with the reviewer that this statement is self-evident and have removed it from the revised manuscript.
27	Line 292 (p. 13)	"Figure" missing?	We thank the reviewer for pointing out this typographical error. This has been fixed in the revised manuscript.
28	Line 293 (p. 13)	Capital D	We thank the reviewer for pointing out this typographical error. This has been fixed in the revised manuscript.
29	Line 294 (p. 14)	Sentence incomplete	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^{**}."**
30	Table 3 caption (p. 15)	How does one calculate an NSE for a single value in each location? Doesn't that simplify just to a normalized bias error? Or is this NSE calculated while summing over all basins? If so, what is that normalized by?	We thank the reviewer for pointing out this ambiguity. The NSE values in Table 3 are calculated across all 589 analysis locations, rather than separately at each location. The denominator is the sum of squared deviations of observed extreme anomalies from their spatial mean. We have clarified this in the revised Table 3 caption.
31	Table 3 note (p. 15)	I find "perfect" too strong of a term to describe what's going on here. Getting the timing of the driest/wettest year in a 30-year sequence to happen in the right year should, probably, be more of a baseline expectation than something that deserves to be called "perfect", i.e. "being entirely without fault or defect" (Merriam-Webster).	We agree with the reviewer that "Perfect" is too strong, as the metric evaluates agreement in the timing of extreme anomaly years rather than perfect reproduction of the event. We have replaced "Perfect" with "Exact-year match" in Table 3 and the corresponding text.
32	Line 308 (p. 15)	It may be consistent, but the differences are also fairly small	We agree with the reviewer that, although the differences are consistent, their magnitude is relatively small for some metrics. We have revised the text in Section 4.3 to acknowledge this.

33	Lines 319-320 (p. 16)	I don't follow this. How does the higher percentage indicate that GHMs identify the same (incorrect) year as the most extreme?	We agree with the reviewer that the original statement was misleading. A Top-5 match indicates that the observed extreme anomaly year is among the five most extreme simulated years, not necessarily that it is identified as the single most extreme year. We have revised the sentence in Section 4.3 to clarify this distinction.
34	Lines 325-326 (p. 16)	This is not wrong, but it's also trivial (of course these things are different) and it's not clear at all to me how the paper actually shows this	We agree with the reviewer that the original statement was too general. We have revised the opening of Section 5.1 to clarify that our results demonstrate differences in GHM performance depending on whether annual streamflow or its anomalies are evaluated. These findings are based on the 589 selected analysis locations and their respective historical evaluation periods, and the two evaluation approaches are therefore presented as providing complementary information rather than establishing overall model superiority.
35	Lines 334-335 (p. 16)	I'm not sure I follow this. Surely the intended output of a hydrological model is to accurately simulate hydrology. How is evaluating streamflow magnitude not in line with that?	We agree with the reviewer that accurately reproducing streamflow magnitude is important for hydrological applications where absolute streamflow values are required. Our intention was not to suggest that streamflow magnitude is irrelevant to climate change impact assessments. We have revised Section 5.1 to clarify that Q-based and anomaly-based evaluations assess different aspects of historical GHM performance and provide complementary information, particularly for applications concerned with relative hydrological changes.

36	Lines 335-339 (p. 16)	This doesn't hold. The paper has applied the de-bias and rescaling transformations and described what happens to the performance scores of three GHMs, but has shown nowhere that this makes the performance assessment more robust. How does ignoring and rescaling specific errors make the assessment more comparable across regions? If the paper wants to argue that using Q^{**} is meaningful it needs to show evidence that this is the case.Also, I'm perfectly happy to accept that GHMs calibration is challenging, but I don't see why that's an excuse for simply ignoring the bias, and then also messing with the interannual variability error scaling. If put bluntly, the argument here is "if we ignore what the models are bad at, they're good" and surely that's not the case we should be making.	We thank the reviewer for this comment. As discussed in our response to major comment AC4 and the supplementary comment on Lines 256–257, we agree that removing or rescaling certain errors does not improve the model itself or demonstrate that anomaly-based evaluation is more robust. We have therefore removed the claims of greater robustness and revised Section 5.1 to clarify that Q-based and anomaly-based evaluations provide complementary information on historical GHM performance. GHMs remain important tools for assessing water resources at the global scale, despite substantial biases in simulated streamflow magnitude and the challenges associated with their calibration. We argue that poor performance in reproducing absolute streamflow magnitude should not, by itself, necessarily lead to excluding a GHM from multi-model ensembles, particularly when the intended application concerns relative hydrological changes. A model may reproduce the direction and proportional magnitude of observed interannual streamflow deviations even when its simulated mean streamflow is biased. This capability provides useful information about historical model behaviour that is not fully characterized by conventional streamflow evaluation.Our intention is therefore not to disregard model errors, but to identify different aspects of model performance and consider their relevance to the intended application. The broader implication of our study is that GHM evaluation and the interpretation of multi-model ensembles should consider both model strengths and limitations, with the relative importance of different performance characteristics depending on the quantity of interest in the climate change impact assessment.
37	Lines 356-357 (p. 17)	The paper keeps claiming this but it consistently fails to provide the reasoning to support this. Why is performance more relevant if we ignore magnitudes?	We agree with the reviewer that ignoring streamflow magnitude does not inherently make a performance assessment more relevant. As discussed in our response to major comment AC4, our intention is not to disregard magnitude errors or suggest that anomaly-based evaluation is universally preferable. We have revised Section 5.2 to clarify that the relevance of different performance characteristics depends on the quantity of interest. For climate change impact assessments concerned with relative hydrological changes, a model's ability to reproduce proportional interannual streamflow variability provides additional historical information that may not be adequately reflected by Q-based evaluation alone. However, this does not imply that magnitude errors are unimportant or that good historical anomaly-based performance guarantees accurate future hydrological changes.
38	Lines 357-359 (p. 17)	There is a ton of literature on bias-correction methods in climate models. If this paper wants to claim that those aren't good enough, please provide a reason for saying so	We agree with the reviewer that the original statement was too general and could be interpreted as questioning the effectiveness of existing climate-model bias-correction methods without sufficient supporting evidence. Evaluating these methods is beyond the scope of our study. We have therefore removed the statement from Section 5.2.

39	Lines 364-366 (p. 17)	This doesn't follow. Streamflow anomalies are calculated from streamflow, which requires routing to the gauge.	We agree with the reviewer that streamflow at a gauging station requires routing. Our intention was to highlight a potential extension of anomaly-based evaluation to GHMs without routed streamflow or with different river-routing networks. In such cases, annual basin-aggregated runoff could be considered as an alternative evaluation variable. We have revised Section 5.2 to clarify that this is a potential application requiring further investigation, rather than an approach demonstrated in the present study.
40	Line 367 (p. 17)	So what if it is? "it's a lot of work" is a pretty poor argument, particularly when standalone routing models exist	We thank the reviewer for this comment. When GHMs use different drainage direction maps, establishing consistent correspondence between model grid cells and observed gauging stations can require model-specific mapping, which presents a practical challenge in global multi-model evaluations. We acknowledge that standalone routing models offer an alternative approach. Our intention was to highlight the potential applicability of anomaly-based evaluation to a wider range of GHMs, rather than to suggest that model-specific mapping cannot be undertaken. We have revised the final paragraph of Section 5.2 to clarify this point.
41	Lines 367-370 (p. 17)	Why couldn't this be done with the raw runoff as well?	We agree that basin-aggregated runoff can also be evaluated without applying the anomaly transformation. Our intention was not to suggest that this possibility is unique to anomaly-based evaluation, but to highlight its applicability to models without routed streamflow or with different river-routing networks. We have revised the final paragraph of Section 5.2 to clarify this distinction and acknowledge that the proposed extension requires further investigation.
42	Lines 384-385 (p. 18)	There has been no attribution of performance differences to process representation whatsoever in this paper	We agree with the reviewer that our analysis does not attribute differences in GHM performance to specific hydrological process representations. Our intention was to highlight that the three GHMs share a common routing framework, which reduces differences associated with routing configurations in the comparison. In the revised manuscript, we have combined the former Sections 5.1 and 5.3 into Section 5.1 and removed the unsupported attribution. We have also clarified in the revised Section 5.1 that the causes of performance differences among the GHMs were not investigated in this study.
43	Lines 395-398 (p. 18)	This does not logically follow. That the highest performance is found for relative anomalies is expected, because important errors are progressively ignored. This does not mean that "conventional evaluation may underestimate the ability of GHMs to represent hydrologic responses to interannual climate variability and thus climate change". If anything, it shows that looking at relative anomalies may hide important parts of model error.	We agree with the reviewer that higher NSE values for relative anomalies do not imply better overall model performance. We have removed the statement that conventional evaluation may underestimate the ability of GHMs to represent hydrological responses to climate variability and climate change. Following the reframing of our interpretation (see AC4), we have substantially revised the Conclusions to emphasize that Q-based and anomaly-based evaluations provide complementary information on historical GHM performance.

44	Figure A1 (annotated basin) (p. 20)	I'm assuming this indicates the single basin discussed in section 4.1. Please indicate this somewhere in this figure's caption.	*Thank you. We have added the following sentence to the caption of Figure A1:* **"The annotated basin corresponds to a randomly selected station in the Macquarie River catchment, Australia (lat: -31.75, lon: 147.75; station ID 65440), for which time series are shown in Fig. 2."**
45	Figure A1 legend (p. 20)	Having gone through, I'm nor sure calling this "raw time series" is accurate. The raw time series is at a sub-annual resolution. This should be called "NSE of annual time series" here, and in the rest of the manuscript.	*Thank you. We have replaced "raw time series" with* ***"annual streamflow time series"*** *in the relevant figure labels, captions and the rest of the manuscript.*
46	Figure A2 legend (p. 21)	I find these colors incredibly hard to interpret because there is no intuitive progression (as far as I can see). Why not simply use two shades (e.g. various intensities of yellow for improvements, various shades of blue for degradations)?	Thank you for the suggestion. We prefer to retain the existing colour scheme because it communicates two aspects of NSE simultaneously. The colours indicate the NSE values, ranging from red for negative NSE (NSE < 0) to blue for high positive NSE, while colour intensity distinguishes locations where anomaly-based NSE is higher (darker) or lower (lighter) than NSE for Q. We consider this representation more informative for interpreting the spatial patterns than using two separate colour scales.