https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

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

Thedini Asali Peiris[1] and Petra Döll[1,2]

[1]Institute of Physical Geography, Goethe University Frankfurt, Frankfurt am Main, Germany
[2]Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt am Main, Germany

Correspondence: Thedini Asali Peiris, Ardhige, ThediniAsaliPeiris@em.uni-frankfurt.de

Abstract.

Global hydrological models (GHMs) are widely used in climate change impact assessments to estimate future changes in freshwater availability, droughts, and floods. Their suitability is often evaluated by comparing simulated daily or monthly streamflow with historical observations. However, since climate change impacts are usually expressed as changes relative to a reference period, biases in flow magnitude do not necessarily prevent GHMs from accurately representing hydrological change.

Here, we propose an alternative method for evaluating the suitability of GHMs for quantifying the impact of climate change on water resources. This method is based on the assumption that GHMs that are more effective in capturing the timing and magnitude of the annual streamflow anomaly during a historical period (which is mainly caused by climate variability) are likely to produce more plausible hydrological responses to climate change. Therefore, it compares simulated and observed absolute and relative streamflow anomalies instead of streamflow magnitudes themselves; it considers annually aggregated streamflow, which, compared with daily or monthly streamflow, is much less impacted by reservoir operations and human water use — two important drivers of streamflow that are difficult to model using GHMs.

In this study, we test this new evaluation method as an example for three GHMs (H08, MIROC-INTEG-LAND, and WaterGAP2.2e), forced by two climate datasets (20CRv3–ERA5 and 20CRv3–W5E5). We use streamflow observations from 589 gauging stations worldwide to evaluate the method. Magnitude-based evaluation shows large performance differences among the GHMs, with negative median Nash–Sutcliffe Efficiency (NSE) values for the two uncalibrated ones due to large biases. In contrast, all GHMs achieve positive NSE for annual streamflow anomalies. Regarding relative anomalies, the land surface model MIROC-INTEG-LAND performs worst with an NSE of about 0.3, whereas the two water resources models perform very similarly and achieve median NSE values above 0.5. Both models show very similar correlation, but WaterGAP simulates the standard deviation of the absolute and relative annual streamflow anomaly better than H08. The differences in performance between the two climate forcings are smaller than the differences among the models. In terms of the ability of GHMs to simulate the years in which extreme wet and dry anomalies occur, there is exact agreement between the observed and simulated extreme years at only 7–13% of analysis locations. Meanwhile, the observed extreme year is identified among the five most extreme simulated years at 40–59% of analysis locations. MIROC-INTEG-LAND also shows the lowest performance regard-

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

ing extreme anomalies. GHM evaluation based on annual streamflow anomalies, particularly relative anomalies, is suitable for assessing how GHMs translate annual climate variability and, to a certain extent, climate change into hydrological changes.

However, the proposed GHM evaluation method does not consider the vegetation response to increased atmospheric CO_2 concentrations and climatic changes that may strongly affect the hydrological response to climate change. Therefore, multi-model ensemble assessments of hydrological impacts of climate change should include even lower-performing GHMs, provided that these GHMs take the vegetation response into account.

1 Introduction

Global hydrological models (GHMs) are essential tools for assessing freshwater availability for humans and ecosystems as well as drought and flood hazards (Fung et al., 2011; Hagemann et al., 2013; Haddeland et al., 2014; Döll et al., 2018). Their ability to simulate large-scale hydrological processes, in particular streamflow and anthropogenic alterations of natural water flows and storages, makes them indispensable tools for global water resources assessments, international policy analyses and climate change impact studies (Veldkamp et al., 2018; Telteu et al., 2021; Scanlon et al., 2018; Zhao et al., 2025). Therefore, the reliability of GHM-simulated streamflow directly affects the robustness of water-related decision-making.

Numerous studies have reported substantial discrepancies between GHM-simulated and observed streamflow, particularly in basins strongly affected by human water use and artificial reservoirs (Scanlon et al., 2018; Krysanova et al., 2018). These discrepancies reflect persistent challenges related to model structure, model parameterization, scale mismatch, and the representation of human water management. Improving the reliability and interpretability of GHM outputs remains a central objective in large-scale hydrological modeling research.

In the context of climate change impact assessments, hydrological projections are commonly derived from multi-model ensembles, in which multiple hydrological models are driven by bias-adjusted climate model outputs to account for structural and forcing uncertainties (Döll et al., 2016). While global climate models (GCMs) are generally regarded as equally plausible representations of the climate system, this is not necessarily the case for GHMs (Krysanova et al., 2018). Krysanova et al. (2018) suggested excluding hydrological models that are unable to reproduce historical observations satisfactorily from multi-model climate change impact assessments. They showed that projected streamflow and other hydrological variables differ substantially between well-calibrated basin-scale models and uncalibrated or only roughly calibrated GHMs, concluding that projections from well-calibrated basin-scale models are generally more plausible than projections from GHMs. However, good performance under historical conditions does not necessarily imply suitability for climate change impact studies. Models may reproduce historical streamflow while failing to represent key processes relevant under future conditions, such as the impacts of increasing atmospheric CO_2 concentrations and climate-driven vegetation changes on evapotranspiration (Peiris and Döll, 2023). Even if models represent these processes, the simulated impacts of these processes on water resources vary between

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

models (e.g., Reinecke et al. (2021)). More generally, calibration under stationary historical conditions does not guarantee robust performance when models are applied outside their calibration domain (Seibert, 2003). In summary, the capability of GHMs for assessing climate change impacts cannot be judged comprehensively, in particular because 1) past climatic changes are still small as compared to potential future changes and 2) historic streamflow has been affected by various non-climatic changes such as land and water use changes and the operation of artificial reservoirs, which are challenging to quantify. Still, a partial evaluation may be useful. However, methods for such evaluations should be improved compared to the current model evaluation approaches (Hattermann et al., 2017; Krysanova et al., 2018, 2020).

Previous evaluations of GHMs have primarily relied on comparisons of simulated and observed streamflow across diverse hydro-climatic regions by comparing simulated and observed streamflow, mostly monthly time series, using metrics such as the Nash–Sutcliffe efficiency (NSE), Kling–Gupta efficiency (KGE), percent bias (PBIAS), and correlation coefficients (Gudmundsson et al., 2011, 2012a, b; Hattermann et al., 2017; Krysanova et al., 2018, 2020). These evaluations reveal substantial variability in model skill across catchment types, flow regimes, and temporal scales. Heinicke et al. (2024) showed that seven out of nine ISIMIP3a GHMs overestimated both mean and extreme streamflow at most of the 644 globally distributed streamflow gauging stations considered in this model performance study, and consistently overestimated both mean and extreme high and low streamflows, a bias also reported in earlier intercomparison studies (Hattermann et al., 2017; Zaherpour et al., 2018). With the median correlation of daily streamflow ranging between 0.17 and 0.68, the study showed that GHMs are not suited for predicting daily streamflow. Veldkamp et al. (2018) determined that including human impact parametrisations, in particular artificial reservoirs and human water use, substantially improves the simulation of monthly streamflow and high-flow extremes. Collectively, these studies indicate that while GHMs capture large-scale hydrological patterns and variability, output uncertainties remain large, particularly in semi-arid and snow-dominated regions, as well as regions where streamflow is strongly altered by human activities.

To our knowledge, studies that aim to evaluate specifically the suitability of GHMs for climate change impact studies are limited. Krysanova et al. (2020) aimed at providing an evaluation approach for GHMs that could be used to determine weights for computing multi-model ensemble means, so that the output of better-performing GHMs would count more when producing multi-model ensemble mean changes. Three out of the four performance metrics they applied are strongly sensitive to the bias of the long-term average streamflow. Consequently, comparing six GHMs across 57 large catchments, they found that the only calibrated model, WaterGAP2, performed best in more than half of the basins, even though it is only calibrated against long-term mean streamflow. They then compared weighted and non-weighted ensemble means of absolute streamflow changes, i.e., the difference between the streamflow during a future period and the reference period.

Climate change impact studies typically focus on changes relative to a reference period rather than absolute future values (e.g., average streamflow at a certain location will be 1000 $\mathrm{m^3/s}$ in 2041–2070). Relative changes, such as percentage changes in precipitation or renewable water resources, are widely reported in climate change impact assessments, including IPCC studies, because they provide a scale-independent measure that facilitates comparison across regions with different baseline conditions. For example, a 5% decrease in precipitation may correspond to a reduction of only about 3 $\mathrm{mm\,yr^{-1}}$ in a dry region but around 60 $\mathrm{mm\,yr^{-1}}$ in a humid region. While the absolute reduction may be larger in humid regions, the relative change en-

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

ables comparison of proportional shifts in hydro-climatic conditions across regions with very different water availability. At the same time, projections of absolute hydrological conditions are subject to substantial uncertainty arising from biases in climate model inputs, even after bias adjustment (Döll et al., 2016). Absolute and relative changes therefore provide complementary information: absolute changes are essential for applications that depend on accurate streamflow magnitude, such as flood risk assessment and water management, whereas relative changes provide a more consistent basis for comparing climate-driven changes across regions. In this context, GHMs, which are designed for large-scale assessments and are known to exhibit biases in streamflow magnitude, are particularly suited for analysing patterns of relative change rather than precise local conditions. This highlights the need for evaluation approaches that emphasise the ability of GHMs to represent climate-driven variability.

Therefore, to evaluate the suitability of impact models such as GHMs in climate impact studies, the ability of impact models to reliably simulate relative changes should be evaluated together with their representation of variability. Biases in simulated long-term mean streamflow can be expected to be rather similar under reference climatic conditions, for which the GHMs are evaluated against streamflow observations, and different climatic conditions in a future time period. Therefore, GHM performance metrics that are strongly affected by bias may be less suitable when the objective is to assess the ability of models to project climate change impacts.

As climate change leads to changes in the seasonality of streamflow, GHMs should be able to represent the seasonality of streamflow correctly, which can be tested by computing the goodness-of-fit of monthly time series or of the 12 calendar month means (Krysanova et al., 2020). However, in many regions, seasonal streamflow dynamics are strongly shaped by reservoir operations and human water use, which are either not fully represented in GHMs or with large uncertainties (Veldkamp et al., 2018). Consequently, model performance based on monthly streamflow may be dominated by factors and actors unrelated to climate-driven hydrological responses. In contrast, aggregation to annual streamflow strongly reduces the influence of short-term human interventions and seasonal mismatches on simulated streamflow. Year-to-year differences in annual streamflow are expected to be less affected by the difficult-to-simulate reservoir operations and the seasonality of irrigation and to be mostly driven by climatic variations. Therefore, a good representation of year-to-year differences of annual streamflow can serve as one indication of the ability of a model to translate not only year-to-year climate variability but also long-term climatic changes into changes in streamflow and water resources. Based on the study of Hattermann et al. (2017), we expect that hydrological models vary significantly in their ability to correctly translate climate variability into streamflow variability; their Fig. 2 shows that 18 hydrological models, whether calibrated basin models or GHMs, translate the same relative anomaly of annual river basin rainfall into very different relative anomalies of annual streamflow.

To assess model robustness under changing climatic conditions, Krysanova et al. (2018) proposed evaluation frameworks such as differential and generalized split-sample tests, sliding-window tests, and multi-period, multi-criteria conditioning approaches, including the FARM test (Krysanova et al., 2018; Euser et al., 2013). These methods are conceptually well-suited for impact-oriented evaluation under ideal circumstances, where long, continuous observational records are available and non-climatic impacts on streamflow are either negligible or can be represented well. Basin-scale benchmarking studies have successfully applied such approaches by evaluating calibrated large-scale hydrological models alongside watershed-scale models to assess historical performance and future responses under warming scenarios (Shrestha et al., 2025). However, their applica-

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

tion at the global scale is challenging. Observational streamflow records are often limited in length and continuity, constraining the applicability of multi-period and regime-based evaluation frameworks. The temporal development of human alterations can only be represented with very high uncertainty. Moreover, climate change signals typically emerge over long time horizons, whereas comparisons between adjacent or relatively short historical periods, as required by split-sample or sliding-window tests, may not adequately represent model behavior under long-term climatic change.

Against this background, this paper proposes to evaluate GHMs in terms of their ability to project the impact of climate change on streamflow, and thus on water resources, by determining their ability to translate climate variability into streamflow variability on an annual timescale. In other words, the paper will test the ability of GHMs to reproduce the change in observed annual streamflow from one year to the next, i.e., the anomaly of annual streamflow with respect to its mean. We assume that GHMs that are more effective in capturing the timing and magnitude of the annual streamflow anomaly during a historical period, which is mainly caused by climate variability, will likely produce more plausible hydrological responses under future climate conditions, as was also stated by Hattermann et al. (2017). Such an evaluation approach is more suitable for comparing and potentially weighting GHMs within multi-model ensembles used for climate change impact assessments than previous GHM performance evaluations. Importantly, this evaluation approach does not take into account the different capabilities of GHMs to simulate the impact of increasing atmospheric CO_2 on actual evapotranspiration, which strongly impacts future streamflow (Gerten et al., 2014; Peiris and Döll, 2023) and other hydrological impacts of climate change (Reinecke et al., 2021; Wada et al., 2013).

In this study, we evaluate three GHMs (H08, MIROC-INTEG-LAND, and WaterGAP2.2e) against observed streamflow records at 589 gauging stations worldwide. To account for the influence of climate forcing on model performance, each GHM is driven by two state-of-the-art historic climate datasets (20CRv3–ERA5 and 20CRv3–W5E5). The evaluation is conducted over station specific 30 year periods to ensure a robust assessment of annual streamflow anomalies.

2 Models and observations

2.1 Global hydrological models (GHMs)

The GHM evaluation was performed using monthly streamflow outputs from three GHMs that participated in the ISIMIP3a model intercomparison round, WaterGAP2.2e, MIROC-INTEG-LAND, and H08 (experiment specifiers obsclim/histsoc, https://protocol.isimip.org/#/ISIMIP3a/water_global). These three GHMs are implemented on a 0.5° × 0.5° grid using the DDM30 drainage direction map for streamflow routing between grid cells, ensuring consistent spatial representation of river networks and comparability of computed streamflow between the models.

The Water and Global Assessment and Prognosis model version 2.2e (WaterGAP2.2e, hereafter 'WaterGAP') was developed to quantify renewable water resources and human water use for all continental land areas except Antarctica and Greenland (Müller Schmied et al., 2024). It computes net water abstractions from rivers, lakes, artificial reservoirs, and groundwater, which then affect water storages and flows, including streamflow. PET is calculated using the Priestley–Taylor method, which relates PET to net radiation and air temperature through a fixed coefficient (α). To minimize model bias, WaterGAP is calibrated in a

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

basin-specific manner for both climate forcings used in this study at 1,509 gauged basins against long-term average streamflow by adjusting 1-3 model parameters (Müller Schmied et al., 2024).

Similar to WaterGAP, the H08 model was developed to simulate water resources and human water use; It is composed of six submodules that represent land surface hydrology, river routing, reservoir operation, crop growth, water abstraction, and environmental flow requirements (Hanasaki et al., 2008a, b). Streamflow is altered by artificial reservoirs and water abstractions. PET is computed with a bulk approach based on saturation deficit and wind velocity, while also modelling the surface energy balance. Sets of four model parameters are optimized for eleven Köppen-Geiger climate zones against monthly time series of streamflow at 777 stations worldwide (Yoshida et al., 2022).

The MIROC Integrated Land Surface model (MIROC-INTEG-LAND v1, hereafter 'MIROC') is a land surface model designed to represent interactions among hydrological, ecological, and human systems (Yokohata et al., 2020).MIROC-INTEG-LAND combines several submodules, including the land surface and human water management component HiGWMAT, the global crop growth model PRYSBI2, the terrestrial ecosystem model VISIT, and the land-use decision-making model TeLMO. Streamflow is generated within HiGWMAT, which computes runoff and soil moisture dynamics based on the MATSIRO land surface model and takes into account reservoir operation and human water use, applying the H08 approach. MATSIRO estimates the exchange of energy, water vapor, and momentum between the land surface and the atmosphere on a physical basis. Different from WaterGAP and H08, which compute the soil water balance in a single layer, MIROC-INTEG-LAND has a large number of soil layers (Telteu et al., 2021), and simulates the response of vegetation and thus actual evapotranspiration to increased atmospheric CO_2 concentrations and climatic changes (Yokohata et al., 2020). MIROC-INTEG-LAND is not calibrated at the basin level (Yokohata et al., 2020).

2.2 Climate forcing

All GHM simulations follow the ISIMIP3a setup "obsclim" and "histsoc", in which climate forcing represents observed climate and greenhouse gas concentrations, while human forcing accounts for historical water use and reservoir operations (Frieler et al., 2024). Each model is driven by two reanalysis-based climate datasets, Twentieth Century Reanalysis version 3 (20CRv3; Compo et al., 2011; Slivinski et al., 2019): 20CRv3–ERA5 and 20CRv3–W5E5. Both datasets combine 20CRv3 for the period prior to 1979 with a second dataset for the period from 1979 onwards, when ERA5 becomes available. To ensure temporal consistency, the pre-1979 data are bias-adjusted to the post-1979 dataset using the ISIMIP bias adjustment framework (Lange, 2019; Mengel et al., 2021).

The two forcings differ in the treatment of the post-1979 period. In 20CRv3–ERA5, ERA5 reanalysis data are used without bias adjustment, whereas 20CRv3–W5E5 employs W5E5, a bias-adjusted version of ERA5 (Hersbach et al., 2020; Cucchi et al., 2020; Lange et al., 2021). Using both forcings allows assessing the sensitivity of model performance to differences in climate forcing, particularly related to systematic biases in the input data. For all three GHMs, monthly time series are available for the period 1901–2019.

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

2.3 Observed streamflow data

This study utilizes observed streamflow datasets from the Global Runoff Data Center (GRDC), the Global Streamflow Indices and Metadata Archive (GSIM), and the African Database of Hydrometric Indices (ADHI). Combining and homogenising these data sets, Müller Schmied and Schiebener (2022) identified 1,509 streamflow gauging station that fit the following criteria: an upstream area of at least 9,000 km^2, a time series of at least four complete calendar years (not necessarily consecutive) with no more than two missing days per month, and a minimum inter-station catchment area of 30,000 km^2. The stations were carefully co-registered to the DDM30 drainage direction map using reported upstream area, geographic location, and additional attributes such as river and station names (Müller Schmied and Schiebener, 2022). This comprehensive dataset was later used by Müller Schmied et al. (2024) for calibrating the WaterGAP2.2e model used in this study.

We selected stations with a continuous 30-year time series of annual streamflow data, essential for analysing interannual variability. To improve spatial representation, particularly in the USA, Europe, and Australia regions, we allowed one month of missing data over the total 30-year analysis period. The missing month was subsequently filled using the average of the available months within the same year. Applying this criterion, we identified 589 stations (Figure 1). For each station, the selected 30-year period corresponds to the most recent span of continuous monthly data available within its record, covering the period from 1901 to 2019.

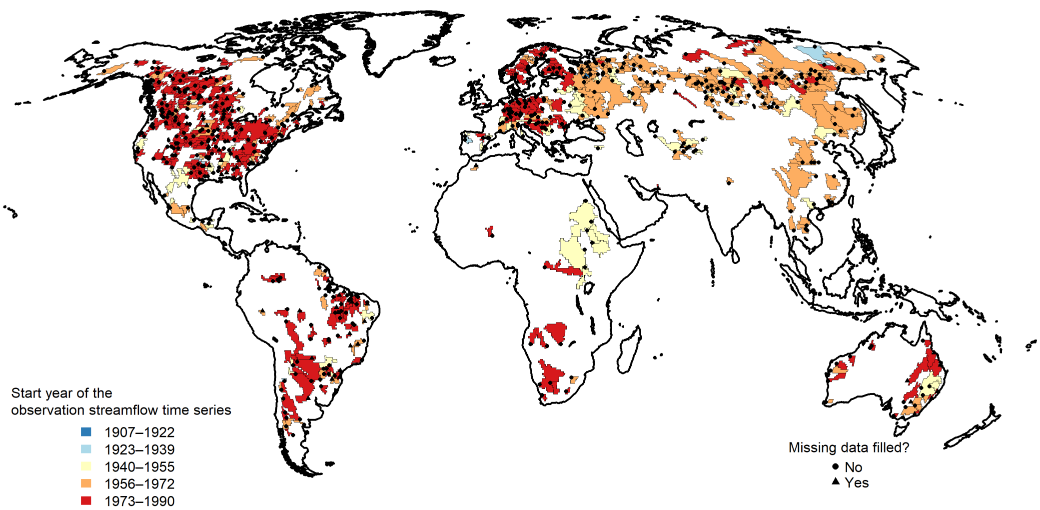

Figure 1. Spatial distribution and temporal coverage of the selected GHM evaluation locations used in this study. Each point represents a streamflow gauging station, with the symbol shape indicating whether missing data were filled (circle = no, triangle = yes). The color of the basin corresponds to the start year of the analysis period. The stations and time periods were selected based on data completeness, allowing at most one missing month for the latest possible continuous 30-year period.

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

3 Model evaluation methods

3.1 Streamflow anomaly time series

In addition to raw annual streamflow (Q), this study evaluates streamflow anomaly time series to assess how GHMs reproduce interannual variability independently of long-term magnitude biases. Because climate change impact assessments interpret hydrological change relative to baseline conditions, two anomaly representations were analysed: absolute anomalies and relative anomalies (hereafter 'configurations'), both derived from annual streamflow. While relative anomalies might be better suited to climate change impact analysis due to the lack of bias-free climate simulations by GCMs, some impact studies, such as Hattermann et al. (2017), prefer to analyse absolute anomalies.

Absolute anomalies quantify deviations from the long-term mean annual streamflow (Q_m). For year i, absolute anomaly streamflow (Q_i^*) is computed as:

$$Q_i^* = Q_i - Q_m \tag{1}$$

This transformation removes systematic offsets between simulated and observed streamflow while preserving the magnitude of year-to-year deviations.

Relative anomalies normalise Q_i^* by the long-term mean. For year i, relative anomaly streamflow (Q_i^{**}) is computed as:

$$Q_i^{**} = \frac{Q_i - Q_m}{Q_m} \tag{2}$$

Q_i^{**} expresses deviations as fractions of the mean annual streamflow, also facilitating comparison between basins with different flow regimes.

3.2 Model evaluation metrics

Annual time series of streamflow and streamflow anomalies of the six climate-GHM combinations were evaluated by four complementary performance metrics: Nash–Sutcliffe efficiency (NSE), Kling Gupta Efficiency (KGE), Pearson correlation coefficient (r), and the variability ratio (γ). Their formulation and interpretation are summarised in Table 1.

3.3 Simulation of the wet and dry years

To assess the capability of three GHMs to reproduce streamflow conditions in extremely dry and wet years, relative anomalies Q^{**} were analysed. Extreme anomaly years were identified from each 30-year time series. The dry anomaly metric (Q_{90}^{**}), i.e., the relative streamflow anomaly that is exceeded in nine out of ten years, was defined as the 3$^{\text{rd}}$ smallest anomaly value of the 30-year time series, representing a strong negative (dry) anomaly. The wet anomaly metric (Q_{10}^{**}), i.e., the streamflow that is exceeded in only one out of ten years, was defined as the 3$^{\text{rd}}$ largest anomaly value.

For each analysis location, the occurrence year of the observed Q_{90}^{**} and Q_{10}^{**} was recorded. Model performance in reproducing extreme anomaly occurrence was evaluated using two event matching metrics. A perfect match was recorded when the observed and simulated extreme anomaly occurred in the same year. In addition, an event-matching metric was calculated

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

Table 1. Performance metrics used to evaluate annual streamflow and annual streamflow anomalies. Var stands for Q, Q^*, or Q_i^{**}

.

Metric	Definition	Range / Ideal	Interpretation for anomaly time series
NSE	$1 - \frac{\sum(Var_i^{\mathrm{sim}} - Var_i^{\mathrm{obs}})^2}{\sum(Var_i^{\mathrm{obs}} - \overline{Var}^{\mathrm{obs}})^2}$	$[-\infty, 1]$ / 1	Measures overall agreement between simulated and observed time series (Nash and Sutcliffe, 1970). NSE = 0 indicates that the model's explanatory power is equivalent to the mean of the observations, while NSE < 0 indicates that the model performs worse than this benchmark.
r	$\frac{\sum(Var_i^{\mathrm{sim}} - \overline{Var}^{\mathrm{sim}})(Var_i^{\mathrm{obs}} - \overline{Var}^{\mathrm{obs}})}{\sqrt{\sum(Var_i^{\mathrm{sim}} - \overline{Var}^{\mathrm{sim}})^2}\sqrt{\sum(Var_i^{\mathrm{obs}} - \overline{Var}^{\mathrm{obs}})^2}}$	$[-1, 1]$ / 1	Measures the temporal agreement in the year-to-year variations between simulated and observed time series. Indicates how well models reproduce the timing of wet and dry years.
γ	$\frac{\sigma^{\mathrm{sim}}}{\sigma^{\mathrm{obs}}} = \sqrt{\frac{\sum(Var_i^{\mathrm{sim}} - \overline{Var}^{\mathrm{sim}})^2}{\sum(Var_i^{\mathrm{obs}} - \overline{Var}^{\mathrm{obs}})^2}}$	$[0, \infty]$ / 1	Measures the bias of the simulated variability as compared to observations. $\gamma > 1$ indicates overestimation and $\gamma < 1$ underestimation.
KGE	for $Q:\ 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$ for $Q^*, Q^{**}:\ 1 - \sqrt{(r-1)^2 + (\gamma-1)^2}$	$[-\infty, 1]$ / 1	Combined performance metric integrating correlation, bias, and variability (Gupta et al., 2009).KGE values greater than -0.41 indicate that a model improves upon the mean flow benchmark (Knoben et al., 2019). For anomaly time series (Q^*, Q^{**}), the mean is zero by construction, and the bias term (β) is therefore not defined.

Note: Var_i^{sim} and Var_i^{obs} denote simulated and observed values in year i; $\overline{Var}$ denotes the long-term mean; σ denotes standard deviation; $\beta = \frac{\overline{Var}^{\mathrm{sim}}}{\overline{Var}^{\mathrm{obs}}}$; $\sum$ denotes summation over the number of years (here 30 years).

based on whether the observed extreme year was among the five most extreme simulated years. For dry anomalies, the observed Q_{90}^{**} year was considered successfully reproduced if it was among the five driest years in the simulated relative anomaly time series. Similarly, for wet anomalies, the observed Q_{10}^{**} year was considered successfully reproduced if it was among the five

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

wettest simulated years. The percentage of analysis locations fulfilling these criteria was then calculated for each ensemble member. In addition to the event matching metrics, model performance was evaluated by comparing simulated and observed Q_{90}^{**} as well as Q_{10}^{**} values at all 589 analysis locations using NSE.

4 Results

4.1 Illustrative example for the performance of annual streamflow of GHMs as compared to that of absolute and relative anomalies of annual streamflow

The relationship between simulated and observed annual streamflow and its anomalies is illustrated for a station (Figure 2) in the Macquarie River catchment, a sub-basin of the Murray–Darling Basin (28,610 km^2). The raw streamflow time series (Q; upper panels) show that H08 and MIROC-INTEG-LAND substantially overestimate streamflow under both climate forcings, while WaterGAP shows minimal bias, reflecting its calibration to long-term mean streamflow for this analysis location (see Section 2). Correspondingly, NSE values are closer to one for WaterGAP (NSE $\approx$ 0.00 – 0.14), whereas H08 and MIROC-INTEG-LAND show strong negative values due to the large overestimation of mean streamflow.

When evaluated using absolute anomalies (Q^*; middle panels), NSE improves for all models, although H08 and MIROC-INTEG-LAND remain negative. H08 continues to overestimate variability (e.g., around 1990), while WaterGAP shows only minor changes relative to the Q configuration. Correlation (r) and the variability ratio (γ) remain unchanged between Q and Q^*, reflecting that subtracting the mean removes magnitude bias but does not change the temporal structure or the variability of the time series.

Under the relative anomaly configuration (Q^{**}; lower panels), performance of H08 and MIROC-INTEG-LAND improved significantly and, under 20CRv3-ERA, exceeded that of WaterGAP, whose performance remained relatively stable. In contrast to Q^*, relative anomalies modify the scaling of variability, resulting in changes in γ, while r remains unchanged. WaterGAP shows comparatively small changes across all the configurations, whereas H08 and MIROC-INTEG-LAND show a stronger response to anomaly transformation.

Differences between the two climate forcings are small. Overall, this station example illustrates that large discrepancies in mean streamflow do not necessarily imply poor representation of interannual variability.

4.2 Global-scale performance of the GHM annual streamflow anomaly time series

Table 2 summarizes the GHM performance at all 589 analysis locations. For each climate-GHM combination, the 5th percentile (P5), median (P50), and 95th percentile (P95) of the NSE and KGE values are listed.

Differences between the two climate forcings are small. For both climate forcings, median NSE values of raw streamflow (Q) are negative for H08 and MIROC-INTEG-LAND (e.g., -6.07 and -1.04 under 20CRv3–ERA5), whereas WaterGAP shows positive median NSE values (0.43–0.44). Considering streamflow anomalies, median NSE values become positive for all climate-GHM combinations and increase further for relative anomalies (Q^{**}), reaching 0.51 for H08 and 0.29-0.30 for

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

Figure 2. Annual streamflow time series at station 65440 in the Macquarie River catchment, Australia (lat: -31.75, lon: 147.75). Raw time series (top row), absolute anomaly (middle row) and relative anomaly (bottom row), forced by two climate datasets: 20CRv3-ERA5 (left column) and 20CRv3-W5E5 (right column). Observed streamflow (Obs) is shown with dotted black lines, while simulated streamflow from H08, MIROC-INTEG-LAND, and WaterGAP is shown with solid colored lines. Below each panel, the metrics NSE, r, and γ as well as the mean streamflow values are listed for the three GHMs.

MIROC-INTEG-LAND. In contrast, WaterGAP shows smaller improvements to NSE values for Q^{**} 0.52–0.56), reflecting its comparatively small bias in Q. KGE is not as sensitive to bias as NSE, as reflected by positive median KGE values of Q even for H08 and MIROC-INTEG-LAND. Still, median KGE values are also highest in the case of Q^{**} as compared to Q and Q^{*}, with the lowest values again for MIROC-INTEG-LAND (0.44-0.45), and similar values for HO8 (0.60-0.62) and WaterGAP (0.67-0.69).

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

Table 2. Summary statistics (P5, P50, P95) of NSE and KGE for annual streamflow (Q), absolute anomaly (Q^*), and relative anomaly (Q^{**}) for each ensemble member under 20CRv3-ERA5 and 20CRv3-W5E5 climate forcings.

GHM	Config	NSE						KGE					
		20CRv3-ERA5			20CRv3-W5E5			20CRv3-ERA5			20CRv3-W5E5		
		P5	P50	P95	P5	P50	P95	P5	P50	P95	P5	P50	P95
H08	Q	-146.83	**-6.07**	0.55	-75.55	**-2.25**	0.70	-4.42	**0.16**	0.77	-3.70	**0.37**	0.82
	Q*	-8.77	**0.24**	0.82	-8.20	**0.40**	0.86	-1.52	**0.53**	0.87	-1.41	**0.61**	0.87
	Q**	-0.72	**0.51**	0.84	-1.56	**0.51**	0.88	-0.01	**0.60**	0.88	-0.08	**0.62**	0.88
MIROC-INTEG-LAND	Q	-82.68	**-1.04**	0.49	-52.43	**-1.67**	0.38	-3.39	**0.28**	0.68	-3.17	**0.25**	0.67
	Q*	-4.72	**0.20**	0.70	-4.17	**0.21**	0.65	-0.89	**0.41**	0.77	-0.63	**0.37**	0.76
	Q**	-1.13	**0.29**	0.71	-1.64	**0.30**	0.66	-0.12	**0.44**	0.75	-0.14	**0.45**	0.75
WaterGAP2.2e	Q	-1.24	**0.44**	0.82	-2.01	**0.43**	0.83	0.02	**0.67**	0.88	-0.08	**0.66**	0.87
	Q*	-1.00	**0.53**	0.87	-1.62	**0.52**	0.86	0.05	**0.68**	0.89	-0.06	**0.67**	0.89
	Q**	-0.92	**0.56**	0.87	-1.61	**0.52**	0.86	0.05	**0.69**	0.89	-0.06	**0.67**	0.89

While the summary statistics provided an overall assessment of model performance, the spatial distribution of NSE reveals substantial differences. In contrast to the calibrated WaterGAP, the spatial pattern of NSE for the annual streamflow time series (Q) shows negative values for most of the 589 analysis locations for H08 and MIROC-INTEG-Land (Appendix A1).

Considering absolute anomalies of annual streamflow (Q^*), NSE becomes positive for MIROC-INTEG-LAND and more so for H08 for a number of basins in both hemispheres (Appendix A1, Figure 3, Appendix A2). Under Q^{**}, both H08 and WaterGAP show mainly positive NSE across many analysis locations, although WaterGAP retains a more heterogeneous spatial pattern compared to H08 (Figure 3, Appendix A2). Under Q^{**} evaluation, positive NSE values are widespread across many basins in North America, Europe, and northern Asia for all climate–GHM combinations, whereas performance remains more spatially heterogeneous in Africa, South America, and Australia.

The fraction of analysis locations showing improved performance is high for all models, exceeding 93% for (Q^*) and 83% for (Q^{**}) (Figure 3; Appendix A2). Here, improvement refers to an increase in NSE relative to the Q; that is, (NSE(Q^*) > NSE(Q)) and (NSE(Q^{**}) > NSE(Q)), respectively. The fraction of improved locations is generally higher for H08 and MIROC-INTEG-LAND than for WaterGAP, reflecting the comparatively better performance of WaterGAP for (Q).

Spatial patterns obtained using 20CRv3–W5E5 forcing (Appendix A2) closely follow those for 20CRv3–ERA5 (with only minor differences in median NSE and the fraction of improved locations), indicating that the observed performance patterns are robust to the choice of climate forcing.

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

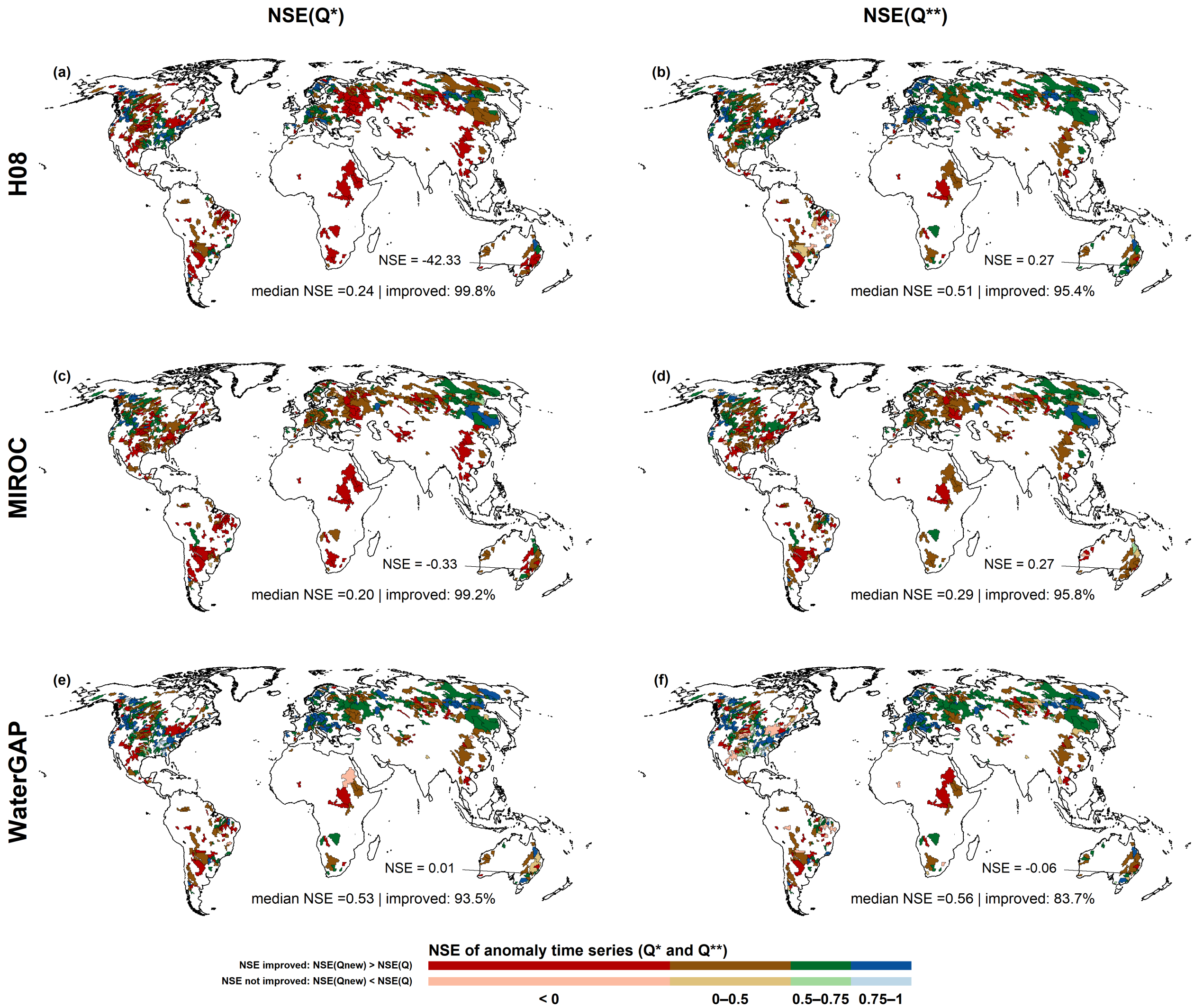

Figure 3. Global basin-scale Nash–Sutcliffe Efficiency (NSE) for anomaly-based streamflow time series driven by 20CRv3–ERA5 forcing. Left panels show NSE for Q^*, right panels for Q^{**}. Rows correspond to GHMs (H08, MIROC-INTEG-LAND, and WaterGAP). Colours indicate change in performance relative to the absolute time series (Q): darker colours denote improved NSE and lighter colours denote reduced NSE. Median NSE and the fraction of basins with improved performance are shown for each panel.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. In the legend, NSE(Qnew) refers to the NSE of the anomaly-based streamflow time series (Q^* or Q^{**}, depending on the panel.

The NSE and KGE behaviour is consistent with their underlying components (Figure 4, Appendix Table B1). The box plots in 4 summarise the central range of model performance, excluding extreme outliers, while full summary statistics are provided in Appendix B1. deviation ratio (γ) remains unchanged between Q and Q^* and moves closer to its optimal value of one for

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

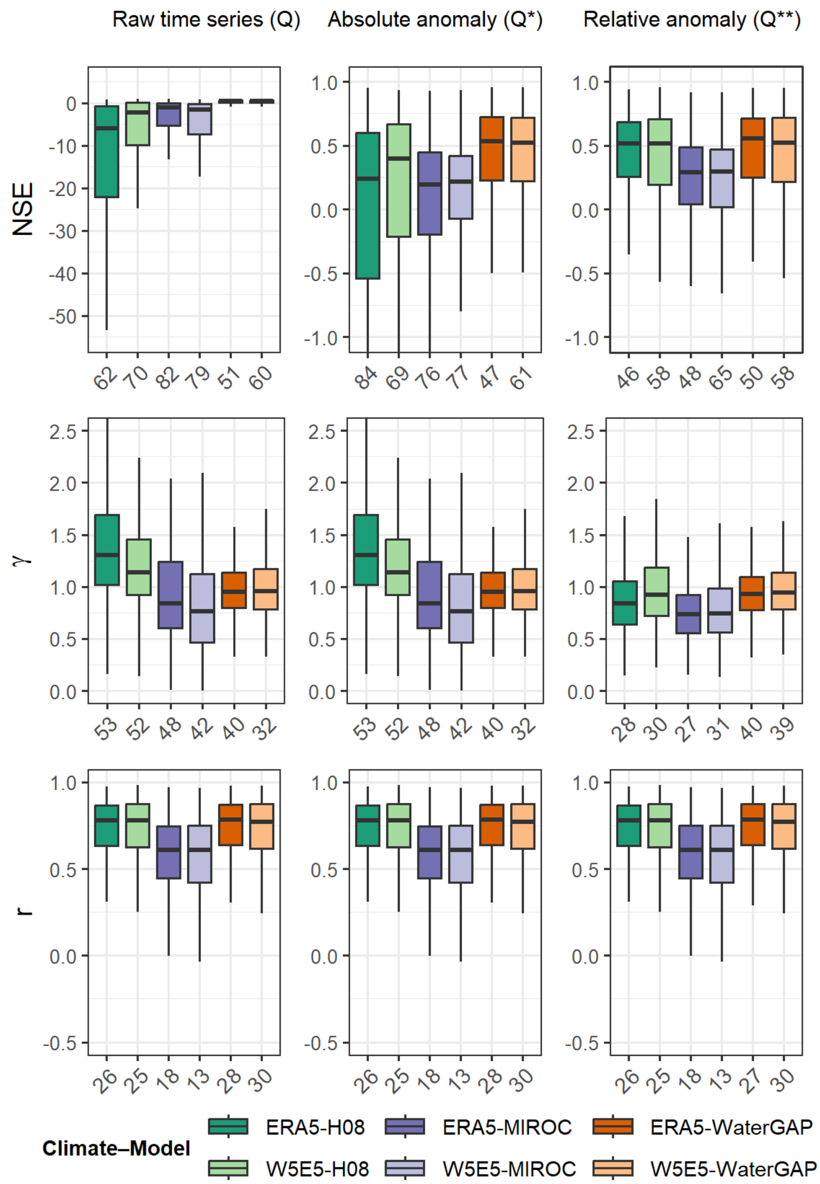

Figure 4. Model evaluation of six ensemble members (20crv3-era5 and 20crv3-w5e5 forced, with three GHMs: H08, MIROC-INTEG-LAND and WaterGAP) at 589 analysis locations using four evaluation metrics Nash-Sutcliffe Efficiency (NSE), γ (amplitude bias) and Pearson's r (correlation) for annual raw streamflow time series (Q) and Absolute (Q^*) and relative (Q^{**}) anomaly of streamflow time series. Outliers (outside 1.5 × inter-quartile range) are excluded, but the number of stations that are defined as outliers is indicated on the x-axis.

295 Q^{**}. Correlation (r) remains unchanged among. This reflects that relative anomalies decrease the bias-related discrepancies of the simulated and observed standard deviations as compared to absolute anomalies. For all three GHMs, correlation does

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

not differ between the two climate forcings. However, the climate forcing 20CRv3-ERA5 leads, on average, to an increase in simulated variability as compared to 20CFv3-W5E5, but only for H08 and MIROC-INTEG-LAND. MIROC-INTEG-LAND shows an appreciably worse correlation than the other two GHMs and, regarding Q, a worse standard deviation ratio, too.

While Table 2 and Appendix Table B1 summarise the full distribution of NSE, r, and γ respectively, Figure 4 focuses on the central range of performance after excluding extreme outliers (defined as 1.5 x inter-quartile range). The box plots show clear differences among GHMs. H08 generally exhibits γ values greater than one under Q and Q^*, indicating overestimation of variability, whereas MIROC-INTEG-LAND tends to underestimate variability. Under Q^{**}, H08 is slightly underestimating but closer to the optimal value compared to the other two configurations. In contrast, MIROC-INTEG-LAND remains below one and shows a small further reduction in γ. WaterGAP shows median γ values closest to one under both climate forcings. H08 and WaterGAP also exhibit higher median correlations than MIROC-INTEG-LAND.

4.3 GHM performance in capturing wet and dry anomaly years

Table 3. Performance of six climate-GHM ensemble members in reproducing dry (Q^{**}_{90}) and wet (Q^{**}_{10}) extreme streamflow anomalies across 589 analysis locations. NSE quantifies agreement in extreme anomaly magnitude, while P and Top-5 match quantify reproduction of extreme anomaly occurrence.

Ensemble	Q^{**}_{90} (dry extreme anomaly)			Q^{**}_{10} (wet extreme anomaly)		
	NSE	Perfect (%)	Top-5 match (%)	NSE	Perfect (%)	Top-5 match (%)
H08–20CRv3–ERA5	0.36	11.71	52.97	0.36	13.24	54.50
MIROC–20CRv3–ERA5	0.06	8.83	39.56	-0.08	8.66	46.69
WaterGAP–20CRv3–ERA5	0.58	12.22	53.99	0.16	13.07	57.39
H08–20CRv3–W5E5	0.42	11.54	51.61	0.41	12.90	57.39
MIROC–20CRv3–W5E5	0.02	7.30	42.11	0.08	11.04	49.92
WaterGAP–20CRv3–W5E5	0.57	11.88	52.46	0.38	13.41	58.57

Perfect = percentage of analysis locations where the observed and simulated extreme anomaly occur in the same year. Top-5 match = percentage of analysis locations where the observed extreme anomaly year is among the five most extreme simulated years in the simulated time series.

NSE computed across 589 analysis locations is generally higher for dry anomalies (Q^{**}_{90}; 0.02-0.58) than for wet anomalies (Q^{**}_{10}; -0.08-0.41) (see Table 3). For Q^{**}_{90}, WaterGAP consistently shows the highest performance across all metrics, with NSE values exceeding 0.5 as well as the highest percentages of perfect and Top-5 matches. In contrast, MIROC-INTEG-LAND shows the lowest performance (Q^{**}_{90}; 0.02-0.06). For Q^{**}_{10}, H08 exhibits the highest NSE values (0.36-0.41), whereas WaterGAP shows the highest Top-5 match percentages (57-59%), MIROC-INTEG-LANG again shows the lowest performance, including negative NSE values under 20CRv3-ERA5.

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

In addition to reproducing extreme anomaly magnitudes, model performance was evaluated based on the occurrence of extreme anomaly years. Exact agreement between observed and simulated extreme anomaly years is achieved at approximately 7-13% of analysis locations across all ensemble members (Table 3). When a broader event matching criterion is applied, the agreement increases to approximately 40-59% of the analysis locations. For example, a Top-5 match value of 52% indicates that, at 52% of the analysis locations, the observed extreme anomaly year is identified among the five most extreme years in the simulated time series; it does not imply that 52% of all extreme events are reproduced exactly.

The significantly higher Top-5 match percentage compared to exact matches indicates that GHMs often identify the same extreme years, even when the observed extreme anomaly year is not reproduced exactly. Differences among GHMs are more pronounced than differences between climate forcings, with WaterGAP and H08 generally showing higher agreement than MIROC-INTEG-LANG for both NSE and event matching metrics.

5 Discussion

5.1 Understanding GHM performance through anomaly-based evaluation

This study shows that agreement in absolute streamflow magnitude and agreement in interannual variability represent fundamentally different aspects of model performance and should not be interpreted interchangeably. Long-term mean streamflow at a given location reflects the cumulative effects of runoff generation, evapotranspiration, storage processes, routing, human water use, and climate forcing uncertainty, all of which differ among the GHMs. Consequently, structural assumptions and parameterisation choices can lead to substantial differences in simulated streamflow magnitude. In contrast, evaluation of annual streamflow anomalies isolates temporal fluctuations around that mean, mostly driven by climate, providing a more direct assessment of how models translate climate variability into hydrological variability.

This distinction is particularly relevant for climate change impact assessments, where hydrological change is typically interpreted relative to a reference period rather than in terms of absolute streamflow values. Evaluation based solely on streamflow magnitude therefore emphasises aspects of model performance that are less directly aligned with the intended application of the model output. Relative anomaly-based evaluation (Q^{**}) focuses on deviations from the mean state and provides a more consistent measure of model responsiveness to climate variability. While absolute anomaly configuration (Q^{*}) removes mean bias, they remain sensitive to variability scaling and is more commonly applied in regional studies with calibrated models and bias-corrected climate input. In global hydrological modelling contexts, where such calibration is challenging, relative anomaly-based model evaluation offers a more robust and comparable framework across regions.

The comparatively small differences in the performance of relative streamflow anomalies between the two climate forcings (Table 2) suggest that the representation of interannual variability is less sensitive to forcing choice than the simulation of absolute streamflow. This indicates that anomaly-based evaluation captures a relatively robust aspect of model behaviour and may therefore provide a useful complementary perspective to conventional magnitude-based evaluation.

These findings are consistent with previous large-scale model evaluations. Gudmundsson et al. (2012a) showed that GHMs can exhibit rather high correlations with observed streamflow while still misrepresenting variability. For example, H08 and Wa-

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

terGAP achieved, for annual time series of median daily streamflow, correlation values of $r \approx 0.90$, and relative differences in standard deviation of the annual time series of -0.53 and -0.32, respectively, indicating different degrees of underestimation of the relative interannual variability. The present results support this broader conclusion that GHMs with relatively high correlations do not necessarily reproduce observed variability magnitudes accurately. Evaluation based on relative anomalies therefore provided complementary information on model responsiveness to climate variability that is not captured by correlation alone.

5.2 Recommendations for evaluating GHMs for climate change impact assessments

Krysanova et al. (2020) suggested assigning, in multi-model ensembles for estimating hydrological impacts of climate change, greater weight to models that better reproduce observed streamflow magnitude. Under such criteria, partially calibrated GHMs such as WaterGAP are favoured. However, the present results show that performance differences among GHMs are much less pronounced when models are evaluated in terms of their ability to simulate the mainly climate-driven variability of annual streamflow. Model performance of the relative or absolute annual streamflow anomalies is more relevant if relative or absolute changes of hydrological model variables and not the actual magnitudes (e.g., of Q) are to be quantified. But due to the biases of climate models regarding the computation of historic climate, estimation of actual future magnitudes can only be attempted if bias correction is known to be very good.

In addition to assessing the performance of the time series of annual streamflow anomalies, it should be evaluated how well GHMs can capture unusually wet and dry years. Our extreme anomaly analysis showed different model rankings for dry and wet years. For example, WaterGAP consistently showed the highest performance in capturing dry anomalies, whereas for wet anomalies, H08 achieved the highest NSE values and WaterGAP the highest Top-5 match percentages.

An additional advantage of evaluating annual streamflow anomalies as compared to evaluating daily or monthly time series of streamflow is that it can also be applied to models that only compute runoff, without streamflow routing, or models using different drainage direction maps. In the latter case, a consistent comparison of these models requires a model-specific mapping of grid cells to gauging stations with observed streamflow, which would be very cumbersome to do. Our anomaly-based evaluation could be applied to annual runoff aggregated over the upstream area of a gauging station instead of annual streamflow because annual runoff integrates the hydrological response of the upstream basin while being less dependent on routing assumptions than daily or monthly values. Thus, if models without any streamflow routing or with different routing maps are to be included in the model intercomparison, we suggest evaluating the aggregated annual basin runoff anomalies. Then, the performance of a larger number of models that serve to quantify hydrological responses to climate change could be consistently compared.

5.3 Caveats and applicability of anomaly-based evaluation

The interpretation of anomaly-based model evaluation should be considered in relation to how “change” is defined in this study. Here, anomalies are calculated relative to the long-term mean of the historical period, whereas climate change impact assessments typically quantify change between distinct time periods, such as historical and future periods. Moreover, the magnitude of climate change signals in historical observations is generally smaller than in projected future conditions. The present

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

analysis evaluates how consistently models translate annual climate variability into hydrological response, which is a key prerequisite for credible climate change impact assessments. While this provides information directly relevant to model suitability for climate change studies, the accuracy of future projections will additionally depend on how model process representations perform under future climatic conditions and on the uncertainty in future climate forcing.

Regarding process representation, our analysis was restricted to three large-scale models that share a common routing framework, allowing differences in performance to be attributed primarily to differences in hydrological process representation rather than river routing. Only MIROC-INTEG-LAND simulates the impact of changing CO_2 and climate on vegetation processes that affect evaporation and thus runoff. Although MIROC-INTEG-LAND generally showed lower performance than H08 and WaterGAP in the historical evaluation, an appropriate simulation of vegetation dynamics is likely to become more relevant with increasing climate change. Therefore, models with different process structures should be included in multi-model ensembles to adequately represent structural uncertainty in future hydrological projections, even if historic simulations are sub-optimal.

6 Conclusions

This study evaluated the ability of three global hydrological models (H08, MIROC-INTEG-LAND, and WaterGAP2.2e) to translate interannual climate variability into streamflow variability using annual streamflow observations from 589 globally distributed catchments. The results show that 1) all three models can estimate both absolute and relative interannual streamflow variability better than annual streamflow magnitude, and 2) their anomaly performance as measured by KGE and NSE is much more similar than their magnitude performance. The highest performance is achieved for relative anomalies of annual streamflow, as they are insensitive to the often large bias of GHMs that are not calibrated against the streamflow at the gauging stations that supply the observations. This suggests that the conventional evaluation of streamflow magnitude may underestimate the ability of GHMs to represent hydrological responses to interannual climate variability and thus climate change.

We recommend including, in assessments of the suitability of GHMs for climate change impact studies, the evaluation of model results against observed mean annual streamflow anomalies and not relying on comparisons to streamflow magnitude. Because climate change impacts are typically quantified as changes relative to a reference period, evaluation of relative anomalies of annual streamflow against observations is most relevant. Both correlations and standard deviation discrepancies of the annual anomaly time series should be analysed, and, additionally, the anomaly performance regarding dry and wet extremes. Even though the impact of different historic climate forcings on the evaluation result was found to be relatively small, consideration of more than one climate forcing is preferable.

The GHM WaterGAP shows a slightly better performance of relative annual streamflow anomaly than the H08 due to its better simulation of the standard deviation, while the MIROC-INTEG-LAND, a land surface model, has a distinctly lower performance. However, it is the only model of the three to simulate the important vegetation response to CO_2 increases and climate change, which is why it should not be excluded from a multi-model ensemble.

Future work should extend the presented anomaly-based GHM evaluation to a larger range of hydrological and land surface models by comparing, instead of modelled annual streamflow anomalies, anomalies of runoff aggregated over the upstream

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

basin of the gauging stations with available observations of annual streamflow. Such an approach would allow a consistent assessment of the ability of many different models to translate interannual climate variability across models with different or no routing schemes. The 589 observational 30-year time series selected for this study are suitable for performing such an extended performance analysis.

Code and data availability. Monthly streamflow simulations from H08, MIROC-INTEG-LAND, and WaterGAP2.2e forced with the 20CRv3–ERA5 and 20CRv3–W5E5 climate forcings were obtained from the ISIMIP3a archive (https://data.isimip.org/search/tree/ISIMIP3a/OutputData/water_global/). The observed streamflow records used in this study are those compiled by (Müller Schmied and Schiebener, 2022) for the calibration and evaluation of WaterGAP. These data can be provided upon request to the authors. Extracted observed data for the 589 analysis locations and R scripts used for simulated data processing and analysis are available through Zenodo at: https://doi.org/10.5281/zenodo.20487152

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

Appendix A: NSE Maps

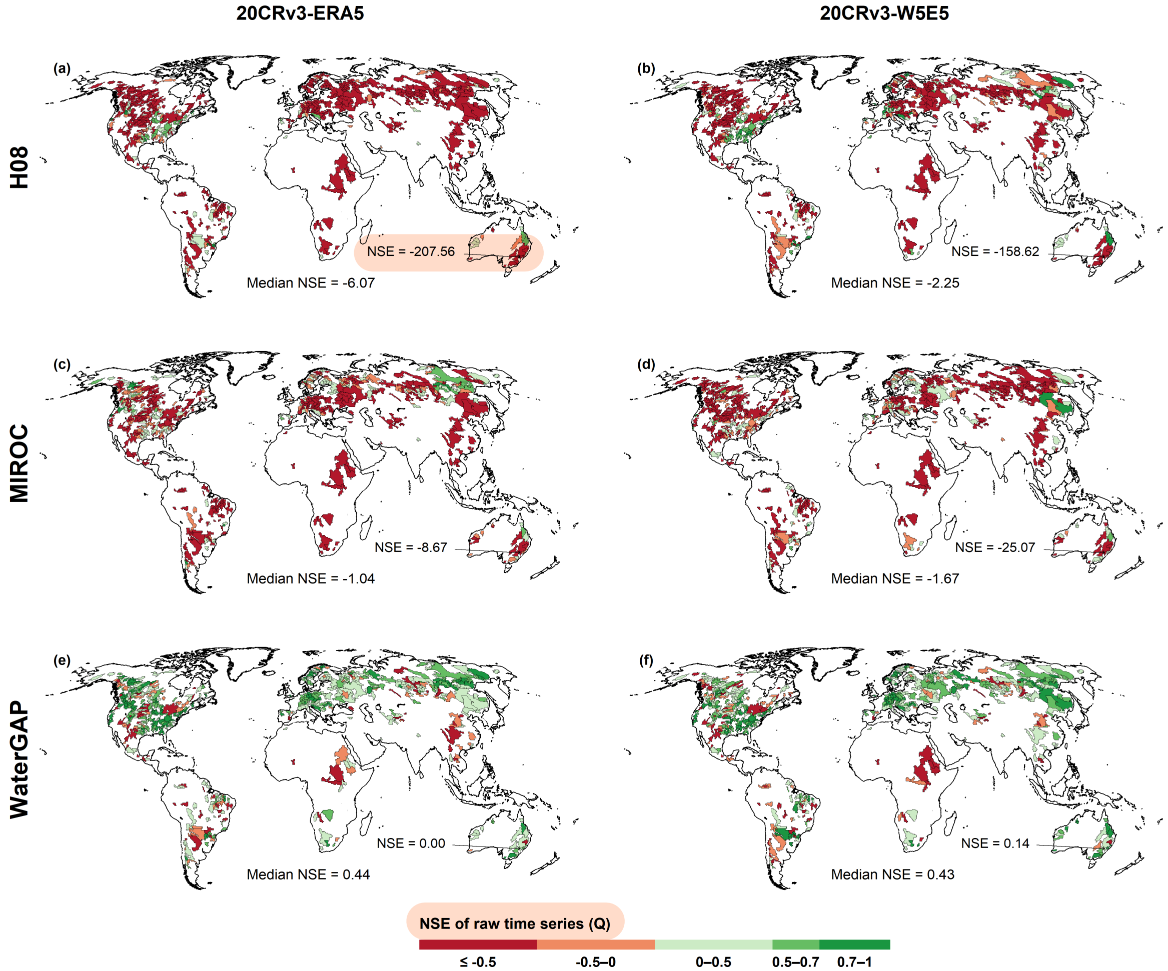

Figure A1. Global basin-scale Nash–Sutcliffe Efficiency (NSE) for raw streamflow time series driven by 20CRv3–ERA5 forcing (Left panel) and 20CRv3-W5E5 (Right panel). Rows correspond to GHMs (H08, MIROC-INTEG-LAND, and WaterGAP). Colours indicate NSE values of each basin correspond to each analysis location. Median NSE shown for each panel.

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

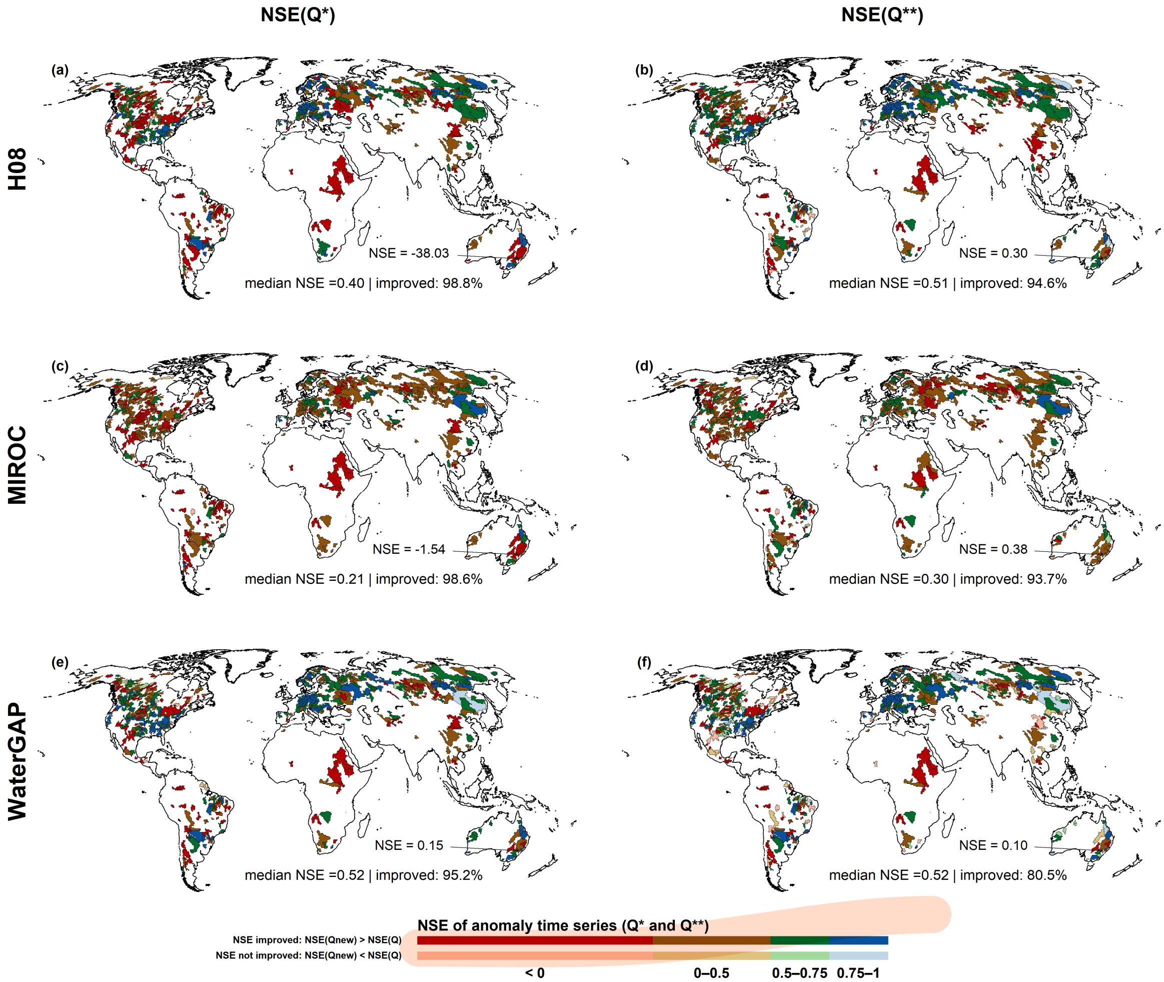

Figure A2. Global basin-scale Nash–Sutcliffe Efficiency (NSE) for anomaly-based streamflow time series driven by 20CRv3–W5E5 forcing. Left panels show NSE for Q^*, right panels for Q^{**}. Rows correspond to GHMs (H08, MIROC-INTEG-LAND, and WaterGAP). Colours indicate change in performance relative to the absolute time series (Q): darker colors denotes improved NSE and lighter colors denotes reduced NSE. Median NSE and the fraction of basins with improved performance are shown for each panel. 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. In the legend, NSE(Qnew) refers to the NSE of the anomaly-based streamflow time series (Q^* or Q^{**}, depending on the panel).

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

Appendix B: Summary statistics additional results

Table B1. Summary statistics (P5, P50, P95) of variability ratio (γ) and correlation (r) and for annual streamflow (Q), absolute anomaly (Q^*), and relative anomaly (Q^{**}) for each ensemble member under 20CRv3-ERA5 and 20CRv3-W5E5 climate forcings.

GHM	Config	γ						r					
		20CRv3-ERA5			20CRv3-W5E5			20CRv3-ERA5			20CRv3-W5E5		
		P5	P50	P95	P5	P50	P95	P5	P50	P95	P5	P50	P95
H08	Q	0.68	**1.30**	3.51	0.59	**1.14**	3.38	0.33	**0.78**	0.93	0.30	**0.78**	0.95
	Q*	0.68	**1.30**	3.51	0.59	**1.14**	3.38	0.33	**0.78**	0.93	0.30	**0.78**	0.95
	Q**	0.38	**0.84**	1.66	0.44	**0.92**	1.86	0.33	**0.78**	0.93	0.30	**0.78**	0.95
MIROC-INTEG-LAND	Q	0.28	**0.84**	2.82	0.16	**0.76**	2.59	0.13	**0.61**	0.87	0.08	**0.61**	0.85
	Q*	0.28	**0.84**	2.82	0.16	**0.76**	2.59	0.13	**0.61**	0.87	0.08	**0.61**	0.85
	Q**	0.32	**0.74**	1.47	0.33	**0.74**	1.62	0.13	**0.61**	0.87	0.08	**0.61**	0.85
WaterGAP2.2e	Q	0.56	**0.95**	1.74	0.59	**0.96**	1.78	0.31	**0.78**	0.94	0.23	**0.77**	0.94
	Q*	0.56	**0.95**	1.74	0.59	**0.96**	1.78	0.31	**0.78**	0.94	0.23	**0.77**	0.94
	Q**	0.58	**0.93**	1.71	0.59	**0.94**	1.76	0.31	**0.78**	0.94	0.23	**0.77**	0.94

Author contributions. ASP and PD developed the study and the methodological framework. ASP conducted the literature review, processed the data, performed the analyses, prepared the figures and tables, interpreted the results, and wrote the manuscript. PD contributed to the interpretation of the results and provided substantial revisions and guidance throughout the study and manuscript preparation.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We would like to thank the authors of Müller Schmied and Schiebener (2022) for providing the station-corrected observation data.

Financial support. The authors thank the Open Access Publication Fund of Goethe University Frankfurt am Main for supporting the open access publication. The majority of work has been carried out in the framework of the Co-development of Methods to utilise uncertain

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

multi-model based Information on freshwater related hazards of Climate Change (CO-MICC) project, funded by the German Federal Ministry of Education and Research (BMBF) (grant no. 690462 — ERA4CS — H2020-SC5-2014-2015/H2020-SC5-2015-one-stage 51) in the framework of the European Research Area for Climate Services (ERA4CS).

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

References

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., et al.: The twentieth century reanalysis project, Quarterly Journal of the Royal Meteorological Society, 137, 1–28, https://doi.org/10.1002/qj.776, 2011.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., Hersbach, H., and Buontempo, C.: WFDE5: bias-adjusted ERA5 reanalysis data for impact studies, Earth System Science Data, 12, 2097–2120, https://doi.org/10.5194/essd-12-2097-2020, 2020.

Döll, P., Douville, H., Güntner, A., Schmied, H. M., and Wada, Y.: Modelling freshwater resources at the global scale: challenges and prospects, Surveys in Geophysics, 37, 195–221, https://doi.org/10.1007/s10712-015-9343-1, 2016.

Döll, P., Müller Schmied, H., Schuh, C., Portmann, F. T., and Eicker, A.: Global-scale assessment of groundwater depletion and related groundwater abstractions, Hydrology and Earth System Sciences, 22, 4479–4497, https://doi.org/10.1002/2014WR015595, 2018.

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

Frieler, K., Volkholz, J., Lange, S., Schewe, J., Mengel, M., del Rocío Rivas López, M., Otto, C., Reyer, C. P., Karger, D. N., Malle, J. T., et al.: Scenario setup and forcing data for impact model evaluation and impact attribution within the third round of the Inter-Sectoral Model Intercomparison Project (ISIMIP3a), Geoscientific Model Development, 17, 1–51, https://doi.org/10.5194/gmd-17-1-2024, 2024.

Fung, F., Lopez, A., and New, M.: Water availability in+ 2 C and+ 4 C worlds, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 369, 99–116, https://doi.org/10.1098/rsta.2010.0293, 2011.

Gerten, D., Betts, R., and Döll, P.: Cross-chapter box on the active role of vegetation in altering water flows under climate change, in: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects, edited by Field, C. B., Barros, V. R., Dokken, D. J., Mach, K. J., Mastrandrea, M. D., Bilir, T. E., Chatterjee, M., Ebi, K. L., Estrada, Y. O., Genova, R. C., Girma, B., Kissel, E. S., Levy, A. N., MacCracken, S., Mastrandrea, P. R., and White, L. L., pp. 157–161, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/CBO9781107415379, 2014.

Gudmundsson, L., Tallaksen, L. M., and Stahl, K.: Spatial cross-correlation patterns of European low, mean and high flows, Hydrological Processes, 25, 1034–1045, https://doi.org/10.1002/hyp.7807, 2011.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., et al.: Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe, Journal of Hydrometeorology, 13, 604–620, https://doi.org/10.1175/JHM-D-11-083.1, 2012a.

Gudmundsson, L., Wagener, T., Tallaksen, L., and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resources Research, 48, https://doi.org/10.1029/2011WR010911, 2012b.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Haddeland, I., Heinke, J., Biemans, H., Eisner, S., Flörke, M., Hanasaki, N., Konzmann, M., Ludwig, F., Masaki, Y., Schewe, J., et al.: Global water resources affected by human interventions and climate change, Proceedings of the National Academy of Sciences, 111, 3251–3256, https://doi.org/10.1073/pnas.1222475110, 2014.

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., Hanasaki, N., Heinke, J., Ludwig, F., Voss, F., et al.: Climate change impact on available water resources obtained using multiple global climate and hydrology models, Earth System Dynamics, 4, 129–144, https://doi.org/10.5194/esd-4-129-2013, 2013.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources–Part 1: Model description and input meteorological forcing, Hydrology and Earth System Sciences, 12, 1007–1025, https://doi.org/10.5194/hess-12-1007-2008, 2008a.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources–Part 2: Applications and assessments, Hydrology and Earth System Sciences, 12, 1027–1037, https://doi.org/10.5194/hess-12-1027-2008, 2008b.

Hattermann, F. F., Krysanova, V., Gosling, S. N., Dankers, R., Daggupati, P., Donnelly, C., Flörke, M., Huang, S., Motovilov, Y., Buda, S., et al.: Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins, Climatic Change, 141, 561–576, https://doi.org/10.1007/s10584-016-1829-4, 2017.

Heinicke, S., Volkholz, J., Schewe, J., Gosling, S. N., Schmied, H. M., Zimmermann, S., Mengel, M., Sauer, I. J., Burek, P., Chang, J., et al.: Global hydrological models continue to overestimate river discharge, Environmental Research Letters, 19, 074 005, https://doi.org/10.1088/1748-9326/ad52b0, 2024.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Knoben, W. J., Freer, J. E., and Woods, R. A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., and Kundzewicz, Z. W.: How the performance of hydrological models relates to credibility of projections under climate change, Hydrological Sciences Journal, 63, 696–720, https://doi.org/10.1080/02626667.2018.1446214, 2018.

Krysanova, V., Zaherpour, J., Didovets, I., Gosling, S. N., Gerten, D., Hanasaki, N., Müller Schmied, H., Pokhrel, Y., Satoh, Y., Tang, Q., et al.: How evaluation of global hydrological models can help to improve credibility of river discharge projections under climate change, Climatic Change, 163, 1353–1377, https://doi.org/10.1007/s10584-020-02840-0, 2020.

Lange, S.: Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1. 0), Geoscientific Model Development, 12, 3055–3070, https://doi.org/10.5194/gmd-12-3055-2019, 2019.

Lange, S., Menz, C., Gleixner, S., Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Müller Schmied, H., Hersbach, H., Buontempo, C., and Cagnazzo, C.: WFDE5 over land merged with ERA5 over the ocean (W5E5 v2.0), https://doi.org/10.48364/ISIMIP.342217, https://data.isimip.org/10.48364/ISIMIP.342217, 2021.

Mengel, M., Treu, S., Lange, S., and Frieler, K.: ATTRICI v1. 1–counterfactual climate for impact attribution, Geoscientific Model Development, 14, 5269–5284, https://doi.org/10.5194/gmd-14-5269-2021, 2021.

Müller Schmied, H. and Schiebener, L.: Assessing the suitability of streamflow station observations for consistent evaluation of simulated river discharge data of the ISIMIP global water sector, Tech. rep., PROCLIAS Project, https://proclias.eu/output/papers/2022-assessing-suitability-streamflow-station-observations/view, 2022.

Müller Schmied, H., Trautmann, T., Ackermann, S., Cáceres, D., Flörke, M., Gerdener, H., Kynast, E., Peiris, T. A., Schiebener, L., Schumacher, M., et al.: The global water resources and use model WaterGAP v2.2e: description and evaluation of modifications and new features, Geoscientific Model Development, 17, 287–326, https://doi.org/10.5194/gmd-17-8817-2024, 2024.

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Peiris, T. A. and Döll, P.: Improving the quantification of climate change hazards by hydrological models: a simple ensemble approach for considering the uncertain effect of vegetation response to climate change on potential evapotranspiration, Hydrology and Earth System Sciences, 27, 3663–3686, https://doi.org/10.5194/hess-27-3663-2023, 2023.

Reinecke, R., Müller Schmied, H., Trautmann, T., Andersen, L. S., Burek, P., Flörke, M., Gosling, S. N., Grillakis, M., Hanasaki, N., Koutroulis, A., et al.: Uncertainty of simulated groundwater recharge at different global warming levels: a global-scale multi-model ensemble study, Hydrology and Earth System Sciences, 25, 787–810, https://doi.org/10.5194/hess-25-787-2021, 2021.

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P., Wiese, D. N., Wada, Y., Long, D., Reedy, R. C., et al.: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data, Proceedings of the National Academy of Sciences, 115, E1080–E1089, https://doi.org/10.1073/pnas.1704665115, 2018.

Seibert, J.: Reliability of model predictions outside calibration conditions: Paper presented at the Nordic Hydrological Conference (Røros, Norway 4-7 August 2002), Hydrology Research, 34, 477–492, https://doi.org/10.2166/nh.2003.0019, 2003.

Shrestha, R. R., Cannon, A. J., Hoffman, S., Whibley, M., and Lima, A.: Benchmarking historical performance and future projections from a large-scale hydrologic model with a watershed hydrologic model, Hydrology and Earth System Sciences, 29, 2881–2900, https://doi.org/10.5194/hess-29-2881-2025, https://doi.org/10.5194/hess-29-2881-2025, 2025.

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., Allan, R., Yin, X., Vose, R., Titchner, H., et al.: Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system, Quarterly Journal of the Royal Meteorological Society, 145, 2876–2908, https://doi.org/10.1002/qj.3598, 2019.

Telteu, C.-E., Müller Schmied, H., Thiery, W., Leng, G., Burek, P., Liu, X., Boulange, J. E. S., Seaby Andersen, L., Grillakis, M., Gosling, S. N., et al.: Understanding each other's models: a standard representation of global water models to support improvement, intercomparison, and communication, Geoscientific Model Development, 14, 3843–3870, https://doi.org/10.5194/gmd-14-3843-2021, 2021.

Veldkamp, T. I. E., Zhao, F., Ward, P. J., de Moel, H., Aerts, J. C., Schmied, H. M., Portmann, F. T., Masaki, Y., Pokhrel, Y., Liu, X., et al.: Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study, Environmental Research Letters, 13, 055 008, https://doi.org/10.1088/1748-9326/aab96f, 2018.

Wada, Y., Wisser, D., Eisner, S., Flörke, M., Gerten, D., Haddeland, I., Hanasaki, N., Masaki, Y., Portmann, F. T., Stacke, T., et al.: Multimodel projections and uncertainties of irrigation water demand under climate change, Geophysical research letters, 40, 4626–4632, https://doi.org/10.1002/grl.50686, 2013.

Yokohata, T., Kinoshita, T., Sakurai, G., Pokhrel, Y., Ito, A., Okada, M., Satoh, Y., Kato, E., Nitta, T., Fujimori, S., et al.: MIROC-INTEG-LAND version 1: a global biogeochemical land surface model with human water management, crop growth, and land-use change, Geoscientific Model Development, 13, 4713–4747, https://doi.org/10.5194/gmd-13-4713-2020, 2020.

Yoshida, T., Hanasaki, N., Nishina, K., Boulange, J., Okada, M., and Troch, P.: Inference of parameters for a global hydrological model: Identifiability and predictive uncertainties of climate-based parameters, Water Resources Research, 58, e2021WR030 660, https://doi.org/10.1029/2021WR030660, 2022.

Zaherpour, J., Gosling, S. N., Mount, N., Schmied, H. M., Veldkamp, T. I., Dankers, R., Eisner, S., Gerten, D., Gudmundsson, L., Haddeland, I., et al.: Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts, Environmental Research Letters, 13, 065 015, https://doi.org/10.1088/1748-9326/aac547, 2018.

https://doi.org/10.5194/egusphere-2026-3412
Preprint. Discussion started: 24 July 2026

Zhao, F., Nie, N., Liu, Y., Yi, C., Guillaumot, L., Wada, Y., Burek, P., Smilovic, M., Frieler, K., Buechner, M., et al.: Benefits of calibrating a global hydrological model for regional analyses of flood and drought projections: A case study of the Yangtze River Basin, Water Resources Research, 61, e2024WR037 153, https://doi.org/10.1029/2024WR037153, 2025.