Global hydrological models continue to overestimate river discharge

To cite this article: Stefanie Heinicke et al 2024 Environ. Res. Lett. 19 074005

View the article online for updates and enhancements.

You may also like
- Intercomparison of global river discharge simulations focusing on dam operation—multiple models analysis in two case-study river basins, Missouri-Mississippi and Green–Colorado
  Yoshimitsu Masaki, Naota Hanasaki, Hester Biemans et al.

- Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study
  T I E Veldkamp, F Zhao, P J Ward et al.

- Characterizing the multisectoral impacts of future global hydrologic variability
  Abigail Birnbaum, Ghazal Shabestanipour, Mengqi Zhao et al.
Global hydrological models continue to overestimate river discharge

Stefanie Heinicke, Jan Volkholz, Jacob Schewe, Simon N Gosling, Hannes Müller Schmied, Sandra Zimmermann, Matthias Mengel, Inga J Sauer, Peter Burek, Jinfeng Chang, Sian Kou-Giesbrecht, Manoli Grillakis, Luca Guillaumot, Naota Hanasaki, Aristeidis Koutroulis, Kedar Otta, Wei Qi, Yusuke Sato, Tobias Stacke, Tokuta Yokohata and Katja Frieler

Abstract

Global hydrological models (GHMs) are widely used to assess the impact of climate change on streamflow, floods, and hydrological droughts. For the ‘model evaluation and impact attribution’ part of the current round of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP3a), modelling teams generated historical simulations based on observed climate and direct human forcings with updated model versions. Here we provide a comprehensive evaluation of daily and maximum annual discharge based on ISIMIP3a simulations from nine GHMs by comparing the simulations to observational data from 644 river gauge stations. We also assess low flows and the effects of different river routing schemes. We find that models can reproduce variability in daily and maximum annual discharge, but tend to overestimate both quantities, as well as low flows. Models perform better at stations in wetter areas and at lower elevations. Discharge routed with the river routing model CaMa-Flood can improve the performance of some models, but for others, variability is overestimated, leading to reduced model performance. This study indicates that areas for future model development include improving the simulation of processes in arid regions and cold dynamics at high elevations. We further suggest that studies attributing observed changes in discharge to historical climate change using the current model ensemble will be most meaningful in humid areas, at low elevations, and in places with a regular seasonal discharge as these are the regions where the underlying dynamics seem to be best represented.

1. Introduction

The water cycle is particularly susceptible to climate change, leading to changes in river flow with far-reaching consequences for water availability for humans (e.g. hydrological droughts), for climatic hazards such as river floods, but also for ecosystems (Schewe et al 2014, Maxwell et al 2019, Gudmundsson...
et al 2021, Thompson et al 2021, Van Vliet 2023). Fluvial floods led to the death of more than 200,000 people and incurred damages of 790 billion USD from 1980 to 2016 (Munch 2016). In addition, it has been estimated that floods and droughts resulted in 25 million people living in extreme poverty (Hallegatte et al 2017). Model-based projections show a strong increase in land area and population exposed to river floods and droughts due to increased global warming (Lange et al 2020).

Hydrological models are an important tool for decision-making in flood and drought management and preparedness and are used to make projections under different climate, land-use, and management scenarios. Global hydrological models (GHMs) in particular have been used, for example, to assess the impact of global warming on the availability of water resources (Schewe et al 2014, Liu et al 2017a, Pokhrel et al 2021), and on flood hazards (Dankers et al 2014, Hirabayashi et al 2021). In addition, GHMs can be used to attribute changes in the hydrological system to climate change. For example, Gudmundsson et al (2021) showed that observed changes in river flow are consistent with climatic changes, and Sauer et al (2021) found a climate signal in the trends in damages caused by river floods.

The evaluation of hydrological models is a critical first step towards impact attribution and future projections as for both purposes we need to understand to what degree models capture the processes determining discharge at individual locations, regions or globally. To this end, GHMs have been forced by observational climate data and observational direct human forcings (e.g. land use, location of dams and reservoirs) in Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2) already. These previous ISIMIP2 simulations have been evaluated, for example, regarding the influence of river routing on simulated discharge (Zhao et al 2017), the incorporation of human impact parameterizations (Liu et al 2017b, Veldkamp et al 2018), drought characteristics (Kumar et al 2022), or the seasonality of mean and extreme runoff (Zaherpour et al 2018). It has been shown that GHMs can simulate the extent of flooded areas, but for some events, GHMs overestimate the flood extent (Mester et al 2021). In addition, studies on economic damages caused by river floods showed that interannual variability of observed damages can be captured at least in some large scale areas (Sauer et al 2021). Deviations at the damage level could be due to either the discharge simulated by GHMs or from other sources in the modelling chain (e.g. assumed river protection levels, estimated distribution of assets, local physical vulnerabilities, Sauer et al 2021).

For ISIMIP3a, the climate input data and information on direct human forcings have been updated. For example, the data cover three additional years (2017–2019), accordingly more recent observational data has been used for bias adjustment, and an update in the bias adjustment method reduced excessively high daily maximum temperature values (Lange et al 2021, table 1 in Frieler et al 2024 shows details on data provided for daily climate, land use, irrigation, dams and reservoirs, and water abstraction). This study comprehensively evaluates the performance of nine GHMs that have contributed discharge data to ISIMIP3a so far. Five of these models were updated in terms of improving the representation of hydrological processes, and three models improved the representation of land cover (details in table 1). The effect of these improvements has been evaluated elsewhere (Müller Schmied et al 2023, Tsilimigkras et al 2023, Boulangé et al 2023, Yoshida et al 2022).

The aim of this study is to intercompare model performance regarding daily and maximum annual discharge as well as low flow and compare the performance of the models’ internal routing schemes to discharge generated from runoff by CaMa-Flood (Yamazaki et al 2011). We investigate which catchment properties correlate with model performance to suggest areas of model development and to identify stations that are particularly suitable for potential attribution studies or assessments of changes under future global warming as considered in ISIMIP3.

2. Methods

2.1. Simulated runoff and river discharge

We use daily runoff and river discharge provided by nine modelling groups based on the ISIMIP3a simulation round (Frieler et al 2024): CLASSIC, CWatM, H08, HydroPy, JULES-W2, MIROC-INTEG-LAND, ORCHIDEE-MICT, WaterGAP2-2e, and WEB-DHM-SG (table S1). WaterGAP2-2e is calibrated with observed discharge data (Müller Schmied et al 2023), and H08 uses parameters optimized to reproduce observed discharge in each climatic zone (Yoshida et al 2022). Simulated data are from the ‘obsclim + histsoc’ experiment that is designed for model evaluation and uses observation-based daily atmospheric climate forcing (GSWP3-W5 × 105, Lange et al 2023) and varying direct human forcings provided by ISIMIP (Frieler et al 2024). According to ISIMIP model documentation (www.isimip.org), all GHMs use land-use data, five models use data on water-use (CWatM, H08, HydroPy, MIROC-INTEG-LAND, WaterGAP2-2e) and four models use data on dams and reservoir (CWatM, H08, MIROC-INTEG-LAND, WaterGAP2-2e). Simulations are done on a 0.5° × 0.5° grid and are available for the years 1901–2019.

It has been shown that modelled peak river discharge using the river routing model CaMa-Flood
<table>
<thead>
<tr>
<th>Model</th>
<th>Hydrological processes</th>
<th>Land-cover representation</th>
<th>Comment</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASSIC</td>
<td>Change in calculation of evaporation using Penman-Monteith equation with CO2 response of vegetation</td>
<td>Did not participate in ISIMIP2a</td>
<td>Did not participate in ISIMIP2a</td>
<td>Melton et al. (2020)</td>
</tr>
<tr>
<td>CWatM</td>
<td>Change in calculation of evaporation using Penman-Monteith equation with CO2 response of vegetation</td>
<td>Using time varying land use instead of static</td>
<td>Using time varying land use instead of static</td>
<td>Burek et al. (2020)</td>
</tr>
<tr>
<td>H08</td>
<td>Land surface model parameters have been optimized at each climatic zone</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HydroPy</td>
<td>Land surface model parameters have been optimized at each climatic zone</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JULES-W2</td>
<td>(1) snow process module upgraded to 10-layer scheme; (2) soil hydrology scheme was changed to TOPMODEL scheme</td>
<td>Plant functional type (PFT) representation was changed from 5 PFTs to 13 PFTs, (2) used time varying land use instead of static</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORCHIDEE-MICT</td>
<td>Improvements for high-latitude processes including snow and permafrost, specifically: (1) soil freezing and snow processes; (2) soil hydrology and SOM-dependent soil thermal and hydraulic parameters for permafrost representation; (3) soil hydraulic representation; (4) volume of soil layers above the permafrost table; (5) drainage and storage capacity of soil layers</td>
<td></td>
<td>Guimberteau et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>MIROC-INTEG-LAND</td>
<td>Updates input for sectoral water use models, updated and extended data base for calibration, modified calibration routine to consider uncertainty in observation data; updated input for surface water bodies (e.g. GRAnD 1.3), implemented river water abstraction procedure</td>
<td></td>
<td>Müller Schmied et al. (2021, 2023)</td>
<td></td>
</tr>
<tr>
<td>WaterGAP2-e</td>
<td>Test model for the evaluation of water models, includes improved land use data and updated water abstraction routines</td>
<td></td>
<td></td>
<td>Qi et al. (2022)</td>
</tr>
<tr>
<td>WEB-DHM-SG</td>
<td>Used time varying land use instead of static</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.** Model improvements implemented since ISIMIP2a for models evaluated in this study. More details on models in table S1.
Figure 1. The location of the 644 GRDC stations used for model evaluation.

(Yamazaki et al. 2011) can perform better than the model’s internal routing scheme (Zhao et al. 2017). CaMa-Flood differs from other routing schemes in that it explicitly also parameterizes flood inundation dynamics and provides water depth and inundation extent as output variables (Yamazaki et al. 2011). It is therefore often used in combination with GHMs for flood modelling. CaMa-Flood provides its own routing scheme on a 15′ grid. However, CaMa-Flood does not include the lakes and reservoirs routines simulated by GHMs, which can result in some models performing better without CaMa-Flood. Thus, in addition to discharge routed with each model’s internal routing scheme (specified in table S1), we use each model’s simulated runoff to drive CaMa-Flood and then derive river discharge. Daily runoff is provided by all nine models listed above, and discharge modelled with the model’s internal routing scheme is provided by seven models. JULES-W2 provides discharge routed with the ‘native JULES’ river topography (Total Runoff Integrating Pathways scheme) that includes spatially distributed meandering and velocity data (Tsilimigkras et al. 2023), and has been shown to improve the accuracy of river flow simulations (Tsilimigkras et al. under review). In addition, JULES-W2 provides discharge routed with the DDM30 river topology.

2.2. Station selection and observed discharge data
To evaluate model performance, we use daily discharge data from the Global Runoff Data Centre (GRDC). GHMs use different river routing schemes (detailed in table S1) and the spatial coordinates of stations with observational data often do not match the coarser gridded river networks (Müller Schmied and Schiebener 2022). To ensure comparability, we base the selection of stations on a previously published report that analysed how well ISIMIP GHMs’ routing schemes align with the location of a selection of GRDC stations and are thus suitable for evaluating river discharge (Müller Schmied and Schiebener 2022). 1096 out of 1509 stations were found to be compatible (Müller Schmied and Schiebener 2022). Of those, we select stations for which daily data was available for at least five years, and we only retain years with less than 10 d of missing data, which are the same criteria as used by Zhao et al. (2017). In the end, we use 644 stations for model evaluation (figure 1), corresponding to 106 basins according to HydroBASINS level 3 from the HydroSHEDS database (figure S1, Lehner and Grill 2013).

2.3. Model evaluation
We analyse the extent to which model simulations can reproduce variability in daily discharge and maximum annual discharge. For both these metrics, and for each station and each GHM, we compare observed discharge to discharge routed with the model’s internal routing scheme, and discharge routed with CaMa-Flood, respectively. To evaluate model performance, we calculate the Kling–Gupta efficiency (KGE) and its three components: correlation ($r$), bias ratio ($\beta$), and variability ratio ($\gamma$) (Kling et al. 2012). The KGE is used to evaluate hydrological models (e.g. Krysanova et al. 2017, Veldkamp et al. 2018) and measures the ability of a model to reproduce observed values. It is calculated by equally weighing the three components, where $r$ is the linear correlation between observed and simulated values, $\beta$ is the ratio of the mean simulated value to the mean observed value, and $\gamma$ is the ratio of the standard deviation of simulated values to the standard deviation of observed values (Knoben et al. 2019). The KGE metric is dimensionless. We calculate the KGE for daily and maximum annual discharge for each model and each station.
For daily discharge, we also evaluate timing differences by testing whether shifting the simulated time series by up to 31 d (back- and forward) improves correlation, as has been suggested by Zhao et al (2017). For maximum annual discharge, we quantify which part of the distribution is over- or underestimated. For each station, we first identify the lower 50% of observational values and count the years where the simulated values are lower or higher, and then we do the same for the upper 50% of observed values.

To assess the capability of models to simulate low flow, we first determine the tenth percentile of daily discharge, i.e. the amount of discharge that is exceeded 90% of the time (Q90, Gosling et al 2017, Krysanova et al 2017). We then divide the difference between simulated tenth percentile and observed tenth percentile by observed tenth percentile. Thus, a value of zero implies a perfect match between observation and simulation, negative values imply that the simulated low flow is too low, while positive values indicate that the simulated low flow is simulated as too high. This is implemented for discharge routed with the model’s internal routing scheme, and discharge routed with CaMa-Flood.

To depict spatial heterogeneity in model performance, we calculated for each model the mean low flow index and mean KGE for maximum annual discharge for each of the 106 basins (figure S1, Lehner and Grill 2013).

2.4. Station characteristics

We use station characteristics data from the Global Streamflow Indices and Metadata Archive (GSIM) that provides metadata for more than 30 000 stations (Do et al 2018, Gudmundsson et al 2018). GSIM data is only available for a subset of stations in this study. We use the following catchment properties from the database (number of stations with data available in parentheses): clay content (432), drainage density (379), elevation (436), irrigated area [%] (436), nightlight development index (176), number of dams upstream (436), population count (435), population density (435), sand content (432), silt content (433), slope (436), storage volume (total upstream storage volume, 436), and topographic index (436). The upstream catchment area from the gauge is taken from the GRDC database. To characterize how dry or wet the area is where a station is located, we used data from the Global Aridity Index and Potential Evapotranspiration Database (Zomer et al 2022). This aridity index is defined as the ratio of precipitation to potential evapotranspiration and is unitless (Zomer et al 2022). Both, catchment area and aridity index are available for all stations.

To further investigate model performance at (sub-)arid stations, we first identified all stations with an aridity index less than or equal to 0.5 (Zomer et al 2022). For these 236 stations, we calculated the KGE and its three components for daily discharge, maximum and mean annual discharge, mean monthly and long-term mean monthly discharge.

3. Results

3.1. Daily discharge

In general, models can reproduce the variability of daily river discharge at most stations (e.g. time-series plot with corresponding KGE values in figures S2 and S3). The median KGE across all stations ranges from −0.43 for model WEB-DHM-SG to 0.46 for WaterGAP2-2e (figure 2(a), table S2). The median correlation ranges from 0.17 for the model CLASSIC to 0.68 for WaterGAP2-2e (figure 2(b), table S2). Most models (seven out of nine) tend to overestimate daily discharge at a majority of stations (bias ratio \( \beta \) larger than one), but there is a large variation in the magnitude of overestimation between models, as well as between stations (figure 2(c), table S2).

Results for the variability ratio (\( \gamma \)) are mixed, with two (out of nine) models overestimating variability (e.g. HydroPy), while the remaining models tend to underestimate variability (e.g. MIROC-INTEGRAL, figure 2(d), table S2).

Models tend to perform well at stations in wetter areas (i.e. higher aridity index, table S3, figure 3), and less well at stations at higher elevations (table S3, figure 4), while in drier areas and at lower elevations there is a large spread across stations including good and poor model performance. The bias ratio tends to be larger at stations in drier areas (figure S4, table S5). For several models (six out of nine), the variability ratio is larger at higher catchment elevations (table S6, figure S5) and steeper slopes (table S6, figure S6). Stations for which models perform best (i.e. highest KGE and correlation across all models) are, for example, located at the Mekong River in Southeast Asia, and the Amazon River in South America.

Shifting the time series can lead to improved correlation (figure S7), but the more shifting is required the worse model performance of shifted time-series is (figure S8).

Regarding the use of river routing schemes, discharge routed with CaMa-Flood performs better for three out of seven models, especially for ORCHIDEE-MICT and WEB-DHM-SG (figure 2(a)). For these cases, KGE and especially its first component (correlation) is higher (figures 2(a) and (b)), but at the same time bias and variability are more strongly overestimated (figures 2(c) and (d)). In contrast, for HydroPy and to a lesser extent also for CWatM, discharge routed with CaMa-Flood tends to be worse compared to the model’s internal routing scheme.
Figure 2. Evaluation of daily discharge routed by the model’s internal routing scheme (Mir) and CaMa-Flood (CaMa) across 644 stations for Kling–Gupta efficiency (KGE) and its three components. JULES-W2 provides discharge routed with two routing schemes (Mir, Mir2, details in table S1). Thick lines: median, box: first and third quartile, whiskers: 10th and 90th percentile. Details table S2.

3.2. Low flow
Most models (six out of nine) do not capture the strength of low flow events, i.e. their magnitude of low flow is simulated too high (figures 5, S9 and S10). However, HydroPy and JULES-W2 are often close to observed low flow, while WEB-DHM-SG tends to underestimate low flow (figure S9). Consistent with the above results, low flow is overestimated in arid regions, especially in the Australian and African basins (figures 5 and S10). The exception is WEB-DHM-SG, which tends to underestimate low flow in these regions as well.

3.3. Maximum annual discharge
Model performance is often worse for maximum annual discharge compared to daily discharge (seven of nine models, figures 6, S11 and S12). The median KGE of annual maximum discharge across all stations ranges from −0.94 for WEB-DHM-SG to 0.24 for WaterGAP2-2e (figure 6(a), table S8), and median correlation ranges from 0.31 for CLASSIC to 0.58 for WaterGAP2-2e (figure 6(b), table S8). There is a tendency to overestimation (six of nine models), and for
Figure 3. Relationship between aridity and model performance for daily discharge routed by the model’s internal routing scheme (Mir) and CaMa-Flood (CaMa) across 644 stations. The aridity index is the ratio of precipitation to potential evapotranspiration (unitless, lower values correspond to more arid conditions, details in Zomer et al 2022).

Figure 4. Relationship between mean catchment elevation and model performance for daily discharge routed by the model’s internal routing scheme (Mir) and CaMa-Flood (CaMa) across 436 stations.

Five models, river routing with CaMa-Flood leads to a stronger overestimation compared to the model’s internal routing scheme (figure 6(c), table S8). The maximum annual discharge is more strongly overestimated at stations in dry areas (figure S13, table S11). Spatial analysis shows that eight out of nine models have a low KGE for most African basins, but a high KGE for most basins in Asia (figures 7 and S14). At the same time, performance for the other regions is more heterogeneous.

The analysis of which part of the distribution of annual maxima is over- or underestimated shows that five models (of nine) overestimate low and high maximum discharge (figure 8). Except
for MIROC-INTEG-LAND, discharge routed with CaMa-Flood tends towards a stronger overestimation when compared to the model’s internal routing scheme. There are some cases where models overestimate low values of annual maximum discharge and underestimate high values. In those cases, the average correlation between observed and simulated discharge is around 0.5. There are only individual stations where models consistently overestimate low values of annual maxima and underestimate high values (example in figures S15 and S16).

3.4. Arid stations
Analysis of model performance across arid stations shows that all models perform better for mean monthly and mean annual discharge than for daily discharge due to a higher correlation coefficient (tables S13–S17, figures S17–S21). The correlation is highest for all models for long-term monthly discharge (table S15, figure S19), suggesting that models can reproduce the seasonal flows.

4. Discussion
We demonstrate that in general, the evaluated models are able to reproduce observed time series of daily and maximum annual discharge. The performance of ISIMIP3a models is similar to evaluation studies of earlier simulation rounds. For example, Veldkamp et al (2018) estimated for monthly discharge simulated by ISIMIP2a models a median correlation of more than 0.6, and Hou et al (2023) derived very similar results for monthly runoff simulated by ISIMIP2a GHMs.

In line with previous studies (Veldkamp et al 2018), the bias ratio is the evaluation parameter that contributed most to a reduced performance in KGE. We find that most models overestimate discharge and that the overestimation tends to be stronger for maximum annual discharge. For ISIMIP2a models, Hattermann et al (2017) also identified that models tend to overestimate monthly discharge, and similarly Zaherpour et al (2018) found that models overestimate extreme runoff. This suggests that applications further down the modelling chain, for example, for flooded areas or flood damages could be affected as well as other sectors such as agriculture or energy production. This is in line with a study showing that simulations with some GHMs led to an overestimation of flooded areas (Mester et al 2021).

Our further investigation of model performance for maximum annual discharge reveals that both low and high values are either over- or underestimated (figure 8) as has been found by Zaherpour et al (2018) for ISIMIP2a models. Thus, contrary to the suggestion by Mester et al (2021), we identify only a few cases where a model overestimates low values and underestimates high values, i.e. a ‘too flat’ simulation. This implies that we find no evidence for the proposal by Mester et al (2021) that the flood return period by GHMs is simulated too short.

Our evaluation setup of including seven GHMs with discharge routed by both, the model’s internal routing scheme and CaMa-Flood, allows for investigating the role of the routing scheme. While Zhao et al (2017) detected that routing with CaMa-Flood led to a lower multi-year mean maximum discharge (ISIMIP2a models), we here find that CaMa-Flood routing results in a higher proportion of years being overestimated for maximum annual discharge (figure 8). Also, except for WEB-DHM-SG, routing with CaMa-Flood has a higher bias ratio than the model’s internal routing (figure 6(c)). Other studies found mixed performances for CaMa-Flood. Yang et al (2019) demonstrated that CaMa-Flood performs well for the amplitude of peak discharge but not the
Figure 6. Evaluation of maximum annual discharge routed by the model’s internal routing scheme (Mir) and CaMa-Flood (CaMa) across 644 stations for Kling–Gupta efficiency (KGE) and its three components. JULES-W2 provides discharge routed with two routing schemes (Mir, Mir2, details in table S1). Thick lines: median, box: first quartile and third quartile, whiskers: 10th and 90th percentile. Details table S8.

timing. In another study, the magnitude of 100 years floods was overestimated in Western and Central USA but underestimated in Eastern USA (Devitt et al 2021). However, in these studies, either CaMa-Flood was used as the only routing model (Yang et al 2019) or combinations of different GHMs, routing schemes, and climate forcing data were compared (Devitt et al 2021). Therefore, it is not clear which components of the modelling chain contribute errors. Our finding that CaMa-Flood tends to overestimate peak discharge, especially at higher elevations with steep slopes, suggests that too much water is transported in the river and floodplain during peak discharges. This may be explained by the fact that evapotranspiration over floodplains and transmission losses are not included in CaMa-Flood (Zhao et al 2017). In addition, routing by CaMa-Flood could be too efficient (i.e. water reaches a station too quickly), suggesting that flow velocity in the river channel or floodplain is overestimated. Unlike kinematic wave approaches, CaMa-Flood includes a diffusion term that allows the wave to spread spatially. While capturing the spatial variability of flow velocity within the channel, the diffusive wave equation tends to concentrate the flow more rapidly in the case of rapid changes in flow, e.g. due to sudden changes in channel slope or rapid increases in discharge. As a result, peak discharges may be overestimated, particularly in higher elevation
areas. In addition, flow velocity is influenced by river channel and floodplain characteristics, some of which are derived empirically in CaMa-Flood. This leads to uncertainties and potential biases in the estimation of peak discharges (Yamazaki et al 2011).

When analysing which catchment properties are linked to model performance we find that two natural features, and not anthropogenic ones, are especially relevant. First, it is well established that hydrological models perform less well in arid and semi-arid regions (Zaherpour et al 2018, Hou et al 2023). We find that this is also the case for ISIMIP3a GHMs (figure 3). Dry areas have a low runoff coefficient meaning that a large proportion of the precipitation evaporates, and thus a small underestimation of evapotranspiration by a model can lead to a strong overestimation of discharge (Hattermann et al 2017). As we find that discharge is more strongly overestimated in arid areas compared to humid areas (figure S4), this could mean that evapotranspiration may be underestimated particularly in arid areas. Given the relevance of hydrological modelling for projecting climate change impacts on drought prevalence in (semi-) arid areas (Wang et al 2022), reducing...
bias in simulating discharge under dry conditions is an important area of future model improvement. However, as we find that the models show a high correlation for long-term mean monthly discharge, GHMs could be used to analyse changes in seasonal flows in (semi-) arid areas.

Second, we find that models perform less well at higher elevations, as has been shown for ISIMIP2a (Yang et al 2019). This suggests that cold dynamics, i.e. glaciers, permafrost, and snowmelt, are important hydrological processes that are not yet adequately included in many models and are thus a further relevant area of model development (Gädeke et al 2020).

As the output of GHMs is used to answer a range of different research questions including the impact of climate change on various aspects of the water cycle (Krysanova et al 2020), it is important to assess additional hydrological variables. For example, there is ongoing work evaluating terrestrial water storage and soil moisture simulated by GHMs from ISIMIP3a (Tiwari et al under review). A next step would be to validate other GHM output variables, such as evapotranspiration and groundwater recharge, as has been done in ISIMIP2 assessment studies (Wartenburger et al 2018, Pokhrel et al 2021, Gnann et al 2023).

We conclude that our results help to identify areas where we have greater confidence in the ability of GHMs to simulate hydrological processes, particularly in humid areas, at low elevations and in areas with strong and regular seasonality. Stations in these areas are therefore likely to be best suited to studies aimed at attributing historical climate change based on the output of the investigated GHMs. We also identify areas where improved process simulation is needed, i.e. evapotranspiration, transmission losses and cold dynamics.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://data.isimip.org/.

Acknowledgments

This research has received funding from the German Federal Ministry of Education and Research (BMBF) under the research projects QUIDIC (01LP1907A) and ISIAccess (16QK05). Support for this publication was provided by COST Action CA19139 PROCLIAS (PROcess-based models for CLimate Impact Attribution across Sectors), supported by COST (European Cooperation in Science and Technology; www.cost.eu).

ORCID iDs

Stefanie Heinicke (https://orcid.org/0000-0003-0222-5281)
Jacob Scheve (https://orcid.org/0000-0001-9455-4159)
Simon N Gosling (https://orcid.org/0000-0001-5973-6862)
Hannes Müller Schmied (https://orcid.org/0000-0001-5330-9923)
Sandra Zimmermann (https://orcid.org/0009-0005-0130-0528)
Matthias Mengel (https://orcid.org/0000-0001-6724-9685)
Inga J Sauer (https://orcid.org/0000-0002-9302-2131)
Peter Burek (https://orcid.org/0000-0001-6390-8487)
Jinfeng Chang (https://orcid.org/0000-0003-4463-7778)
Sian Kou-Giesbrecht (https://orcid.org/0000-0002-4086-0561)
Manoli Grillakis (https://orcid.org/0000-0002-4228-1803)
Naota Hanasaki (https://orcid.org/0000-0002-5092-7563)
Aristeidis Koutoulis (https://orcid.org/0000-0002-2999-7575)
Kedar Otta (https://orcid.org/0000-0002-2540-9879)
Yusuke Satoh (https://orcid.org/0000-0001-6419-7330)
Tobias Stacke (https://orcid.org/0000-0003-4637-5337)
Katja Frieler (https://orcid.org/0000-0003-4869-3013)

References

Boulange J, Yoshida T, Nishina K, Okada M and Hanasaki N 2023 Delivering the latest global water resource simulation results to the public Clim. Serv. 30 100386
Devitt L, Neal J, Wagener T and Coxon G 2021 Uncertainty in the extreme flood magnitude estimates of large-scale flood hazard models Environ. Res. Lett. 16 064013
Frieler K et al 2024 Scenario setup and forcing data for impact model evaluation and impact attribution within the third round of the inter-sectoral model intercomparison project (ISIMIP3a) Geosci. Model Dev. 17 1–51

Gädeke A et al 2020 Performance evaluation of global hydrological models in six large Pan-Arctic watersheds Clim. Change 163 1329–51

Gnann S et al 2023 Functional relationships reveal differences in the water cycle representation of global water models Nat. Water 1 1079–90

Gosling S N et al 2017 A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C Clim. Change 141 577–95

Gudmundsson I et al. 2021 Globally observed trends in mean and extreme river flow attributed to climate change Science 371 1159–62

Hattermann F F et al 2017 Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins Clim. Change 141 561–76


Kumar A et al 2022 Multi-model evaluation of catchment- and global-scale hydrological model simulations of drought characteristics across eight large river catchments Adv. Water Resour. 165 104212

Lange S et al 2020 Projecting exposure to extreme climate impacts across six event categories and three spatial scales Earth’s Future 8 2020EF001616

Lange S et al 2021 WFDEiE5 over land merged with ERA5 over the ocean (W5E5 v2.0) ISIMIP Repository (https://doi.org/10.48364/data.isimip.org)

Lange S, Mengel M, Treu S and Büchner M 2023 ISIMIP3a atmospheric climate input data (v1.2). ISIMIP repository ISIMIP Repository

Lehner B and Grill G 2013 Global river hydrography and network routing: baseline data and new approaches to study the world’s large rivers systems Hydro. Process. 27 2171–86

Liu J et al 2017a Water scarcity assessments in the past, present, and future Earth’s Future 5 545–59


Melton J R, Arora V K, Wisserig-Cojoc E, Seiler C, Forttier M, Chan E and Teckentrup L 2020 CLASSIC v1.0: the open-source community successor to the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM)—Part 1: model framework and site-level performance Geosci. Model Dev. 13 2825–50


Müller Schmied H et al 2021 The global water resources and use model WaterGAP v2.2d: model description and evaluation Geosci. Model Dev. 14 1037–79

Müller Schmied H et al 2023 The global water resources and use model WaterGAP v2.2c: description and evaluation of modifications and new features Geosci. Model Dev: Discuss. 1 46


Stacke T and Hagemann S 2021 HydroPy (v1.0): a new global hydrology model written in Python Geosci. Model Dev. 14 7795–816

Thompson J R, Gosling S N, Zahebourn J and Laizé C L R 2021 Increasing risk of ecological change to major rivers of the world with global warming Earth’s Future 9 e2021EF002048

Tiwari A D et al 2021 Review of Simulated River Discharge Data of the ISIMIP Global Water Sector (TRIP) scheme for improved river routing at the global scale Geosci. Model Dev. 14 226–33

Tsihimikras A, Clark D B, Hartley A J, Burke E J, Grillakis M G and Koutoulis A G under review Spatially-varying parametrization of the total runoff integrating pathways (TRIP) scheme for improved river routing at the global scale Advances in Water Resources

Tsilimigkris et al 2022 Review of Simulated River Discharge Data of the ISIMIP Global Water Sector (TRIP) scheme for improved river routing at the global scale Geosci. Model Dev. 14 226–33

Van Vliet M T H 2023 Complex interplay of water quality and water use affects water scarcity under droughts and heatwaves Nat. Water 1 992–4


Wartenburger R et al 2018 Evapotranspiration simulations in ISIMIP2a—evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets Environ. Res. Lett. 13 075001

Yamazaki D, Kanae S, Kim H and Oki T 2011 A physically based description of floodplain inundation dynamics in a global river routing model Water Resour. Res. 47 1–21


Yokohata T et al 2020 MIROC-INTEG-LAND version 1: a global biogeochemical land surface model with human water management, crop growth, and land-use change Geosci. Model Dev. 13 4713–47

Zhao F et al 2017 The critical role of the routing scheme in simulating peak river discharge in global hydrological models Environ. Res. Lett. 13 065015

Zaherpour J et al 2018 Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts Environ. Res. Lett. 13 075003

Zomer R J, Xu J and Trabucco A 2022 Version 3 of the global aridity index and potential evapotranspiration database Sci. Data 9 409