Published in partnership with CECCR at King Abdulaziz University

6

https://doi.org/10.1038/s41612-025-01134-5

## Finer resolutions and targeted process representations in earth system models improve hydrologic projections and hydroclimate impacts

Check for updates

Puja Das<sup>1,2</sup> & Auroop R. Ganguly<sup>1,2,3</sup>

Earth system models inform water policy and interventions, but knowledge gaps in hydrologic representations limit the credibility of projections and impacts assessments. The literature does not provide conclusive evidence that incorporating higher resolutions, comprehensive process models, and latest parameterization schemes, will result in improvements. We compare hydroclimate representations and runoff projections across two generations of Coupled Modeling Intercomparison Project (CMIP) models, specifically, CMIP5 and CMIP6, with gridded runoff from Global Runoff Reconstruction (GRUN) and ECMWF Reanalysis V5 (ERA5) as benchmarks. Our results show that systematic embedding of the best available process models and parameterizations, together with finer resolutions, improve runoff projections with uncertainty characterizations in 30 of the largest rivers worldwide in a mechanistically explainable manner. The more skillful CMIP6 models suggest that, following the mid-range SSP370 emissions scenario, 40% of the rivers will exhibit decreased runoff by 2100, impacting 850 million people.

Earth System Models (ESMs) are critical tools for understanding climate science, supporting climate adaptation, and informing water resources management. The latest generation of these ESMs enables more refined analysis of hydroclimate responses<sup>1–3</sup>, yet significant challenges remain in ensuring the credibility of their projections, particularly for hydrologic cycle components like runoff at stakeholder-relevant spatiotemporal resolutions<sup>4,5</sup>. As water is essential for human sustenance, agriculture, energy production, and ecosystem maintenance, accurate projections of the global hydrologic cycle (GHC) are crucial<sup>6</sup>, especially given the centrality of water to all 17 United Nations sustainable development goals (UN SDGs)7. However, integrating hydrologic and hydroclimate processes into ESMs get complicated by the heterogeneity of the process, data and environment<sup>1,8</sup>, leading to notable gaps in GHC projections<sup>9,10</sup>. Addressing these challenges involves determining whether improvements can be achieved through systematic incorporation of model processes, critical parameterization, and finer resolution. However, the combined impact of these three advancements in watershed hydrology has not been thoroughly analyzed in large watersheds. Here, we show that a targeted advance in each of these categories can lead to improvements that align with a mechanistic understanding of hydroclimatology. This mechanistic perspective refers to how these processes and parameterizations influence hydrologic responses, providing a physically grounded explanation for model improvements. Our analysis of the historic performance of ESMs over the last two generations reveals that mean runoff projections in the current generation have improved statistically significantly over the previous generation, with 95% confidence limits, particularly in the larger watersheds of the world. Based on our analysis, countries with low GDP per capita and low Human Development Index (HDI) are being impacted by decreasing runoff with heavy population, and future projections also indicate that 40% of the world's largest river basins could experience decreasing runoff. Our analysis suggests that further improvements in runoff projections are achievable and may help in the analysis of hydroclimate impacts, inform the design of water resources infrastructures, policies and interventions, and enable risk-informed decisions to mitigate the impacts.

The primary hypothesis of this study is that targeted improvements in ESMs can enhance their ability to generate critical information for water resource management in the largest river basins in the world. We evaluate runoff simulations from the latest generation of ESMs, specifically,

<sup>1</sup>Sustainability and Data Sciences Laboratory, Northeastern University, Boston, MA, USA. <sup>2</sup>The Institute for Experiential AI and Roux Institute, Northeastern University, Boston, MA, USA. <sup>3</sup>Pacific Northwest National Laboratory, Richland, WA, USA. <sup>(C)</sup> e-mail: a.ganguly@northeastern.edu

ensembles members from the Coupled Model Intercomparison Project version 6 (CMIP6)<sup>11</sup>. While prior research has assessed CMIP6 performance for runoff generation and its advancements over the previous generations, specifically CMIP5<sup>12-15</sup>, corresponding underlying factors driving these variations remain understudied. Recent literature even suggests a possible degradation in the performance of CMIP6 models in some cases<sup>13,16,17</sup>, even raising questions about whether these models are approaching their current limits<sup>18</sup>. Analogies can be drawn from the field of weather forecasting, where periods of relative stagnation in numerical weather prediction modelling were followed by improvements, such as in recent years through the use of artificial intelligence<sup>19-21</sup>. The rationale for our hypothesis regarding ESM performance is partially based on these analogies. ESMs simulate physical processes (e.g., atmosphere, ocean, land, sea ice) at varying resolutions and may include optional biophysical and biogeochemical components using diverse parameterizations<sup>22</sup>. Previous studies have shown that improved resolution refines dynamical and physical parameterizations across atmospheric, oceanic, land, and sea ice systems, enhancing overall model coupling<sup>23-28</sup>. Additionally, numerous studies from different modelling groups have shown that comprehensive processes and enhanced parameterization are crucial for overall model improvements<sup>23,29-32</sup>. Despite the intuitive understanding of the importance of these elements, our study attempts to address a key gap by systematically investigating their roles by examining the physics, parameterizations, and resolutions that the process models have incorporated for runoff projections. We explore which combinations of parameterizations and resolutions yield the most accurate hydrologic projections, providing valuable insights for future model development and refinement.

Here, we focus on assessing runoff in major river basins, which are crucial due to their relatively massive scales and notable impacts on human populations<sup>33,34</sup>. Our findings show that as of 2020, approximately 2.8 billion people reside within the larger river basins. Moreover, due to the coarse resolution of the ESMs, many land surface processes are difficult to resolve in small river basins, so we have selected 30 large river basins based on discharge to capture these effects more accurately. We evaluated the performance of both the multi-model ensemble (MME) and individual models from CMIP6 and CMIP5, focusing on their statistical alignment with reference runoff datasets and the quantification of associated uncertainties. The runoff projections of CMIP6 and CMIP5 are compared using gridded runoff from Global Runoff Reconstruction (GRUN)35 and ECMWF Reanalysis V5 (ERA5)<sup>36</sup> as benchmarks. Furthermore, we explore the potential impacts of future runoff changes under two different Shared Socioeconomic Pathways [SSP-1: Sustainability and SSP-3: Regional Rivalry], and investigate how future population dynamics may be affected by changes in surface runoff. Additionally, we have highlighted how discrepancies in future projections of the best and worst performing models can influence impact assessments.

#### Results

#### Watershed hydrology and demographics

To understand the relationship between water availability and population growth, we have studied the hydrological and demographic changes in 30 major river basins across the globe from 1970s to 2010s. We used the Global Runoff Reconstruction (GRUN)<sup>35</sup> as gridded ground truth for runoff and the Gridded Population of the World (GPW) data from Columbia University<sup>37</sup> for determining hydrologic and demographic dynamics over the last 40 years. We also focused on the income and HDI data from 2020 to better understand the socioeconomic vulnerabilities in conjunction with changes in runoff. For GDP per capita and HDI, we have used the UNDP Human Development Report<sup>38</sup> and World Bank data,<sup>39</sup> respectively. Figure 1 provides a detailed visualization of the differences and trends in runoff alongside changes in population density as well as population count, GDP per capita, and HDI in 2020. Increases in runoff are observed in only three river basins (Orinoco, Paraná, and Zambezi), potentially due to changes in regional precipitation patterns. However, a widespread decline in runoff is evident in 67% of the world's largest river basins, particularly in tropical regions of Africa and South Asia, where major population increases have been observed. These regions, characterized by low- to middle-income countries with a low HDI, are especially vulnerable. The primary reasons behind the declining runoff in many major river basins are thought to be climate change and human activities<sup>40–42</sup>. Climate change has led to reduced rainfall in several regions, directly impacting water availability. Additionally, human activities, including changes in land use and land classification, have altered the natural flow of water, further reducing runoff <sup>41,43</sup>. This growing population, combined with decreasing runoff, presents a critical challenge for the future, which arguably makes it even more crucial to examine future projections from ESMs to anticipate further changes in runoff and develop adaptive strategies.

### Skill, consensus and uncertainty quantification of current and previous generation of Earth system models

We evaluate the performance of all available CMIP6 models (25) in terms of their historical projections of annual runoff from 1960 to 2005. We compare the MME mean, median and range of variability across 30 river basins against two reference datasets: Global Runoff Reconstruction and Reanalysis. We have listed all models used in this study in supplementary table S1 with their modelling centers and grid sizes. Our model selection prioritize diversity, including all available models that provide runoff data, while avoiding redundancy by excluding multiple models from the same modeling institute. Additionally, we analyze the MME mean from all available CMIP5 ESMs to provide a comparative framework. Figure 2 shows the MME mean and median annual runoff and the range of model variability from CMIP6 and CMIP5 generation along with reference runoff for the major river basins.

Our findings reveal that CMIP6 models show statistically significant improvements over CMIP5 in simulating historical mean runoff, as evaluated against both ERA5 and GRUN reference datasets. A statistical comparison using the Wilcoxon signed-rank test<sup>44</sup> demonstrates that CMIP6 models consistently outperform CMIP5 models in terms of Kling-Gupta Efficiency (KGE) across all 30 river basins, with p-values < 0.0001, indicating significance at the 95% confidence level. This consistent improvement across diverse hydrological contexts underscores the advancement in simulation skill achieved by the latest generation of ESMs. However, it is important to note that CMIP6 models exhibit a higher spread in their runoff projections compared to CMIP5, particularly in high-discharge river basins such as the Amazon, Congo, Ganges, Brahmaputra, Orinoco, and Rio De La Plata (Fig. 2, Supplementary Table S2). This increased spread suggests that the addition of more models introduce more variability in certain contexts. In other words, CMIP6 MME have higher skills, but models from CMIP5 generation have higher consensus. River basins with lower discharge generally show a smaller spread in CMIP6 model projections, although there are exceptions, such as the Irrawaddy, Mekong, and Yangtze rivers.

To further assess the credibility of CMIP6 improvements over CMIP5, we analyzed a subset of 11 models available in both generations and estimated their mean annual runoff (supplementary figure S1). This subset was selected to evaluate whether developments from CMIP5 to CMIP6 led to improvements in runoff simulations. Our results indicate that CMIP6 performs better than CMIP5 in the majority of river basins, demonstrating advancements in model representation of hydrologic processes. Additionally, we observe that the range of variability is reduced in CMIP6.

To systematically evaluate model performance, we employed four metrics—Percent Bias (PBIAS), Nash-Sutcliffe Efficiency (NSE), Kling-Gupta Efficiency (KGE), and Pearson's Correlation Coefficient (CC)—to compare runoff simulations from these 11 models against reference datasets. The evaluation (supplementary figure S2) reveals that CMIP6 models generally outperform CMIP5 models across most river basins, as indicated by the box plots. These results highlight improved accuracy, reduced bias, and enhanced overall performance in CMIP6 models, demonstrating superior skill in replicating historical runoff across diverse hydrological contexts.

Furthermore, there are inherent concerns regarding the role of internal climate variability in model simulations, particularly when only a single

ensemble member per model is used. Since each ensemble member represents a different realization of climate variability, relying on a single run may not fully capture the range of possible outcomes. To address this, we selected three models from the CMIP5 and CMIP6 generations and analyzed five randomly selected initial conditions per model (supplementary material section A and supplementary figure S3). Our findings reveal that when multiple ensemble members are considered, the range of variability is lower in multiple initial condition ensemble (MICE) than multi model ensemble (MME). This suggests that part of the increased spread in CMIP6 projections arises from structural differences between models rather than internal variability alone.

Lastly, to quantify the uncertainty in runoff projections across river basins we have analyzed the variability within the CMIP5 and CMIP6 model ensembles, implicitly assuming that multi-model variability contributes to uncertainty in the projections. By examining the spread of projections from each model ensemble, we aimed to capture the range of runoff scenarios. In Fig. 3, we present a comprehensive comparison of

uncertainty between CMIP5 and CMIP6 models. Panel A shows a global map of the increases in uncertainty from CMIP5 to CMIP6, categorized into three levels (< 35%, 35-50%, > 50%), with hatching patterns that indicate the status of the capture of reference runoff. Panel B displays violin plots illustrating the distribution of model projections for each river basin. The map reveals distinct spatial patterns in model performance, with CMIP5 models failing to capture reference runoff primarily in South American basins (Amazon, Parana, Orinocco, Rio de la Plata), while CMIP6 models struggle more in higher latitudes of Asia and Europe (Ob, Volga, Danube, Yenisei). We find that CMIP6 models exhibit higher variance in their runoff projections across all basins, reflecting greater uncertainty. This heightened uncertainty in CMIP6 may be attributable to the broader exploration of the model parameter space or a more complex representation of physical processes, potentially leading to a wider range of plausible runoff scenarios. However, despite this increased variance, CMIP6 models performed better in capturing the reference runoff across a greater number of river basins than CMIP5 models when considering all available



# Fig. 1 | Decreasing runoff intersects with increasing and more vulnerable population. Analysis of changes in the 2010s (2005–2014) relative to 1970s (1965–1974) in 30 of the largest river basins suggests decreasing runoff intersecting with increasing population, especially in highly populated low- to middle-income countries with a low Human Development Index (HDI). A Runoff (measured using GRUN data) changes are categorized: diagonal hatching for decreases (< - 5 mm/year), blank for

stable (- 5 to 5mm/year), and vertical hatching for increases (> 5 mm/year). Population changes are marked by blue (decreasing), yellow (stable), and red (increasing), indicating shifts of more than 10 people per square kilometer. Notable declines in runoff are noted in tropical regions of Africa and South Asia, with rising population densities. **B** Global population density, **C** Human Development Index, and **D** GDP per capita in 2020 are shown with varying shades.



Fig. 2 | Improved mean and changed variability for runoff projections in CMIP6 vs CMIP5. Mean and median annual runoff and model variability are compared with reconstruction (GRUN) and reanalysis (ERA5) from 1960 to 2005 for 30 major river basins. Orange and green shaded areas represent the spread of CMIP5 and CMIP6 projections, respectively. Solid lines indicate the MME mean for CMIP5 (orange), CMIP6 (green), GRUN (black), and ERA5 (pink)

and dashed lines indicated MME medians. The latest generation of ESMs (CMIP6) shows significant improvement (at 95%) over CMIP5 in mean runoff for historical projections, when compared against ERA5 and GRUN [confirmed by Wilcoxon signed-rank test (p < 0.00001)]. The spread of model projections in CMIP6 ESMs seems to be much higher, which can be explained by the addition of higher number of models that generate runoff projections.

models from both ensembles for one random initial condition (25 CMIP6 models and 11 CMIP5 models). In simpler terms, the CMIP6 models exhibit better skills while CMIP5 models show better consensus. Notably, when the comparison was restricted to 11 models common to both generations (supplementary figure S4), the uncertainty levels in runoff projections remained comparable, suggesting that the increased variability in

CMIP6 arises primarily from the inclusion of additional models rather than inherent differences in individual model performance.

**Credibility of hydrologic projection**. To assess the credibility of hydrologic projections among CMIP6 models, we developed an aggregated ranking matrix based on the four key performance metrics



Fig. 3 | Improved mean runoff projections but larger uncertainty in CMIP6 in comparison to CMIP5. A Global map showing the spatial distribution of uncertainty increases from CMIP5 to CMIP6 models, with uncertainty increases categorized as less than 35% (blue), 35–50% (yellow), and more than 50% (red). Hatching patterns indicate reference runoff capture status: white indicates both model generations captured reference runoff, dotted pattern shows both failed to

capture references, cross-hatching shows CMIP5 failed to capture references, and

vertical lines show CMIP6 failed to capture references. **B** Violin plots illustrating the variability in model projections for each river basin, with blue and red shaded areas representing the spread of model projections for CMIP6 and CMIP5, respectively. Solid black lines indicate the median variance, while dashed lines represent the 25th and 75th quartiles. CMIP6 models consistently exhibit higher variance in their runoff projections across river basins, indicating greater uncertainty in projections despite improved mean representation.

discussed earlier. These metrics collectively assess different aspects of model performance, specifically, bias (Percent Bias), model efficiency (NSE and KGE), and the strength of the relationship between model predictions and observed data (CC). All performance metrics are calculated using the MME mean and median. For each river basin, all 25 models were ranked for each metric. These rankings were derived by comparing the performance of each model against two runoff reference datasets. To improve the robustness of the ranking and reduce the influence of random variability from a single ensemble member, we selected up to three initial conditions for each model and calculated the multi-initial condition ensemble mean for ranking purposes. For many models, fewer than three initial conditions were available, and in some cases only a single ensemble member could be used. The specific initial conditions utilized for each model are detailed in the supplementary section A. This approach helps mitigate the possibility that a model's ranking is overly influenced by a single ensemble member that happens to match observed historical sequences due to chance rather than inherent model skill. Supplementary Figures S5 and S6 illustrate the rankings of each model for each river basin in comparison to GRUN and ERA5, respectively, with results visualized using color intensity–darker shades indicating better performance. The rankings revealed that no single CMIP6 model consistently outperforms others across all river basins. However, certain models, such as MRI-ESM2, MIROC-ES2L, E3SM-1-0, consistently achieved high ranks, demonstrating robust performance

across multiple river systems. Subsequently, we created an aggregated rank by combining the results from all eight rankings–four metrics compared across two reference datasets<sup>45</sup>. All metrics were given equal weight in the creation of these aggregated ranks. Supplementary figure S7 displays these aggregated rankings. This ranking is valuable for both water resource managers and earth system modelers or data analysts. For water resource managers, the rankings provide insight into which models are more reliable for informing mitigation and adaptation strategies in their respective river basins. For modelers, these rankings highlight which models perform better in specific basins, allowing them to focus on targeted improvements.

According to the aggregated ranks, some models outperform others overall. To understand why certain models excel, we analyzed the interplay between the physical science implementations, the incorporation of critical parameterizations for runoff through the participation of key MIPs, and the spatial resolution of land surface models within ESMs. We plotted the models from best to worst performance and examined the physical processes they modeled and their corresponding resolutions. Figure 4 provides a detailed analysis of 25 CMIP6 models, showing that advancements in physical science processes, incorporation of critical runoff parameterizations through participation in key MIPs, and finer spatial resolutions significantly improve runoff projections. A model with comprehensive physical and biogeochemical representation typically includes eight key processes: Atmosphere, Aerosol, Ocean, Land, Ocean Biogeochemistry, Atmospheric Chemistry, Land Ice, and Sea Ice. We also considered model participation in key CMIP6-endorsed Model Intercomparison Projects (MIPs) relevant to runoff-such as the Cloud Feedback MIP (CFMIP), Land Use MIP (LUMIP), Land Surface Snow and Soil Moisture MIP (LS3MIP), and Global Monsoon MIP (GMMIP)-which, while not parameterizations themselves, guided the development, testing, or refinement of related physical parameterizations within the models<sup>11</sup>.

Our findings reveal that the top-performing models are distinguished by their incorporation of a greater number of these processes and parameterizations, along with finer spatial resolutions and participation in CMIP6 endorsed MIPs. Panel A of the figure displays the total count of processes and participation in CMIP6 ensodres MIPs considered, while Panel B shows the specific processes and MIPs incorporated by each model. Notably, almost all models include the four core processes-Land Biogeochemistry, Atmosphere, Ocean, and Sea Ice-so these are not highlighted in Panel B. MIROC-ES2L has performed well despite having 250 km resolution, with the incorporation of critical parameterizations. Interestingly, CanESM5 performed well despite its coarser 500 km spatial resolution, likely due to its comprehensive inclusion of physical processes. On the other hand, models like GISS-E2-1-G and CMCC-CM2-SR5, despite having finer resolutions (100 km) and extensive parameterizations, underperformed due to the absence of critical elements such as Ocean Bio-geo Chemistry and Land Ice in their simulations.

This analysis underscores the potential importance of both resolution and the breadth of process representation in producing credible hydrologic projections. However, it is important to emphasize that these relationships are correlative rather than causal. While structural characteristics appear to align with model performance, isolating the specific effects of individual processes or configurations would require targeted sensitivity experiments or controlled intercomparison studies.

For greater statistical robustness, we conducted both mean-based and median-based analyses of model performance across river basins and evaluation metrics. A Spearman rank correlation analysis revealed a moderate positive correlation ( $\rho = 0.567$ , p = 0.0032) between meanbased and median-based rankings, indicating that while there is some consistency between methods, the choice of central tendency metric significantly affects model evaluation outcomes. The statistically significant *p*-value (p < 0.01) confirms that this relationship is unlikely to have occurred by chance, supporting the need for both metrics when assessing model performance. Supplementary figure S8 compares these rankings to highlight differences in model behavior. As shown in Fig. S8, several models exhibit shifts in ranking when evaluated using medians rather than means. For example, INM-CM5-0 and TaiESM1 rank substantially higher using median-based evaluation, while models like FGOALS-g3 and EC-Earth3 rank considerably lower. This shift reveals a critical nuance that some models perform well consistently across basins, while others achieve high mean scores due to strong performance in a limited number of regions. Median-based rankings favor consistency and resilience across diverse hydrologic regimes. Notably, while some highranking models (e.g., MRI-ESM2-0) feature extensive physical process representations, others (e.g., INM-CM5-0) perform well despite having fewer parameterizations or lower spatial resolution. These results suggest that model performance is not solely a function of structural complexity, but also of how well specific processes are implemented and calibrated. This complementary analysis underscores the importance of evaluating model skill through multiple statistical lenses, and indicates that targeted improvements to key processes may be more effective than simply increasing model complexity. Such insights are vital for guiding the future development and evaluation of Earth System Models.

Impact on Human Population. We have examined the potential impact on populations under changes in future runoff, focusing on projected trends in 30 major river basins globally. The analysis is based on long-term future projections from CMIP6 MME under a certain shared socioeconomic pathway scenario, where we estimated the mean and trends for surface runoff from 2017 to 2100. CMIP6 models use SSPs, which are more realistic representations of future world<sup>46,47</sup>. There are 5 SSP scenarios and among them SSP126 and SSP245 denotes a greener world in future whereas SSP 370 and SSP 585 has a higher radiative forcing in future<sup>11</sup>. In this study, SSP 370 situation has been considered as this signifies a forcing level familiar to several unmitigated SSP baselines<sup>48</sup> and this corresponds to a 7  $w/m^2$  radiative forcing<sup>11</sup> during the end of the century. Furthermore, when we assessed the number of people impacted by changes in runoff across different SSPs within our study area (Fig. 5A), we found that the SSP370 scenario projects the highest population growth. Panel A of the figure highlights the projected population<sup>49</sup> in the study areas under five Shared Socioeconomic Pathways (SSPs) for the years 2030, 2050, and 2100. Projections shows nearly 40% of the total population will be living in the study areas under SSP3 scenario ('Regional Rivalry') highlighting the urgency of understanding the potential impacts of changing hydrological conditions under such a high-risk pathway.

The projections from MME mean from 21 CMIP6 model under SSP 370 indicate that at the end of the century, approximately 40% of the river basins will experience a decreasing trend in their long-term mean runoff. Panel B of figure 5 depicts the runoff mean in these river basins under the SSP 370 scenario, with population density in, 2100 overlaid. Notably, basins such as the Ganges, Brahmaputra, Irrawaddy, and Yukon are expected to show the highest increasing trends, while the Amazon, Orinoco, Colorado, Nelson, Churchill, Volga, Orange, Murray and Danube basins are projected to see a decreasing trend. The majority of densely populated areas are expected to coincide with regions showing an increasing trend in runoff. This represents a different scenario from the past 40 years (1970-2010), during which a widespread decline in runoff has been observed in many river basins, particularly in tropical regions of Africa and South Asia. However, it is important to note that CMIP6 MME models have a tendency to overestimate runoff; historical analysis shows that 22 out of 30 river basins tend to exhibit runoff overestimation. This introduces substantial uncertainty in the projected increasing trends for the majority of river basins, highlighting the need for cautious interpretation and careful planning in the future.

Furthermore, we have assessed the number of people that will be affected by changes in runoff trends. A decreasing trend in runoff could lead to water scarcity, impacting numerous people. According to population projections for 2100 under the SSP3 scenario, approximately 5 billion people will live in the 30 major river basins. Among this, approximately 260 million people will still be affected by decreasing trends in runoff by 2100. This estimate was derived from a 21-member model ensemble (MME). However, when we



Fig. 4 | Analysis of 25 CMIP6 models shows that advancements in physical science processes, critical parameterizations for runoff, and finer resolution improve runoff projections. A The bar chart shows the count of processes (sky blue) and participiation in key MIPs (violet) included in each CMIP6 model. Models are categorized by their spatial resolution, represented by colored dots-blue for 100 km, green for 250 km, and orange for 500 km resolution. The chart highlights that leading models typically incorporate a greater number of processes, participation in key MIPs, and finer spatial resolutions. B The scatter plot highlights model

differences in process representation, parameterization strategies, and spatial resolution. Leading models are distinguished by their participation in CMIP6endorsed MIPs focused on cloud feedback (CFMIP), land use (LUMIP), and land surface snow and soil moisture (LS3MIP), which promoted improvements in relevant physical parameterizations. A star preceding a model's name denotes its presence in both the CMIP6 and CMIP5 phases. Models are arranged from highest-performing (left) to lowest-performing (right).

consider the 5-member ensembles of the best-performing and worstperforming models (based on historical performance), we observe significant differences in the projected impacts. The best-performing models suggest that 9 out of 30 rivers will experience a decreasing trend, potentially affecting 850 million people. Conversely, the worst-performing models indicate that fewer rivers will face decreasing runoff, impacting 250 million people. This discrepancy highlights the need for careful consideration when interpreting model outputs, as the choice of model can significantly influence projected impacts. Moreover, given the tendency of CMIP6 MME models to overestimate runoff, the actual number of people impacted by decreasing runoff trends could be even higher, underscoring the need for careful consideration in future water resource management and planning.

Lastly, while much of our analysis focuses on the SSP3-7.0 scenario which is chosen for its relevance to high-emissions and high-risk futures-it



Fig. 5 | End of century projections suggest that larger population will be at the risk of water deficiency. 9 of the 30 largest watersheds globally, corresponding to 850 Million people by 2100 according to regional rivalry scenario (SSP3), show decreasing runoff trends according to best performing models. However, historical analysis of CMIP6 MME reveals that 22 of these basins tend to overestimate runoff, indicating even more people could face reduced water availability. A Projected population percentages in the study areas under five Shared Socioeconomic

Pathways (SSPs) for 2030, 2050, and 2100, with the highest population under SSP3. **B** Changes in mean runoff from historical observations to future projections (CMIP6 MME from 21 Models, SSP 370) and trends in river basins by 2100. Vertical hatches indicate increasing runoff (> 0.0005 mm/day); diagonal hatches indicate decreasing runoff (< 0 mm/day). Population density is color-coded, with densely populated areas showing increasing runoff trends. Blue and red dots indicate whether CMIP6 MME historically overestimated or underestimated runoff.

is important to consider a broader range of plausible climate pathways. To explore the potential hydrologic benefits of mitigation, we conducted a complementary analysis using the SSP1-2.6 scenario, which represents a low-emissions future aligned with ambitious climate targets. The CMIP6 models included in this analysis are listed in Supplementary Table S1. Here, Fig. S9 quantitatively compares runoff projections between these scenarios across all 30 basins, revealing significant differences in both magnitude and trend direction. While mean runoff values show moderate differences (4.7% higher on average under SSP3-7.0), the runoff trends exhibit striking contrasts - in 7 out of 30 basins, the scenarios predict opposite trend directions, with most instances showing decreasing trends under SSP3-7.0 but increasing trends under SSP1-2.6. Statistical analysis confirms that under SSP1-2.6, the estimated population exposed to declining runoff by 2100 drops by approximately 41%, from 850 million to 500 million people. The inter-model spread is also notably lower under SSP1-2.6 (standard deviation 32% smaller on average), suggesting higher agreement among models for this lower-forcing scenario. These quantitative differences underscore both the potential hydrological benefits of emissions mitigation and the importance of explicitly characterizing scenario uncertainty when evaluating climate change impacts on water resources and associated adaptation needs.

#### Discussion

Over the past 40 years, runoff has been decreasing in many major river basins worldwide, raising concerns about water availability for human consumption, agriculture, and ecosystems. This decline is particularly pronounced in tropical regions of Africa and South Asia, where population growth and economic development place increasing pressure on freshwater resources<sup>50,51</sup>. Research focused on tropical Africa has found decreasing runoff trends in river basins between 1961-1990, with the most severe reductions occurring in regions experiencing rapid land-use change<sup>52</sup>. Similar patterns have been observed in tropical Asia, where studies have demonstrated runoff decreases in major basins<sup>53,54</sup>. These hydrological changes intersect with growing water stress, as research has found that nearly 80% of the global population already faces high levels of threat to water security, with hot spots concentrated in regions experiencing runoff declines<sup>55</sup>. Several factors contribute to this trend, including deforestation, urbanization, and the conversion of natural landscapes for agriculture. The role of land-use change in exacerbating these trends is substantial; studies have estimated that anthropogenic land cover transformations have altered runoff patterns<sup>56,57</sup>, with deforestation in tropical watersheds increasing peak flows while reducing dry-season baseflow<sup>58</sup>. Given these widespread changes, it is critical to investigate the key drivers influencing runoff projections in Earth System Models (ESMs) and assess how well the latest generation of climate models from CMIP6 captures these complex hydrological dynamics.

Our findings highlight that CMIP6 models demonstrate improvements in skill over CMIP5. However, despite their improved mean runoff representation, CMIP6 models exhibit a higher spread in their projections, particularly in high-discharge river basins such as the Amazon, Congo, Ganges, Brahmaputra, Orinoco, and Rio de la Plata. This increased spread suggests that while advancements in process representation and parameterizations have enhanced the models' ability to capture complex hydroclimatic interactions, they have also introduced more variability due to differing implementations across modeling groups. Consequently, CMIP6 models exhibit greater uncertainty compared to CMIP5, highlighting the trade-off between improved skill and increased variability in hydrologic projections. This higher variability in runoff projections raises important questions about whether the spread stems from underlying climate variability or model structural uncertainties. For example, in African basins (Congo, Zambezi and Nile), Conway et al. found that high interannual rainfall variability translates into even greater amplification in river flow responses, with rainfall accounting for approximately 60-80% of river flow variability<sup>52</sup>. Another study confirmed that model uncertainty dominates total projected uncertainties for runoff, with low-latitude regions of Africa showing the greatest uncertainty<sup>59</sup>. This suggests that the observed hydrological spread in CMIP6 projections likely stems from both enhanced climate forcing variability and more diverse model structures, highlighting the ongoing challenge of balancing improved process representation with reduced uncertainty in hydrologic projections.

In terms of uncertainty quantification, when comparing 25 CMIP6 models versus 11 CMIP5 models, we found that CMIP6 models generally captured reference runoff in more river basins than CMIP5, despite their higher variability. CMIP5 models predominantly failed to capture reference runoff in South American basins, while CMIP6 models struggled more in higher latitudes of Asia and Europe. This spatial pattern suggests regional differences in how model improvements have affected performance. However, the variance in CMIP6 projections was consistently higher across all basins, indicating a tradeoff between improved mean representation and increased uncertainty in future projections. This pattern shifted when we restricted our analysis to a direct comparison of 11 models from each generation. When we compared 11 models from both generations, the CMIP5 models outperformed CMIP6 in terms of skills for a larger number of river basins. This finding is consistent with recent literature<sup>13,17</sup>, which suggests that despite advancements in the CMIP6 models, they do not always outperform CMIP5 models in all contexts, and in some cases, they even perform worse. Guo et al.<sup>13</sup> evaluated CMIP5 and CMIP6 runoff using GRUN, ERA5, and other reference datasets, finding that CMIP5 models sometimes demonstrate better performance in specific river basins.

Similarly, Wang et al.<sup>17</sup> observed that for precipitation simulation - a critical driver of runoff - CMIP5 models occasionally show superior performance over CMIP6 in regional basins. Zhu et al.60 further confirmed this pattern when analyzing terrestrial water storage anomalies, noting that increased model complexity doesn't necessarily translate to improved hydrological representation across all regions. One possible explanation for this is the incorporation of new cloud physics parameterizations in 8 out of the 11 CMIP6 models considered here. These parameterizations were introduced to enhance the physical accuracy of the models, in terms of atmospheric feedbacks. However, this modification is known to contribute to higher climate sensitivity and projected warming<sup>61</sup>, which led the IPCC AR6 to assign less weight to these models in warming projections<sup>62</sup>. It is plausible that these parameterizations also impact hydrological variables, such as precipitation, evaporation, and runoff, resulting in less consistency between CMIP6 runoff projections and observed runoff patterns. In the remaining 14 CMIP6 models, only 3 incorporated the cloud physics parameterization. While these models without the updated parameterization exhibit slightly higher uncertainty, they outperform the CMIP5 models in capturing reference runoff. This suggests that while advancements in model physics are essential, they do not uniformly improve model performance in all contexts. A key caveat here is that the higher uncertainty in CMIP6 does not necessarily imply lower accuracy, as it could reflect a more comprehensive exploration of possible future scenarios. Further studies are needed to better understand the implications of this parameterization on hydrological projections and to refine the models to strike a balance between uncertainty and accuracy in runoff predictions. Furthermore, to better understand the specific impacts of cloud-related parameterizations on hydrologic outcomes, future work could employ targeted sensitivity analyses or controlled single-model experiments. Initiatives such as the Cloud Feedback Model Intercomparison Project (CFMIP) and the Aerosols and Chemistry Model Intercomparison Project (AerChemMIP) offer structured frameworks to isolate and evaluate the effects of physical parameter changes. Controlled experiments focusing on cloud physics schemes, similar to approaches outlined by Zelinka et al.<sup>61</sup>, could help clarify how structural changes influence hydrological projections, providing actionable pathways to advance process representation while maintaining model credibility.

Additionally, we developed an aggregated ranking matrix to assess the credibility of hydrologic projections among CMIP6 models. To enhance robustness, we selected up to three initial conditions per model where available and calculated the multi-initial condition ensemble mean to mitigate the influence of any single realization. However, the model rankings remain sensitive to initial condition choices, and some models lacked sufficient ensemble members, limiting the statistical robustness of the rankings. Given that three initial conditions are insufficient to fully capture variability arising from internal climate dynamics, future analyses with a larger number of initial conditions and broader model participation would improve ranking reliability.

Despite these limitations, our results show that model rankings are strongly influenced by structural differences, including parameterization choices and resolution, rather than internal variability alone. Models that incorporated a more comprehensive set of physical processes and critical parameterizations (through participation in CMIP6 endorsed MIPs relevant for runoff), generally performed better in simulating historical runoff. Nevertheless, no single model consistently outperformed others across all river basins, reinforcing the importance of using a multi-model ensemble approach when evaluating future hydrologic changes.

While expanding ensemble sizes could, in principle, help reduce sampling uncertainty, our findings suggest that CMIP6 already exhibits substantial inter-model spread that is not fully resolved through ensemble averaging alone. Therefore, rather than focusing solely on increasing ensemble size, future modeling efforts should prioritize strategic ensemble design-selecting structurally diverse models and incorporating performance-based weighting methods-to minimize ensemble noise and bias. Furthermore, advances in artificial intelligence (AI) and hybrid modeling techniques, such as AI-assisted model tuning and ensemble filtering, offer promising pathways to improve ensemble credibility without relying on brute-force increases in model counts. Integrating lessons from these emerging approaches, could enhance both the robustness and efficiency of future hydrologic projection frameworks.

Our study also provides insight into the broader impact of model structural differences and emissions pathways on future projections. The CMIP6 multi-model ensemble suggests that approximately 40% of the world's largest river basins may experience declining runoff by 2100, potentially affecting up to 850 million people under the SSP3-7.0 scenario, based on estimates from the best-performing models. However, comparative analysis with SSP1-2.6 demonstrates that ambitious climate mitigation could significantly reduce both the magnitude of hydrological impacts and the associated uncertainty in projections. Notably, in several basins, the scenarios predict opposite trend directions, emphasizing how different emissions pathways can fundamentally alter projected hydrological trajectories. While substantial discrepancies exist between the best- and worstperforming models under both scenarios, these findings underscore both the value of continued improvement in model process representation and the potential benefits of lower-emissions pathways for global water security.

While our projection offers an important baseline for understanding climate-related water risks, several limitations should be acknowledged. First, these estimates are derived from a single emissions scenario (SSP3-7.0) and are based on ensemble mean values, without representing the full range of inter-model uncertainty. Second, although runoff is a key indicator of surface water availability, it does not capture other essential components of the hydrologic cycle-such as groundwater dynamics, storage changes, and direct human interventions in water systems. Recent studies underscore how these anthropogenic influences, particularly groundwater depletion and land use change, can compound uncertainty and exacerbate water scarcity beyond what is captured by runoff-based analyses alone<sup>63</sup>. These considerations highlight the importance of interpreting population exposure estimates with appropriate caution and emphasize the need for robust model evaluation prior to applying ensemble projections to longterm water planning. Future work should incorporate multiple emissions pathways, explicitly quantify projection uncertainty using confidence bounds or probabilistic methods, and integrate additional hydrologic components to develop more comprehensive and policy-relevant assessments of water security under climate change.

Finally, while our study provides valuable insights into the improvements and limitations of CMIP6 models, it is important to acknowledge certain limitations. First, our evaluation focused primarily on mean annual runoff rather than extreme hydrologic events. Given the increasing frequency of floods and droughts under climate change, future work should examine whether CMIP6 models also improve the simulation of hydrologic extremes. Second, we observed notable differences between the two reference datasets (GRUN and ERA5), which highlights the need for future studies to consider uncertainty stemming from observational datasets when evaluating model performance. A quantitative comparison revealed that the mean absolute difference between GRUN and ERA5 is 0.33 mm/day across all 30 river basins, with larger discrepancies in high-discharge tropical basins (mean difference of 0.63 mm/day) compared to mid-latitude basins (0.38 mm/day) and snow-dominated basins (0.16 mm/day). The Pearson correlation coefficient between these datasets ranges from 0.13 to 0.88 across basins, indicating variable agreement in temporal patterns with some systematic biases. The largest discrepancies occur in data-sparse regions such as the Irrawaddy, Yangtze, and Ganges-Brahmaputra basins, where GRUN's machine learning approach may be less constrained by observational data. Previous studies have documented that ERA5-Land runoff products tend to overestimate discharge volumes in many regions due to positive bias in precipitation inputs<sup>36,64,65</sup>. Conversely, GRUN has been shown to underestimate runoff in regions with sparse gauging networks, particularly in tropical watersheds during high-flow seasons<sup>35,66</sup>. Third, while this study focused exclusively on 30 of the world's largest river basins to accommodate the coarse spatial resolution of current ESMs, we acknowledge that this scope limits the direct applicability of our findings to regional and local water resource decision-making. Future research could address this limitation by evaluating runoff projections in smaller basins using higher-resolution regional climate models (RCMs) or statistically downscaled CMIP products, where available. Addressing these limitations will be crucial for further refining simulation and evaluation of hydrologic projections and improving their applicability to water resource management and climate adaptation efforts.

Furthermore, future work can be benefitted from incorporating additional benchmarking datasets to further constrain model evaluation uncertainty. The LORA (Linear Optimal Runoff Aggregate) dataset optimally combines multiple land surface model outputs and may offer improved performance where GRUN and ERA5 show limitations<sup>67</sup>. For tropical basins with sparse observation networks, the GSCD (Global Streamflow Characteristics Dataset) provides complementary information derived directly from gauge records<sup>68</sup>. Process-based VIC (Variable Infiltration Capacity) global simulations with multiple meteorological forcings could help distinguish between forcing-related and model structural uncertainties<sup>69</sup>.

In conclusion, our study demonstrates that CMIP6 models represent a significant step forward in improving hydrologic projections, but they also introduce greater variability that must be carefully considered in climate impact assessments. The balance between improved skill and increased uncertainty underscores the complexity of earth system modeling and the need for continued refinements in process representation, parameterization schemes, and uncertainty quantification methodologies. Future work should include a detailed examination of these extremes to better inform water resource management and climate adaptation strategies.

#### Methods

**Dataset**. Earth System Model Projections of Runoff data are collected from CMIP6 and CMIP5 models as well as reference runoff at monthly resolution. For the assessment of performance of CMIP6 models, Experiment ID: Historical and SSP126 as well as SSP 370 (future projection) are selected. All available models with runoff projections ware used in this study, discarding the models with missing data, and multiple models from the same institutions are also not considered. For historical projections, 25 CMIP6 models and 11 CMIP5 models are used, while for future projections, 23 for SSP 126 and 21 for SSP370 CMIP6 models are used in this study. In Table S1, the list of models used for each case and the name of their modelling group and resolutions are listed. All the CMIP6 models with historical experiment have data from 1850 to 2014 and CMIP5 models have data from 1850 to 2005. For future experiment, data from 2015 to 2100 is available. Models provide runoffs in *kg/m<sup>2</sup>s* unit and they are converted into mm/day unit for ease of calculation.

For reference surface runoff in historical timescale, this study considered two sources. One is reanalysis dataset and another one is reconstruction based runoff. For reanalysis dataset, runoff datasets are extracted from the European Union's Earth Observation Program (ERA5)<sup>36</sup>. Reanalysis dataset is created from sparsely available observation data combined with data from climate models or remote sensing. Reanalysis datasets are gridded and the grid size (lat x lon) for ERA5 Runoff model is 1800 x 3600. The dataset was gridded using optimal interpolation. Runoff from ERA5 climate reanalysis dataset was available from 1950 to 2021. Reanalysis datasets are also available at the National Oceanic and Atmospheric Administration (NOAA)<sup>70</sup>. However, the spatial resolution for NOAA Runoff simulations are 94 x 192, and was very coarse in comparison to the other reanalysis dataset. For these reasons, NOAA dataset was not considered in this study. For grid-based observations of monthly runoff data, GRUN<sup>35</sup> dataset has been used in this study, which is available from 1902 to 2014 with a grid size of 360 x 720. Preprocessing was performed for aligning the coordinates of all models and datasets. To maintain corresponding time frame, 1960-2005 was considered as the historical study period. For future runoff projections, runoff data for the period of 2017-2100 was selected. The spatial information of the rivers was extracted from the Global Runoff Data Centre (GRDC). which can be downloaded from their website (https://www. bafg.de/GRDC/EN/02\_srvcs/22\_gslrs/gislayers\_node.html).

Population data plays a vital role in estimating the impact of changes in runoff in the river basins. Gridded global population (v4) data<sup>37</sup> is used in this study for the years 1970, 2010 and 2020. This dataset is available at 1 km spatial resolution and the data are stored in WGS84, geographic coordinate system. For future population projection in this study, gridded population data was used from a recent study where, population was projected for 5 different SSP scenarios. The population projection was simulated based on the WorldPop dataset and other related covariates using Random Forest algorithm. This dataset is available at 1 km spatial resolution covering 248 countries for a 5-year temporal resolution starting from 2020 to 2100<sup>49</sup>. We used population projections for SSP3 scenario in 2020, 2030, 2050, 2070 and 2100. For GDP per capita and HDI data we have used, UNDP, Human Development Report (2024) - with minor processing by Our World in Data. and World Bank (2023) - with minor processing by Our World in Data.

**Model Performance Metrics.** To assess the statistical significance of differences in model performance between CMIP5 and CMIP6, we employed the Wilcoxon signed-rank test, a non-parametric alternative to the paired t-test that does not assume normality of the data. The test was applied to compare performance metrics such as the Kling-Gupta Efficiency (KGE), calculated for each model across 28 river basins using both ERA5 and GRUN reference datasets. By evaluating the median differences in model skill across basins, the Wilcoxon test enabled a robust assessment of whether CMIP6 models showed statistically significant improvements over CMIP5.

Percent Bias (PBIAS) and Conditional Bias (CB) are widely used metrics for evaluating climate variable estimates. These metrics provide insight into the systematic deviation of model predictions from observed data. PBIAS is particularly informative as it expresses the bias as a percentage, allowing for a clear interpretation of the model's tendency to overestimate or underestimate the observed values. A low PBIAS is preferred, as it indicates minimal systematic error, while a high PBIAS suggests high model bias, pointing to the model's failure to accurately capture the true values.

PBIAS = 
$$100 \times \frac{\sum_{i=1}^{n} (x_i - y_i)}{\sum_{i=1}^{n} x_i}$$
 (1)

Where, n = number of years,  $x_i =$  reference runoff and  $y_i =$  CMIP Projections.

The Pearson's Correlation Coefficient (CC) is widely used in climate communities for comparison studies. A CC value of 1 denotes the perfect correlation. The formula of CC is as follows:

$$CC = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$
(2)

Where, CC = correlation coefficient,  $x_i$  = reference runoff,  $y_i$  = CMIP Projections,  $\bar{x}$  = mean of the reference runoff values,  $\bar{y}$  = mean of the CMIP Projections.

Nash Sutcliffe Efficiency (NSE) is a commonly used and potentially dependable statistic for evaluating predictive skills of hydrologic variables<sup>71</sup>. NSE value of 1 shows perfect similarity between observed and predicted variable. An NSE value close to 0 means that the mean of the observed values is as good as predicted values. However, negative NSE values mean that the average of the observed data is a better predictor than the simulated data. The formula is shown below:

NSE = 1 - 
$$\frac{\sum_{t=1}^{t} ((x-y)^2}{\sum_{t=1}^{t} ((x-\bar{x})^2)}$$
 (3)

Where, NSE = Nash Sutcliffe Efficiency, x = reference runoff at time t,  $\bar{x}$  = mean of the reference runoff, y = simulated runoff at time t.

NSE value gives an indication of MSE and correlation between observed and predicted variables. But in most cases, measure of variability is also important to understand the efficiency of any variable projections. For this reason, Modified Kling Gupta Efficiency is also estimated in this study as it considered 3 components: correlation coefficient, bias ratio as well as relative variability<sup>72</sup>. Similar to NSE, KGE value of 1 denote perfect simulations. If KGE value is greater than -0.41 then the model simulations are considered to be better than the mean of the observed values.

KGE = 
$$1 - \sqrt{(R-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$$
 (4)

$$R = \frac{Cov(x, y)}{\sigma_x \sigma_y} \tag{5}$$

$$\beta = \frac{\mu_y}{\mu_x} \tag{6}$$

$$\gamma = \frac{\frac{\sigma_y}{\mu_y}}{\frac{\sigma_x}{\mu_x}} \tag{7}$$

Where, KGE = Kling Gupta Efficiency, R = correlation Coefficient,  $\beta$  = bias ratio,  $\gamma$  = relative variability, x = observed runoff,  $\gamma$  = simulated runoff, Cov(x, y) = covariance of reference and simulated runoff,  $\sigma_y$  = standard deviation of simulated runoff,  $\mu_y$  = mean of simulated runoff,  $\sigma_x$  = standard deviation of observed runoff,  $\mu_x$  = mean of observed runoff.

Uncertainty. The uncertainties associated with runoff projections were estimated by calculating the variance across multiple CMIP5 and CMIP6 model ensembles. For each river basin, we computed the mean annual runoff for each model, grouped the results by year, and then calculated the variance across all models to quantify uncertainty. In addition to variance, we visualized the interquartile range (IQR) and median of the uncertainty distributions for each river basin. This provided a comprehensive view of the spread and central tendency of the model outputs, enabling a more detailed understanding of runoff projection uncertainty. In addition to estimating uncertainty through variance, we also assessed whether the model ensembles could capture the reference runoff within the spread of model outputs. For each river basin, we computed the minimum and maximum runoff values across both CMIP5 and CMIP6 model ensembles and compared these ranges to reference runoff values. Specifically, we evaluated whether the reference runoff from GRUN and ERA5 datasets fell within the range of model predictions. This analysis enabled us to assess the performance of each model ensemble in capturing observed runoff variability, providing further insight into model reliability for each river basin.

Aggregate ranking (AR) of models. For each river basin, all 25 models are ranked for each of the four metrics from 1 (best performance) to 25 (least performance). For each model *i*, we compute the sum of the ranks across all four metrics:

 $SR_{i} = Rank_{PBias,i} + Rank_{NSE,i} + Rank_{KGE,i} + Rank_{CC,i}$ (8)

The Aggregate Ranking (AR) for model *i* is defined as:

$$AR_i = SR_i(GRUN) + SR_i(ERA5)$$
(9)

where, SR<sub>*i*</sub>(GRUN) is Sum of rankings for model *i* across all four metrics when compared against the GRUN dataset. SR<sub>*i*</sub>(ERA5) is Sum of rankings for model *i* across all four metrics when compared against the ERA5 dataset.

To obtain a global or overall score, we average the rankings across all basins for each model:

$$AR = \frac{1}{n} \sum_{i=1}^{n} AR_i$$
(10)

where, *n*: Number of basins considered in the evaluation and AR<sub>i</sub>: Aggregate Ranking for model *i*. This formulation allows for a comprehensive evaluation of model performance across multiple metrics and datasets. By aggregating the performance scores over different river basins and reference datasets, we obtain a balanced metric for comparing models within CMIP6.

#### Data availability

CMIP6 and CMIP5 models datasets are available at World Climate Research Program Website hosted by Lawrence Berkeley National Laboratory in the respective addresses (https://aims2.llnl.gov/search/cmip6/) and (https://aims2.llnl.gov/search/cmip5/). ERA5 and GRUN runoff data can be obtained from ECMWF Website (https://climate.copernicus.eu/ climate-reanalysis) and here (https://doi.org/10.6084/m9.figshare.9228176) respectively. Global Runoff Data can be downloaded from (https://www.bafg. de/GRDC/EN/02\_srvcs/22\_gslrs/gislayers\_node.html). Global Population Data is available at (https://sedac.ciesin.columbia.edu/). Future Projection of Global Population Data dataset is available at (https://www.nature.com/ articles/s41597-022-01675-x#Sec9). Human Development Index (HDI) and GDP per capita data are available at (https://ourworldindata.org/grapher/ human-development-index and https://ourworldindata.org/grapher/gdpper-capita-worldbank?time=2022) respectively.

Received: 3 March 2025; Accepted: 13 June 2025; Published online: 01 July 2025

#### References

- 1. Clark, M. P. et al. Improving the representation of hydrologic processes in earth system models. *Water Resour. Res.* **51**, 5929–5956 (2015).
- Lurton, T. et al. Implementation of the cmip6 forcing data in the ipslcm6a-Ir model. J. Adv. Model. Earth Syst. 12, e2019MS001940 (2020).
- Liu, S., Liu, Z., Duan, Q. & Huang, B. The performance of cmip6 models in simulating surface energy fluxes over global continents. *Clim. Dyn.* 61, 579–594 (2023).
- Hulme, M., Pielke, R. & Dessai, S. Keeping prediction in perspective. Nat. Clim. Change 1, 126–127 (2009).
- Kumar, D., Kodra, E. & Ganguly, A. R. Regional and seasonal intercomparison of cmip3 and cmip5 climate model ensembles for temperature and precipitation. *Clim. Dyn.* 43, 2491–2518 (2014).
- Douville, H. et al. Water cycle changes. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (2021).
- Mugagga, F. & Nabaasa, B. B. The centrality of water resources to the realization of sustainable development goals (sdg). a review of potentials and constraints on the african continent. *Int. Soil Water Conserv. Res.* 4, 215–223 (2016).
- 8. Ganguly, A. Science-integrated artificial-intelligence for flooding and precipitation extremes (safe). Tech. Rep., AI4ESP (2021).
- 9. Gardner, L. R. Assessing the effect of climate change on mean annual runoff. *J. Hydrol.* **379**, 351–359 (2009).
- Singh, P., Arora, M. & Goel, N. K. Effect of climate change on runoff of a glacierized himalayan basin. *Hydrological Process.: Int. J.* 20, 1979–1992 (2006).
- Eyring, V. et al. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Dev.* 9, 1937–1958 (2016).
- Hou, Y., Guo, H., Yang, Y. & Liu, W. Global evaluation of runoff simulation from climate, hydrological and land surface models. *Water Resour. Res.* 59, e2021WR031817 (2023).
- Guo, H., Zhan, C., Ning, L., Li, Z. & Hu, S. Evaluation and comparison of cmip6 and cmip5 model performance in simulating the runoff. *Theor. Appl. Climatol.* **149**, 1451–1470 (2022).
- Milly, P. C., Dunne, K. A. & Vecchia, A. V. Global pattern of trends in streamflow and water availability in a changing climate. *Nature* 438, 347–350 (2005).

- Alkama, R., Marchand, L., Ribes, A. & Decharme, B. Detection of global runoff changes: results from observations and cmip5 experiments. *Hydrol. Earth Syst. Sci.* **17**, 2967–2979 (2013).
- Li, J., Huo, R., Chen, H., Zhao, Y. & Zhao, T. Comparative assessment and future prediction using cmip6 and cmip5 for annual precipitation and extreme precipitation simulation. *Front. Earth Sci.* 9, 687976 (2021).
- Wang, D. et al. Performance evaluations of cmip6 and cmip5 models for precipitation simulation over the hanjiang river basin, china. *J. Water Clim. Change* 13, 2089–2106 (2022).
- 18. Maslin, M. & Austin, P. Climate models at their limit? *Nature* **486**, 183–184 (2012).
- 19. Alley, R. B., Emanuel, K. A. & Zhang, F. Advances in weather prediction. *Science* **363**, 342–344 (2019).
- 20. Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).
- Charlton-Perez, A. J. et al. Do ai models produce better weather forecasts than physics-based models? a quantitative evaluation case study of storm ciarán. *npj Clim. Atmos. Sci.* 7, 93 (2024).
- Döscher, R. et al. The ec-earth3 earth system model for the coupled model intercomparison project 6. geosci model dev 15: 2973–3020 (2022).
- Dunne, J. P. et al. The gfdl earth system model version 4.1 (gfdl-esm 4.1): Overall coupled model description and simulation characteristics. *J. Adv. Model. Earth Syst.* 12, e2019MS002015 (2020).
- Shi, Y., Gong, W., Duan, Q. et al. How parameter specification of an earth system model of intermediate complexity influences its climate simulations. *Progress in Earth and Planetary Science* 6 (2019).
- Dawson, A. & Palmer, T. Simulating weather regimes: Impact of model resolution and stochastic parameterization. *Clim. Dyn.* 44, 2177–2193 (2015).
- Adcroft, A. & Hallberg, R. On methods for solving the oceanic equations of motion in generalized vertical coordinates. *Ocean Model*. 11, 224–233 (2006).
- 27. Held, I. et al. Structure and performance of gfdl's cm4. 0 climate model. *J. Adv. Model. Earth Syst.* **11**, 3691–3727 (2019).
- Zhao, M. et al. The gfdl global atmosphere and land model am4. 0/ Im4. 0: 1. simulation characteristics with prescribed ssts. *J. Adv. Model. Earth Syst.* 10, 691–734 (2018).
- Bao, Y., Song, Z. & Qiao, F. Fio-esm version 2.0: Model description and evaluation. J. Geophys. Res.: Oceans 125, e2019JC016036 (2020).
- Caldwell, P. M. et al. The doe e3sm coupled model version 1: Description and results at high resolution. *J. Adv. Model. Earth Syst.* 11, 4095–4146 (2019).
- Collins, W. et al. Development and evaluation of an earth-system model-hadgem2. *Geoscientific Model Dev.* 4, 1051–1075 (2011).
- Dufresne, J.-L. et al. Climate change projections using the ipsl-cm5 earth system model: from cmip3 to cmip5. *Clim. Dyn.* 40, 2123–2165 (2013).
- Mekonnen, M. M., Gerbens-Leenes, P. & Hoekstra, A. Y. Future electricity: The challenge of reducing both carbon and water footprint. *Sci. total Environ.* 569, 1282–1288 (2016).
- Hogeboom, R. J., De Bruin, D., Schyns, J. F., Krol, M. S. & Hoekstra, A. Y. Capping human water footprints in the world's river basins. *Earth's. Future* 8, e2019EF001363 (2020).
- Ghiggi, G., Humphrey, V., Seneviratne, S. I. & Gudmundsson, L. Grun: an observation-based global gridded runoff dataset from 1902 to 2014. *Earth Syst. Sci. Data* **11**, 1655–1674 (2019).
- Muñoz-Sabater, J. et al. Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. data* 13, 4349–4383 (2021).
- for International Earth Science Information Network CIESIN Columbia University, C. Gridded population of the world, version 4.11 (gpwv4): Population count, revision 11. Tech. Rep., NASA Socioeconomic Data

and Applications Center (SEDAC), Palisades, NY (2018). https://doi. org/10.7927/H4JW8BX5. Accessed August 7, 2024.

- UNDP, Human Development Report (2024) with minor processing by Our World in Data. Human development index [dataset]. https:// ourworldindata.org/grapher/human-development-index (2024).
   UNDP, Human Development Report, "Human Development Report 2023-2024" [original data].
- World Bank (2023) with minor processing by Our World in Data. Gdp per capita - world bank - in constant 2017 international \$ [dataset]. https://ourworldindata.org/grapher/gdp-per-capita-worldbank (2023). World Bank, "World Bank World Development Indicators" [original data].
- Senbeta, T. B. & Romanowicz, R. J. The role of climate change and human interventions in affecting watershed runoff responses. *Hydrological Process.* **35**, e14448 (2021).
- Yang, W., Long, D. & Bai, P. Impacts of future land cover and climate changes on runoff in the mostly afforested river basin in north china. *J. Hydrol.* **570**, 201–219 (2019).
- Guan, X. et al. Past variations and future projection of runoff in typical basins in 10 water zones, china. *Sci. Total Environ.* **798**, 149277 (2021).
- Hu, Y. et al. An integrated assessment of runoff dynamics in the amu darya river basin: Confronting climate change and multiple human activities, 1960–2017. *J. Hydrol.* 603, 126905 (2021).
- Woolson, R. F. Wilcoxon signed-rank test. *Encyclopedia of biostatistics*8 (2005).
- 45. Materials and methods are available as supplementary material.
- Song, Y. H., Shahid, S. & Chung, E.-S. Differences in multi-model ensembles of cmip5 and cmip6 projections for future droughts in south korea. *Int. J. Climatol.* 42, 2688–2716 (2022).
- O'Neill, B. C. et al. The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Dev.* 9, 3461–3482 (2016).
- Keller, D. P. et al. The carbon dioxide removal model intercomparison project (cdrmip): Rationale and experimental protocol for cmip6. *Geoscientific Model Dev.* **11**, 1133–1160 (2018).
- Wang, T., Tu, X., Singh, V. P., Chen, X. & Lin, K. Global data assessment and analysis of drought characteristics based on cmip6. *J. Hydrol.* 596, 126091 (2021).
- 50. Wohl, E. et al. The hydrology of the humid tropics. *Nat. Clim. Change* **2**, 655–662 (2012).
- Dai, A. Historical and future changes in streamflow and continental runoff: A review. *Terrestrial Water Cycle and Climate Change: Natural* and Human-Induced Impacts 17–37 (2016).
- Conway, D. et al. Rainfall and water resources variability in subsaharan africa during the twentieth century. *J. Hydrometeorol.* **10**, 41–59 (2009).
- Shah, H. L. & Mishra, V. Hydrologic changes in indian subcontinental river basins (1901–2012). *J. Hydrometeorol.* **17**, 2667–2687 (2016).
- Wang, X. et al. Attribution of runoff decline in the amu darya river in central asia during 1951–2007. J. Hydrometeorol. 17, 1543–1560 (2016).
- Vorosmarty, C. J., Green, P., Salisbury, J. & Lammers, R. B. Global water resources: vulnerability from climate change and population growth. *science* 289, 284–288 (2000).
- Sterling, S. M., Ducharne, A. & Polcher, J. The impact of global landcover change on the terrestrial water cycle. *Nat. Clim. change* 3, 385–390 (2013).
- Gerten, D., Rost, S., von Bloh, W. & Lucht, W. Causes of change in 20th century global river discharge. *Geophysical Research Letters*35 (2008).
- Bruijnzeel, L. forestation and dry season flow in the tropics: A closer look. *Journal of Tropical Forest Science* 229–243 (1989).
- Wu, Y. et al. Hydrological projections under cmip5 and cmip6: Sources and magnitudes of uncertainty. *Bull. Am. Meteorological Soc.* **105**, E59–E74 (2024).
- Zhu, E., Wang, Y., & Yuan, X. Changes of terrestrial water storage during 1981–2020 over China based on dynamic-machine learning model. *J. Hydrol.* 621, 129576 (2023).

- 61. Zelinka, M. D. et al. Causes of higher climate sensitivity in cmip6 models. *Geophys. Res. Lett.* **47**, e2019GL085782 (2020).
- Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W. & Zelinka, M. Climate simulations: recognize the 'hot model'problem. *Nature* 605, 26–29 (2022).
- Nie, W., Kumar, S. V. & Zhao, L. Anthropogenic influences on the water cycle amplify uncertainty in drought assessments. *One Earth* (2025).
- Liu, L., Yi, Y., Jiang, H., Ran, Y. & Chen, D. Era5-land overestimates runoff coefficient but underestimates runoff recession rate in the central tibetan permafrost region. *J. Hydrol.: Regional Stud.* 53, 101792 (2024).
- Xu, C., Wang, W., Hu, Y. & Liu, Y. Evaluation of era5, era5-land, gldas-2.1, and gleam potential evapotranspiration data over mainland china. *J. Hydrol.: Regional Stud.* 51, 101651 (2024).
- Ibarra, D. E., David, C. P. C. & Tolentino, P. L. M. Evaluation and bias correction of an observation-based global runoff dataset using streamflow observations from small tropical catchments in the philippines. *Hydrol. Earth Syst. Sci.* 25, 2805–2820 (2021).
- Hobeichi, S., Abramowitz, G., Evans, J. & Beck, H. E. Linear optimal runoff aggregate (lora): A global gridded synthesis runoff product. *Hydrol. Earth Syst. Sci.* 23, 851–870 (2019).
- Beck, H. E., De Roo, A. & van Dijk, A. I. Global maps of streamflow characteristics based on observations from several thousand catchments. *J. Hydrometeorol.* **16**, 1478–1501 (2015).
- 69. Lin, P. et al. Global reconstruction of naturalized river flows at 2.94 million reaches. *Water Resour. Res.* **55**, 6499–6516 (2019).
- Kistler, R. et al. The ncep–ncar 50-year reanalysis: monthly means cd-rom and documentation. *Bull. Am. Meteorological Soc.* 82, 247–268 (2001).
- McCuen, R. H., Knight, Z. & Cutter, A. G. Evaluation of the nash–sutcliffe efficiency index. *J. hydrologic Eng.* **11**, 597–602 (2006).
- Kling, H., Fuchs, M. & Paulin, M. Runoff conditions in the upper danube basin under an ensemble of climate change scenarios. *J. Hydrol.* **424**, 264–277 (2012).

#### Acknowledgements

The research was supported by DOD SERDP RC20-1183 and partially by NASA Water Resources Program under Grant 21-WATER21-2-0052 (Federal Project ID: 80NSSC22K1138) and the Northeastern University Office of the Provost through AI for Climate and Sustainability (AI4CaS) of the Institute for Experiential AI (EAI). The authors thank current and former members of Northeastern University's Sustainability and Data Sciences Laboratory (SDS Lab) for helpful discussions.

#### **Author contributions**

P.D. and A.R.G. conceptualized the research and defined the problem, P.D. implemented the research and developed the analyses and assessments, P.D. and A.R.G. interpreted the results and fine-tuned the analyses, assessments, and results, P.D. prepared the first version of the manuscript while P.D. and A.R.G. jointly and iteratively developed the final version.

#### **Competing interests**

The authors declare no competing interests.

#### Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-025-01134-5.

**Correspondence** and requests for materials should be addressed to Auroop R. Ganguly.

Reprints and permissions information is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/bync-nd/4.0/.

© The Author(s) 2025