nature water

Analysis

https://doi.org/10.1038/s44221-023-00160-y

Functional relationships reveal differences in the water cycle representation of global water models

Received: 3 March 2023

Accepted: 13 October 2023

Published online: 27 November 2023

Check for updates

Sebastian Gnann[1,2,17] ✉, Robert Reinecke[1,3,17], Lina Stein[1], Yoshihide Wada[4,5], Wim Thiery[6], Hannes Müller Schmied[7,8], Yusuke Satoh[9], Yadu Pokhrel[10], Sebastian Ostberg[11], Aristeidis Koutroulis[12], Naota Hanasaki[13], Manolis Grillakis[12], Simon N. Gosling[14], Peter Burek[5], Marc F. P. Bierkens[15,16] & Thorsten Wagener[1]

Global water models are increasingly used to understand past, present and future water cycles, but disagreements between simulated variables make model-based inferences uncertain. Although there is empirical evidence of different large-scale relationships in hydrology, these relationships are rarely considered in model evaluation. Here we evaluate global water models using functional relationships that capture the spatial co-variability of forcing variables (precipitation, net radiation) and key response variables (actual evapotranspiration, groundwater recharge, total runoff). Results show strong disagreement in both shape and strength of model-based functional relationships, especially for groundwater recharge. Empirical and theory-derived functional relationships show varying agreements with models, indicating that our process understanding is particularly uncertain for energy balance processes, groundwater recharge processes and in dry and/or cold regions. Functional relationships offer great potential for model evaluation and an opportunity for fundamental advances in global hydrology and Earth system research in general.

Global water models—including hydrological, land surface and dynamic vegetation models[1]—have become increasingly relevant for policy-making and in scientific studies. The Sixth Assessment Report[2] of the Intergovernmental Panel on Climate Change draws heavily on results from global water models, which provide information about climate change impacts on hydrological variables including soil moisture[3], streamflow[4], terrestrial water storage[5] and groundwater recharge[6]. Some of these models are already embedded in global water information services to provide information to a wide array of stakeholders, such as the Global Groundwater Information System[7] or the African

[1]Institute of Environmental Science and Geography, University of Potsdam, Potsdam, Germany. [2]Chair of Hydrology, Faculty of Environment and Natural Resources, University of Freiburg, Freiburg, Germany. [3]Institute of Geography, Johannes Gutenberg University Mainz, Mainz, Germany. [4]Climate and Livability Initiative, Center for Desert Agriculture, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [5]International Institute for Applied Systems Analysis, Laxenburg, Austria. [6]Department of Water and Climate, Vrije Universiteit Brussel, Brussels, Belgium. [7]Institute of Physical Geography, Goethe University Frankfurt, Frankfurt am Main, Germany. [8]Senckenberg Leibniz Biodiversity and Climate Research Centre (SBiK-F), Frankfurt am Main, Germany. [9]Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea. [10]Department of Civil and Environmental Engineering, Michigan State University, East Lansing, MI, USA. [11]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany. [12]School of Chemical and Environmental Engineering, Technical University of Crete, Chania, Greece. [13]National Institute for Environmental Studies, Tsukuba, Japan. [14]School of Geography, University of Nottingham, Nottingham, UK. [15]Department of Physical Geography, Utrecht University, Utrecht, Netherlands. [16]Unit Soil and Groundwater Systems, Deltares, Utrecht, Netherlands. [17]These authors contributed equally: Sebastian Gnann and Robert Reinecke. ✉e-mail: gnann1@uni-potsdam.de

Flood and Drought Monitor[8]. Because measurements of many hydrological variables are very sparse and insufficient for large-scale analyses, global water models are regularly used in scientific studies to provide globally coherent estimates of variables such as groundwater recharge and groundwater storage change[9,10]. Global water models are also an integral part of Earth system models, and a realistic representation of the water cycle is essential for simulating the role of water within and across the different components of the Earth system[11].

The Intergovernmental Panel on Climate Change's Sixth Assessment Report[2] concludes from an analysis of currently available global water model projections that 'uncertainty in future water availability contributes to the policy challenges for adaptation, for example, for managing risks of water scarcity'. Whereas some of this uncertainty stems from projected and observed climatic forcing, considerable uncertainty stems from global water models themselves[4,6,12–14]. For instance, Beck et al.[13] found distinct inter-model performance differences when comparing simulated and observed streamflow for ten global water models driven by the same forcing. To illustrate this uncertainty, we show how 30-year (climatological) averages of actual evapotranspiration, groundwater recharge and total runoff vary globally on the basis of outputs from eight models driven by the same forcing (Fig. 1a–c; Methods). We find substantial disagreement among models, as indicated by high coefficients of variation, particularly for groundwater recharge and total runoff. We further show which model deviates most from the ensemble mean and find that there is not one model that consistently deviates the most (Fig. 1d–f). Whereas this analysis cannot tell us which models perform better or worse, it suggests that it is not straightforward to single out a model for a certain flux or a certain region, which warrants a more in-depth evaluation.

Most evaluation strategies compare model outputs to historical observations over the area for which the observation is representative. This can be at the plot (for example, flux towers), catchment (for example, gauging stations) or grid cell (for example, gridded remote sensing products) scale. Such approaches are necessary but not sufficient to robustly evaluate global models[15]. First, these approaches compare simulated and observed values location by location and are therefore limited to potentially improving a model for that location; however, given that large fractions of the global land area are ungauged, we require methods that can extract and transfer information from gauged to ungauged locations[16]. Second, relevant information for model evaluation might not just lie in comparing the magnitudes of simulated and observed values in a single location but rather in how a variable varies along a spatial gradient[17]. And third, comparison with historical observations does not guarantee that a model reliably predicts system behaviour under changing conditions[18]. Rather than evaluating global models in essentially the same way as catchment-scale models, evidence of different large-scale hydrological relationships presents us with an opportunity for a different evaluation strategy that is inherently large-scale but so far rarely exploited.

Towards evaluation using functional relationships

Reviewing the hydrological literature reveals a range of relationships[19] that, if they appear in empirical data, should also appear in models (and vice versa). Such relationships often capture behaviour that is not prescribed by small-scale processes but rather emerges through the interaction of these processes (or model components) at large scales. The perhaps most prominent example is the Budyko framework[20], which describes the long-term partitioning of precipitation into evapotranspiration and streamflow solely as a function of the aridity index. Another example are so-called elasticities of streamflow to changing climatic drivers (for example, precipitation or temperature), which provide an observation-based constraint on climate change effects on streamflow[21,22]. A third example are empirical relationships between annual rainfall and runoff, which can be affected differently by prolonged drought; in Australia, some catchments have shown similar rainfall–runoff relationships before and after the Millennium Drought, while other catchments have transitioned to a new stable state[23]. The search for robust relationships that characterize the functioning of hydrological systems is in itself a great scientific challenge[19], but such functional relationships also provide an excellent yet poorly explored opportunity for the evaluation of global water models.

We define the term function as the actions of (hydrological) systems on the inputs that enter them, such as partition, storage and release of water and energy[24,25]. Accordingly, we define functional relationships as relationships between two or more variables that characterize these functions. Such relationships often focus on forcing, state and response variables that are expected to be causally related (for example, precipitation and runoff), and they can focus on both temporal variability at a single location and (as used here) spatial variability across multiple locations. Functional relationships need not be uniquely defined and are typically characterized by substantial scatter due to other (secondary) controlling variables, local variability or uncertainty.

Whereas functional relationships have been used before to evaluate land surface, forest and Earth system models—for example, by analysing relationships between soil moisture and evaporation and runoff[26–29] or between precipitation and other atmospheric drivers and vegetation productivity[30–32]—their potential for evaluating global water models has not yet been sufficiently explored. The use of functional relationships is currently scattered among the hydrological literature (for example, refs. 33–35) and has not been formalized into an evaluation framework. There is a pressing need to develop a 'theory of evaluation'[36] that does justice to the nature of global models, the purposes for which they are used and their growing relevance for society[37]. Functional relationships have the potential to be a central building block of such a theory of evaluation, and below we show how they can help shed new light on model behaviour.

Here we focus on functional relationships that capture the spatial co-variability of forcing and response variables. Rather than focusing on a process-by-process comparison that can quickly become unmanageable[28], functional relationships can capture emergent patterns and shift the focus to identifying the dominant controls on the variables of interest. Especially the relationships between water and energy availability and the major water fluxes leaving the land surface—evaporation and runoff—have been frequently studied[20,38], providing an excellent starting point for model evaluation. In addition, functional relationships that focus on spatial patterns offer several advantages. First, such relationships are well suited for the analysis of global models due to their spatially distributed nature, which means that these relationships can be readily obtained from comparing values from multiple grid cells. Second, spatial relationships can be calculated based on long-term averages, which for some variables are often the only observations available (for example, for groundwater recharge[39,40]). And third, such relationships can capture how hydrological variables co-vary across large scales and thus offer the potential for model improvement over large areas, including locations that lack observations.

In this analysis, we investigate how long-term averages of two forcing and three response variables co-vary spatially, leading to six variable pairs overall. The forcing variables are precipitation P and net radiation N (the available water and energy, respectively), and the response variables are actual evapotranspiration E_a, groundwater recharge R and total runoff Q (three key water fluxes). We analyse forcing–response relationships based on 30-year (climatological) averages (1975–2004; all in mm per year) from eight global water models (CLM4.5, CWatM, H08, JULES-W1, LPJmL, MATSIRO, PCR-GLOBWB and WaterGAP2) from phase 2b of the Inter-Sectoral Impact Model Intercomparison Project

https://doi.org/10.1038/s44221-023-00160-y

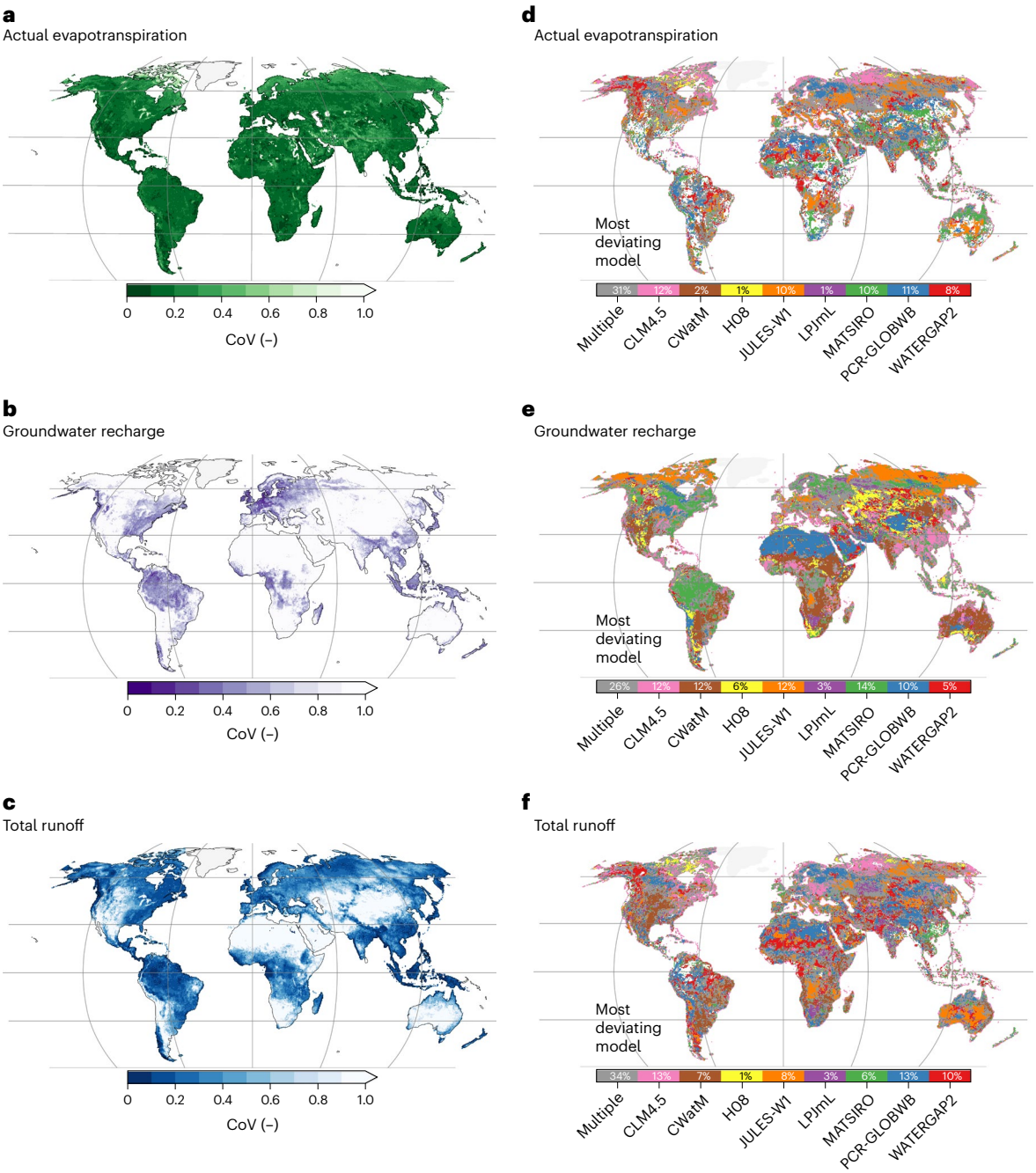


Fig. 1 | Disagreement between global water models for three key water fluxes. **a**–**c**, Left: maps showing the coefficient of variation, calculated per grid cell as the ensemble standard deviation divided by the ensemble mean of eight global water models for different water fluxes: actual evapotranspiration (**a**), groundwater recharge (**b**) and total runoff (**c**). Lighter areas ('blank spaces') indicate high coefficients of variation (CoV) values and thus show where models disagree most. **d**–**f**, Right: maps showing which model deviates most from the ensemble mean for each grid cell for different water fluxes: actual evapotranspiration (**d**), groundwater recharge (**e**) and total runoff (**f**). Dark grey areas in **d**–**f** indicate that multiple models deviate similarly strongly from the ensemble mean. Empty, blank areas in **d**–**f** indicate that no model deviates strongly from the ensemble mean. The percentages shown in **d**–**f** refer to the fraction of grid cells (not land area) covered by each model. Greenland is masked out for the analysis.

(ISIMIP 2b[41]). In addition, we use observational datasets, observation-driven machine learning products and the semi-empirical equation introduced by Budyko[20] to calculate functional relationships between the same variables as for the models as benchmarks (Table 1). To explore regional variability in functional relationships[38], we divide the world into four climatic regions: wet–warm (18% of modelled area), wet–cold (15%), dry–cold (24%) and dry–warm (43%), shown in Fig. 2d. Details can be found in the Methods section.

Disagreement in functional relationships between models

We can visually assess relationships between forcing (P, N) and response variables (E_a, R, Q) by inspecting scatter plots where each point represents one grid cell (or observation); this is shown for precipitation and groundwater recharge in Fig. 2a. We first take a closer look at the shapes of the functional relationships, indicated by the coloured lines in Fig. 2a. Later we will also quantify the strength of the relationships

https://doi.org/10.1038/s44221-023-00160-y

Table 1 | Spearman rank correlations among forcing variables and water fluxes and number of observations based on different observational or observation-driven datasets and the Budyko equation

Flux	Forcing	Source	Nr	Wet–warm (15%)		Wet–cold (23%)		Dry–cold (28%)		Dry–warm (34%)	
				ρ_s	Count	ρ_s	Count	ρ_s	Count	ρ_s	Count
E_a	*P*	Budyko*[20]	1	0.84	m.e.	0.83	m.e.	0.98	m.e.	1.00	m.e.
E_a	*P*	FLUXCOM[43]	2	0.57	m.e.	0.76	m.e.	0.71	m.e.	0.88	m.e.
E_a	*N*	Budyko*[20]	1	0.95	m.e.	0.99	m.e.	0.59	m.e.	0.79	m.e.
E_a	*N*	FLUXCOM[43]	2	0.93	m.e.	0.94	m.e.	0.79	m.e.	0.91	m.e.
R	*P*	MacDonald[40]	3	(0.0)	4	-	0	-	0	0.84	130
R	*P*	Moeck[39]	4	-0.06	234	0.66	83	0.29	100	0.74	4790
Q	*P*	Budyko*[20]	1	0.94	m.e.	0.87	m.e.	0.90	m.e.	0.99	m.e.
Q	*P*	GSIM[47,48]	5	0.62	1259	0.71	1211	0.32	517	0.80	900
Q	*P*	GRUN[49]	6	0.86	m.e.	0.74	m.e.	0.27	m.e.	0.94	m.e.
Q	*N*	Budyko*[20]	1	0.45	m.e.	0.42	m.e.	0.11	m.e.	0.69	m.e.

The percentage of grid cells per climate region is given in brackets. The Budyko equation was forced per grid cell with the same forcing as the models (indicated by *) and thus covers approximately the same extent (except for cells with negative net radiation). The gridded datasets (FLUXCOM, GRUN) are available at the same resolution as the models and thus also cover approximately the same extent (except for non-vegetated areas in the case of FLUXCOM). This is indicated by m.e. for model extent. For datasets without matching precipitation data, we used GSWP3 reanalysis data. Nr corresponds to the numbers used in Fig. 4. The MacDonald rank correlation for the wet–warm region is shown in brackets because of the very small sample size; it is not shown in Fig. 4. Dashes (-) indicate that correlations could not be calculated because no observations were available. ρ_s denotes Spearman rank correlations.

using Spearman rank correlations ρ_s. We limit ourselves to a qualitative discussion, given that fitting an equation would mean that we would have to assume a functional form. We report mean values and slopes (obtained via linear regression) for each region in Supplementary Tables 4–7, which quantitatively support our visual assessment. Figure 3 shows connected binned median values for precipitation and the three water fluxes for all models and observational datasets (Table 1), separated by climate region. A similar plot for net radiation and the three water fluxes is shown in Extended Data Fig. 1.

While the P–E_a relationships look similar in shape, they can differ greatly in magnitude (Fig. 3). They increase rather linearly in dry (water-limited) regions and increase initially in wet (energy-limited) regions and then level off as they reach an energy limit that bounds actual evapotranspiration. The limit differs greatly between models, varying up to about 400 mm per year in wet–warm regions. Because all models are forced with the same total radiation, this difference is related to the way the models translate total radiation into net radiation and how they then use net radiation to calculate actual evapotranspiration. There is no obvious connection between this difference and the different potential evapotranspiration schemes used[42], potentially because the models, while forced with the same climate inputs, differ in the way they parameterize the land surface (for example, land use, soils). In dry regions, actual evapotranspiration is mostly limited by precipitation, a forcing dataset that is the same for all models, resulting in less variability. The Budyko equation and the FLUXCOM[43] dataset suggest, in line with literature estimates[44], that most models underestimate actual evapotranspiration, often greatly so (Supplementary Tables 4 and 5). However, we note that FLUXCOM probably overestimates actual evapotranspiration, especially in dry–warm regions, because it considers only vegetated areas[43]. Overall, the disagreement in modelled actual evapotranspiration, particularly visible in energy-limited regions, suggests substantial differences in the way models estimate the energy available for evapotranspiration.

Most P–R relationships increase monotonically, but the shape, the slope and the threshold at which some models start to produce groundwater recharge are very different (Fig. 3). For instance, in dry–warm regions, some models produce essentially no groundwater recharge even if precipitation is above 1,000 mm per year, while others produce over 200 mm per year. In dry–warm regions, we have by far the most extensive database on groundwater recharge[39,40], and the observations fall (apart from those at very high precipitation values) within the range of the models. In wet–warm regions, we find the largest disagreement between models and observations, which suggest lower (higher) groundwater recharge rates for higher (lower) precipitation. Whereas this shows the benefit of using an ensemble rather than a single model, even a large ensemble spread does not always capture the observed relationships. The large spread further suggests that many models greatly over- or underestimate groundwater recharge rates and consequently greatly over- or underestimate how much groundwater contributes to evapotranspiration and streamflow[45]. These differences in slope, visible for all climate regions, reflect very different spatial sensitivities to changes in precipitation. Whether temporal sensitivities are similar can only be hypothesized given that no global observational dataset with groundwater recharge time series is available but would imply very different responses to projected changes in precipitation.

The P–Q relationships look similar in shape and mostly increase monotonically, especially for wet regions (Fig. 3). The relative differences are larger for dry places, commonly perceived as regions where runoff is more difficult to model[46]. The model and benchmark relationships disagree particularly strongly in dry–cold regions. There, the GSIM[47,48] dataset shows a variable relationship between total runoff and precipitation, whereas the GRUN[49] dataset shows almost no increase with increasing precipitation. Overall, GSIM, GRUN and the Budyko equation indicate, in line with an earlier evaluation[50], that most models produce too much total runoff. This parallels recent findings that Earth system models predict higher runoff increases due to climate change than observations suggest[22]. The overestimation in total runoff is complementary to the underestimation of actual evapotranspiration and shows that most models partition too much precipitation into runoff rather than evapotranspiration.

Diverging dominance of forcing on response variables

To quantitatively compare the strength of the forcing–response relationships, we use Spearman rank correlations ρ_s. A rank correlation close to 1 (or −1) indicates that the spatial variability in the forcing variable almost completely explains the spatial variability in the response variable, as can be seen in Fig. 2a for WaterGAP2. A rank correlation closer to 0 indicates that other factors control the response (for example, other input or model parameters describing the land surface), as can be seen in Fig. 2a for PCR-GLOBWB. We stress that a high correlation is not a measure of goodness of fit. Considerable scatter and correspondingly low correlations might indeed be characteristic for many relationships, and if models overestimate how strongly a forcing

https://doi.org/10.1038/s44221-023-00160-y

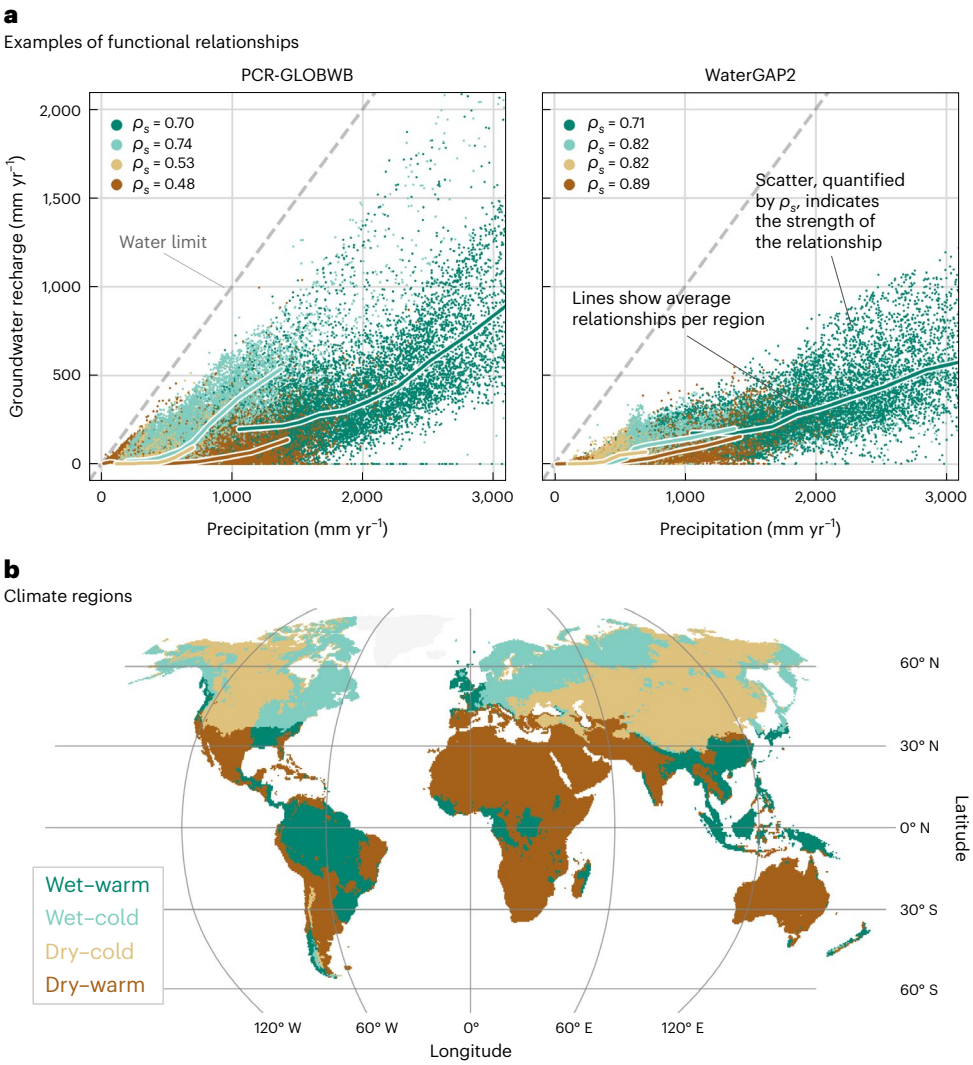


Fig. 2 | Examples of functional relationships. a, Scatter plots between precipitation and groundwater recharge for PCR-GLOBWB and WaterGAP2. Owing to space constraints, we focus on a few examples with differing relationships. Scatter plots for all variable pairs are shown in Supplementary Figs. 15–20. Each dot represents one grid cell and is based on the 30-year average of each flux. Spearman rank correlations ρ_s measure the strength of the relationship between forcing and response variables and are calculated for all grid cells within a climate region. The lines connect binned medians (ten bins along the *x* axis with equal amount of points per bin) for each region. **b**, The climate regions are shown. The grey dashed line shows the 1:1 line, indicating the water limit assuming all water is supplied by precipitation.

variable controls a model output, this also indicates unrealistic behaviour. Calculating rank correlations for all variable pairs, we find that the models differ substantially among each other and in comparison to observations (Fig. 4; rank correlations for all benchmark datasets and models are listed in Table 1 and Supplementary Table 3, respectively).

For precipitation and actual evapotranspiration (Fig. 4a), the models show the same ranking between climate regions and rather small differences in magnitude, indicating that actual evapotranspiration is strongly constrained by the available water in all models. The model-based correlations are higher in dry regions (ρ_s = 0.74–0.98) than in wet regions (0.57–0.83), reflecting water and energy limitations. The Budyko equation assumes complete dependence on aridity (here defined as N/P). It thus predicts higher correlations overall and mainly distinguishes between wet (0.83–0.84) and dry (0.98–1.00) regions but, unlike models and FLUXCOM, not between cold and warm regions. The Budyko equation should thus be seen as a useful comparison but not as the 'correct' model, given that different studies have shown that snow[51], climate seasonality[52], vegetation type[53], inter-catchment groundwater flow[54] and human impacts[55] can affect the long-term water balance beyond aridity.

We find much variability for net radiation and actual evapotranspiration (Fig. 4b). There is no obvious correspondence between the potential evapotranspiration schemes used[42] (for example, Priestley–Taylor for LPJmL and WaterGAP2 or Penman–Monteith for JULES-W1 and CWatM) and the rank correlations, implying that other factors play a more important role (also, refs. 14,56). Both the Budyko equation and FLUXCOM show very high correlations for all wet places (0.93–0.99), indicating a strong energy limitation[57], underestimated by many models (especially CWatM and MATSIRO). FLUXCOM shows a stronger N–E_a relationship (Fig. 4b) in dry–cold places than all models and the Budyko equation, while it shows a weaker P–E_a relationship (Fig. 4a) there. This could be due to an uncertain representation of energy balance processes in cold regions, possibly related to interactions between snow-affected albedo and evapotranspiration[58,59], sublimation[60] or the aerodynamic component of potential evapotranspiration[61].

For precipitation and groundwater recharge (Fig. 4c), some models (CLM4.5, MATSIRO, WaterGAP2 and H08) show high to very high correlations (0.71–0.95) for all climate regions, suggesting that precipitation is the dominant control on groundwater recharge across

https://doi.org/10.1038/s44221-023-00160-y

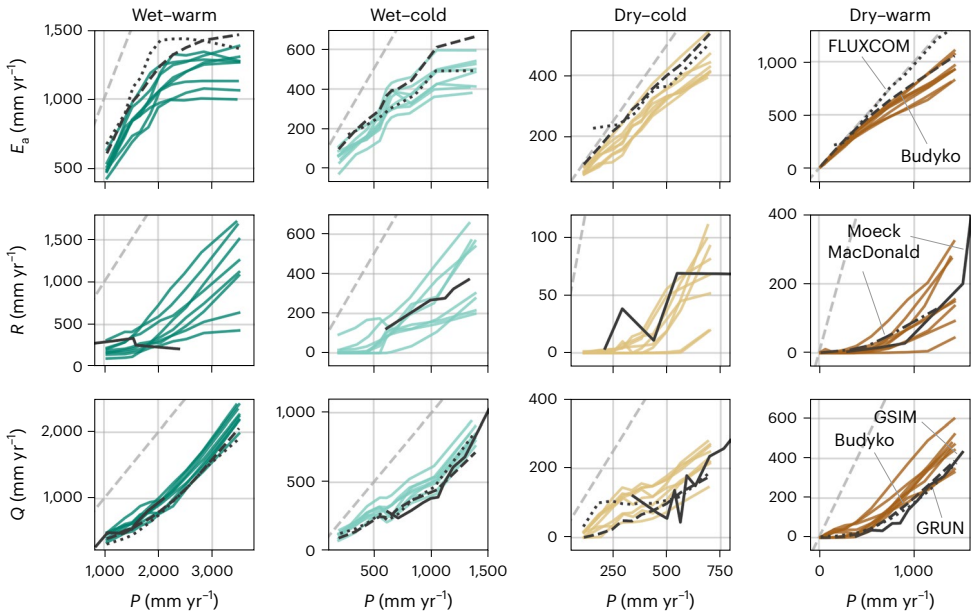


Fig. 3 | Average functional relationships among precipitation and three key water fluxes. Average functional relationships based on models and benchmark datasets among precipitation P and actual evapotranspiration E_a, groundwater recharge R and total runoff Q, respectively. The coloured lines represent one model each, the grey-black lines represent different observational datasets, labelled on the outer-right panels. The MacDonald groundwater recharge dataset contains only enough data values for the dry–warm region and is thus only shown there. The lines connect binned medians (ten bins along the x axis with equal amount of points per bin) for each climate region. The grey dashed line shows the 1:1 line, indicating the water limit assuming all water is supplied by precipitation. Note that the graphs do not show the full range for some curves to better illustrate the model differences.

all climate regions in these models. Other models (CWatM, JULES-W1, LPJmL, PCR-GLOBWB) show much lower and more variable correlations (0.35–0.85), suggesting different controls on groundwater recharge (for example, model structural decisions and parameterizations). H08 and WaterGAP2 use the same approach to calculate groundwater recharge[42] and they show almost identical rank correlations, indicating that the functional relationships might be relatable to the model structure in this case. Recent studies have shown a strong influence of precipitation and aridity on groundwater recharge[39,40,45], and using the same datasets, we also find high to very high correlations in dry–warm regions (0.74–0.84). In these often highly water-limited regions, precipitation appears to be the dominant control on groundwater recharge. Besides climate, perceptual models of groundwater recharge generation usually include soil characteristics, topography, land use and geology[62,63]. This might explain why observations show a more scattered P–R relationship, particularly in wet–warm regions (−0.06).

For precipitation and total runoff (Fig. 4e), WaterGAP2 and PCR-GLOBWB both show lower correlations (0.52–0.75) than the other models (0.58–0.95). WaterGAP2 is the only model here that is calibrated against streamflow observations[42], which might explain why it shows the lowest rank correlations for total runoff. The Budyko framework assumes that long-term runoff only depends on aridity and thus shows higher correlations (0.87–0.99) than the benchmark datasets (0.27–0.94) and most models (0.52–0.95). Because factors other than aridity can influence total runoff[51–54] and given that GSIM tends to show lower correlations overall (0.32–0.80), models that show correlations as high as the Budyko equation probably overestimate how strongly precipitation controls total runoff. Similar to the shapes of the functional relationships (Fig. 3),we generally find the largest differences in both models and datasets in dry–cold regions, where GRUN and GSIM show particularly low correlations (0.27 and 0.32).

For net radiation and both groundwater recharge and total runoff (Fig. 4d,f), we find high variability and mostly positive correlations. The models probably produce more groundwater recharge and total runoff in regions with higher net radiation because precipitation is also higher in these regions (Supplementary Fig. 1). Whereas it is difficult to interpret these correlations, the large variability still suggests considerable differences between models.

Discussion

Focus areas for model improvement

Our analysis has revealed substantial disagreement between models and between models and observations, questioning the robustness of model-based studies and impact assessments, especially if only a single model is used. The energy balance, from total radiation to actual evapotranspiration, appears to be poorly represented, indicated by a different energy limit (Fig. 3), a general underestimation of actual evapotranspiration and widely varying N–E_a relationships (Fig. 4). This warrants a closer look in future studies, as a realistic depiction of energy balance and evaporation processes is critical for climate change studies[57,58]. We find the largest disagreement for groundwater recharge, which is arguably the least understood process and poorly constrained by sparse observations[39,40]. The inter-model differences in groundwater recharge can be much larger than the differences in actual evapotranspiration and must therefore have other reasons. To better constrain the large variability between models, we need to improve our understanding of the dominant controls on groundwater recharge at large scales[64]. This knowledge is important for assessments of sustainable use of groundwater resources[9,10], for groundwater modelling studies that use groundwater recharge from global water models as input[65] and for understanding the sensitivity of groundwater recharge to changing climatic drivers[6]. Most models overestimate total runoff and we find the largest disagreement for total runoff in dry–cold regions. This echoes existing literature[1,12,22,50] and highlights the need for model refinement in dry and/or cold regions, which are under-researched and strongly affected by climate change[46]. To explore more in-depth how snow processes affect the water balance, future studies could focus on functional relationships in snow-dominated regions by specifically delineating these regions using the fraction of precipitation falling as snow or snow cover extents.

https://doi.org/10.1038/s44221-023-00160-y

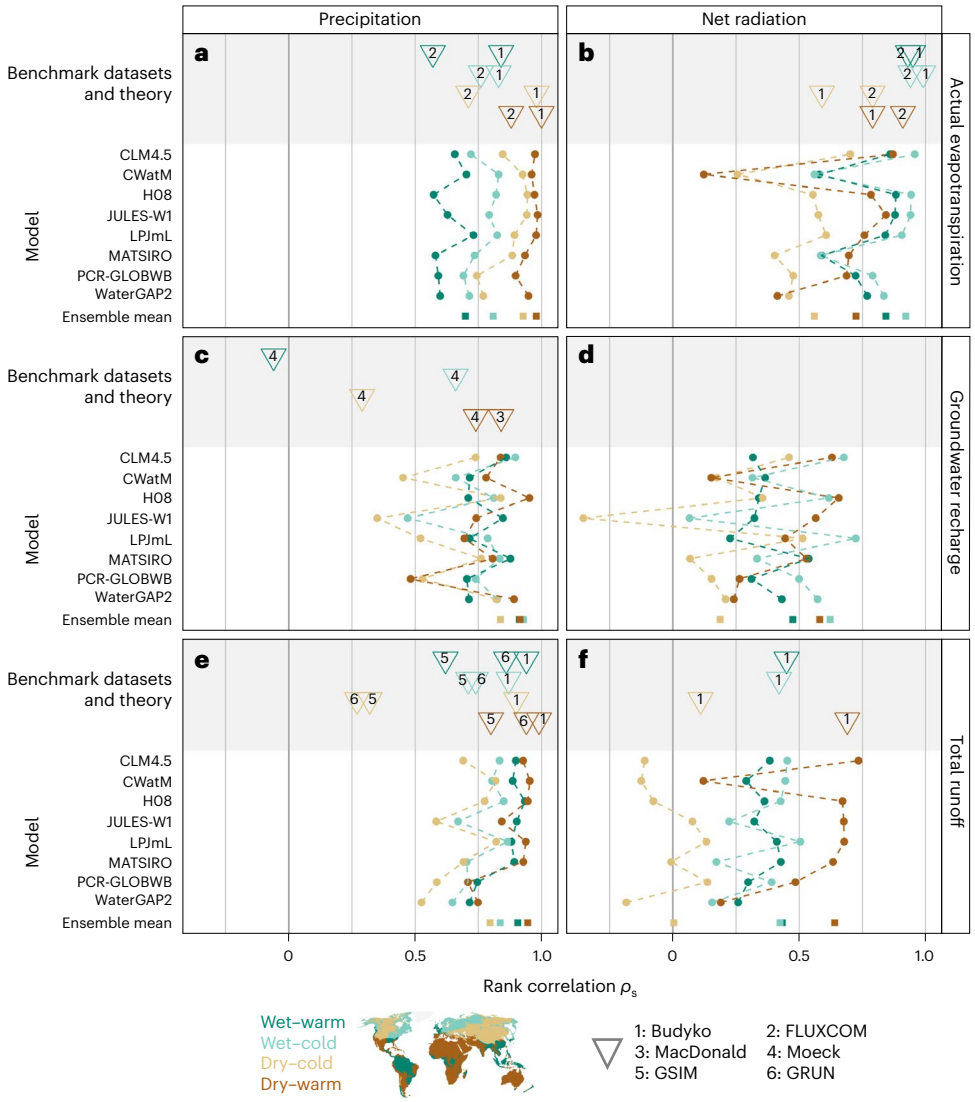


Fig. 4 | Strength of functional relationships for models and benchmark data. **a**–**f**, Spearman rank correlations ρ_s between forcing variables (precipitation (**a**,**c**,**e**), net radiation (**b**,**d**,**f**)) and water fluxes (actual evapotranspiration (**a**,**b**), groundwater recharge (**c**,**d**) and total runoff (**e**,**f**)), divided into different climate regions. Net radiation for LPJmL and PCR-GLOBWB is not available and is estimated as the median of the other models (per grid cell). The lines connecting the dots are only there as a visual aid. The numbered triangles show rank correlations based on benchmark datasets (grey background) and the Budyko equation, with numbers indicating the corresponding data source (Table 1). Observation-based rank correlations are shown only if they are based on more than 50 data points.

Towards an inventory of robust functional relationships

We have used different observational datasets, observation-driven machine learning products and the Budyko equation[20] to derive empirical and theory-based functional relationships, but challenges remain. Observation-driven machine learning products[43,49] are not raw observations and may reflect their upscaling methods rather than the underlying natural distribution but serve as useful benchmarks in the absence of direct observations (for example, because of limited numbers of FLUXNET sites[66]). The Budyko equation[20] is a climate-only model and thus provides a useful benchmark but neglects other influences on the long-term water balance. The observations themselves and the forcing data paired with them are also associated with uncertainty, even though most of the relationships used here appear to be relatively robust (Methods includes an extended discussion). Yet especially for variables with small numbers of observations, it is challenging to provide robust observation-based constraints for certain regions (Table 1). For example, groundwater recharge measurements have almost entirely been made in dry–warm regions (97% of MacDonald data[40] and 92% of Moeck data[39]), leaving groundwater recharge in other regions poorly constrained. On the other hand, most streamflow measurements have been taken in wet regions (60% of GSIM data used here), and globally there is a placement bias of stream gauges towards wet regions[67], even though—according to our classification—short of two-thirds of the global land area are defined as dry. Instead of taking new measurements to understand a specific place, new measurements would have much more leverage if they would help us to also understand other places, for example, by filling an observational gap along a climatic gradient (that is, in functional space). In addition, more quality-controlled datasets with uncertainty estimates[40] are critical to obtain realistic uncertainty estimates for functional relationships. This would ultimately allow us to obtain robust ranges of functional behaviour that we can benchmark our models against.

The functional relationships studied here appear to be robust with respect to modelled human impacts, probably because we investigated long-term averages over large regions where climatic controls on the selected hydrological variables dominate (Supplementary Figs. 26–30).

Yet for different variables, especially when studied at shorter temporal and smaller spatial scales, human impacts might have a considerable effect on functional relationships. The effects of human impacts might be investigated by studying strongly managed and near-natural regions separately[68]. Indeed, comparing functional relationships between human impacted and natural regions would be an excellent strategy to assess the degree of human alteration of the natural water cycle. In addition, relationships that specifically focus on human impacts, such as relationships between irrigated areas and irrigation water withdrawals[69], might be used to better understand the representation of human impacts in models.

Whereas visual comparison (focusing on the shape of the relationships) and rank correlations (focusing on the strength of the relationships) have exposed clear differences between models and observations, our approach here should be seen as a first step. There are other ways to describe the relationships analysed here, for example, by characterizing thresholds or nonlinearities (visible in Fig. 3). Metrics such as rank correlations also require careful interpretation. For example, positive correlations between net radiation and groundwater recharge probably arise because precipitation and net radiation are positively correlated and thus do not imply a causal relationship. The interpretation of empirical relationships should therefore be backed up by process knowledge or extended by methods that allow for discovery of causal relationships[70]. Physics-aware machine learning might be powerful in that respect, as it combines domain knowledge with versatile pattern recognition[71]. Beyond the relationships investigated here, we anticipate that exploring temporal relationships (for example, using elasticities[21,22] or shifts in P–Q relationships[23]), dividing the landscape into additional categories (for example, hydrobelts[72]) and including other variables, such as state variables or stores (for example, soil moisture, terrestrial water storage), will provide additional insights.

Conclusions

As our models grow in complexity, encompassing more processes and covering larger spatial and temporal scales, we need a concurrent development of model evaluation strategies: an evaluation framework for large-scale models. Central to such an evaluation framework should be functional relationships, which shift the focus away from matching historical records in specific locations to a more diagnostic and process-oriented evaluation of model behaviour[36]. Functional relationships allow us to focus on larger-scale assessments, to relate places to each other and to explore if dominant controls in models are consistent with observations, theory and expectations (that is, our perceptual model[73]). This understanding is critical for ensuring that models faithfully represent real-world systems, ultimately leading to more credible projections of environmental change impacts. Eventually, expanding our range of functional relationships in hydrology, constrained by various observational datasets and expert knowledge, would give us a knowledge base of realistic system behaviour that could be used to evaluate models, diagnose model deficiencies and weight model ensembles, comparable to the use of emergent constraints in climate modelling[37].

Both our approach and our findings have implications beyond hydrology. First, the terrestrial water cycle plays a central role in the Earth system and is often strongly coupled to other components, such as the biosphere, lithosphere and atmosphere and human activities (for example, refs. 74–76). More realistic simulations of the global water cycle therefore also enable us to better clarify how it influences and is influenced by other Earth system components. Methodologically, functional relationships are not limited to applications in hydrology. In fact, land surface, forest and Earth system models[26–32] have already been studied in similar fashions, though a broader application of this approach has so far been missing. As indicated by recent studies[76,77], functional relationships provide an excellent opportunity to study the interactions between hydrology and, for example, terrestrial ecosystems, and thus represent a tool that can be used across disciplines.

Beyond model evaluation, functional relationships invite us to think about how the global water cycle functions, what we know, what we do not know and what that means for a future under climate change[73]. Our results suggest that improved process understanding will be particularly important for energy balance processes, groundwater recharge processes and generally in dry and/or cold regions. So how can we improve our process understanding? In 1986, Eagleson[78] stated that 'science advances on two legs, analysis and experimentation, and at any moment one is ahead of the other. At the present time advances in hydrology appear to be data limited'. For some processes, this still seems to be the case. But clearly, we have a wealth of data available and might ask ourselves: are we extracting all of the information from the observations we have? On the basis of the data we have, what and where should we measure next? And are there functional relationships in hydrology yet to be found[19]? Even if the search for such relationships is challenging, it will be a fruitful and exciting endeavour for global hydrology.

Methods

Model data retrieval and processing

We analysed 30-year (climatological) averages (1975–2004) from eight global water models[41]: CLM4.5[79], CWatM[80], H08[81], JULES-W1[82], LPJmL[83], MATSIRO[84], PCR-GLOBWB[85] and WaterGAP2[86]. The model simulations were carried out following the ISIMIP 2b protocol and here we used model outputs forced with the Earth system model HadGEM2-ES under historical conditions (historical climate and CO_2 concentrations). We note that the specific forcing chosen does not appear to influence model-based functional relationships (see below). We used precipitation P (ISIMIP variable name *pr*), net radiation N (not an official ISIMIP output), actual evapotranspiration E_a (ISIMIP variable name *evap*), groundwater recharge R (ISIMIP variable name *qr*) and total runoff Q (ISIMIP variable name *qtot*). Note that Q here refers to runoff generated on the land fractions (and not surface water bodies) of each grid cell and does not include upstream inflows, which allows for comparison to grid cell P. P, E_a, R, and Q were downloaded from https://data.isimip.org/. Net radiation N is not an official ISIMIP output and was provided by the individual modelling groups. It is not available for all models, so we used the ensemble median per grid cell for models without N data. We converted all fluxes to mm per year and removed E_a values larger than 10,000 mm per year and set R values smaller than 0 to 0. Note that our analysis excludes Greenland and Antarctica. A more detailed description is given in the Supplementary Information.

CoV and most deviating model maps

For each grid cell, we used the 30-year averages of the eight models (that is, the model ensemble) and calculated the ensemble standard deviation divided by the ensemble mean. Maps of the standard deviation are shown in the Supplementary Information (Supplementary Figs. 8–10). To see which model dominates the ensemble spread, we checked for each grid cell which model shows the largest absolute difference (denoted by d_1) from the ensemble mean (denoted by μ). To see if multiple models dominate the ensemble spread, we also checked for each grid cell which model shows the second-largest absolute difference (denoted by d_2) from the ensemble mean. If the relative difference between the largest and the second-largest difference is less than 20%, that is $(d_1 - d_2)/d_1 < 0.2$, the grid cell falls into the category 'multiple'. If the relative difference between the most deviating model and the ensemble mean is less than 20%, that is $d_1/\mu < 0.2$, the grid cell is counted as having no most deviating model (empty areas on Fig. 1d–f).

Functional relationships

To visualize the shape of the functional relationships, we binned the data in each climate region into ten bins (along the x axis) with an

equal amount of points, calculated the median per bin and connected the obtained median value. For groundwater recharge, we used only five bins because there are so few values. Note that the non-gridded observational datasets do not have the same spatial distribution as the gridded datasets and the models and thus do not have the same distribution of forcing variables. Their bins can therefore span different ranges of the forcing variables. As a metric for the strength of the functional relationships, we calculate Spearman rank correlations ρ_s between model inputs and outputs per climate region, a measure of the monotonicity between two variables that is robust to outliers. We use the following categories for correlations: negative correlation (<0), no to low correlation (0 to 0.25), medium correlation (0.25–0.5), high correlation (0.5–0.75), very high correlation (0.75–1.0). We also show mean fluxes and slopes obtained through linear regression in Supplementary Tables 4–7.

Climate regions

On the basis of the aridity index (here defined as N/P; where N is model ensemble median), a place is categorized as either wet ($N/P < 1$) or dry ($N/P > 1$). On the basis of how many days per year fall below a 1 °C temperature threshold, a place is categorized as either cold (more than one month below 1 °C) or warm (less than one month below 1 °C). This results in four categories: wet–warm (15% of model grid cells/18% of modelled area), wet–cold (23%/15%), dry–cold (28%/24%) and dry–warm (34%/43%). To test how different decisions affect our climate region classification, we also used the ensemble median of potential evapotranspiration E_p (partially downloaded, partially provided by the modelling groups) to calculate the aridity index (E_p/P), and we used a different threshold for our warm/cold distinction. This resulted in little differences overall, as can be seen in the Supplementary Information (Supplementary Fig. 14).

Benchmark datasets and theory

To benchmark model performance, we used different observational datasets, observation-driven machine learning products and the Budyko equation[20]. If the datasets provide their own forcing data, we used these data. If not, we paired them with GSWP3 P data[87] to have one consistent forcing product. For E_a, we used FLUXCOM data[43] (RS monthly 0.5° from 2001–2015) paired with GSWP3 P data[87] (downloaded from https://data.isimip.org/). For R, we used data from MacDonald et al.[40], which include matching P data, and data from Moeck et al.[39] paired with GSWP3 P data[87]. For Q, we used GRUN data[49] from 1985–2004 paired with GSWP3 P data[87] (the dataset used in the creation of GRUN) and GSIM data[47,48] paired with GSWP3 P data[87]. For GSIM, we only used catchments with areas ranging from 250 to 25,000 km^2 with a minimum of ten years of data between 1985 and 2004 to ensure a sufficient number of catchments that do not differ too much in size from the model grid cells. To obtain theory-based estimates for E_a and Q, we forced the Budyko[20] equation (equation (1)) with HadGEM2-ES P (the same forcing as used for the models) and ensemble median N from the ISIMIP 2b models analysed here.

$$\frac{E_a}{P} = \sqrt{\frac{N}{P}\tanh\left(\frac{P}{N}\right)\left(1-\exp\left(-\frac{N}{P}\right)\right)} \quad (1)$$

More details on data processing and quality checks can be found in the Supplementary Information.

Extended discussion on model forcing and scenario uncertainty

The choice of forcing product and differences in the treatment of human influences (for example, water use and dams) might affect the functional relationships exhibited by the models. To get an idea how much uncertainty this introduces, we compared our results to model runs using WATCH-WFDEI forcing with either variable historical conditions (varsoc) or no human influences (nosoc) for WaterGAP2 and PCR-GLOBWB, carried out following the ISIMIP 2a protocol. The results, shown in the Supplementary Information (Supplementary Figs. 26–30), stay essentially the same, showing that the model-based correlations are robust signatures of model behaviour.

Extended discussion on benchmark dataset uncertainty

Because not all datasets come with matching P data, we sometimes paired the observations with GSWP3 reanalysis data[87]. To get an idea how much uncertainty this introduces, we investigated how different P data sources affect the functional relationships. Correlations calculated using the MacDonald et al.[40] R data with either GSWP3 P data or the accompanying P data are very similar for dry–warm places (0.83 and 0.84; Supplementary Information). Using HadGEM2-ES P (the model forcing) data instead of GSWP3 P data to calculate correlations with FLUXCOM E_a[43], Moeck R[39], GRUN Q[49] and GSIM[47,48], respectively, results in no notable differences. Because most datasets only contain a limited number of years of data, sometimes only one average value[39,40], we used all available years in our analysis. The only observation-driven dataset that contains a long enough time series to analyse functional relationships for two independent 30-year periods is GRUN[49]. Using GRUN data from 1945–1974 instead of 1975–2004 results in virtually no differences. While we cannot rule out that other datasets would lead to different relationships, this analysis indicates that the functional relationships and the rank correlations are relatively robust (Supplementary Figs. 31–42).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The long-term averages created and used in this study are deposited at https://zenodo.org/record/7714885. Correlations and other statistics are available in the Supporting Information. Data used in this study can be downloaded from the following links. ISIMIP 2b data (model outputs and GSWP3 precipitation data) are available from https://www.isimip.org/. FLUXCOM data are available from http://www.fluxcom.org/. MacDonald et al. recharge data are available from https://www2.bgs.ac.uk/nationalgeosciencedatacentre/citedData/catalogue/45d2b71c-d413-44d4-8b4b-6190527912ff.html (contains data supplied by permission of the Natural Environment Research Council (2022)). Moeck et al. recharge data are available from https://opendata.eawag.ch/dataset/globalscale_groundwater_moeck. GSIM data are available from https://doi.pangaea.de/10.1594/PANGAEA.887477 and https://doi.pangaea.de/10.1594/PANGAEA.887470. MSWEP data can be requested for research purposes from http://www.gloh2o.org/mswep/.

Code availability

Python and R codes used to perform the analyses are available at https://github.com/HydroSysPotsdam/GHM_Comparison.

References

1. Gädeke, A. et al. Performance evaluation of global hydrological models in six large Pan-Arctic watersheds. *Climatic Change* **163**, 1329–1351 (2020).
2. IPCC. *Climate Change 2022: Impacts, Adaptation and Vulnerability* (eds Pörtner, H. O. et al.) (Cambridge Univ. Press, 2022).
3. Samaniego, L. et al. Anthropogenic warming exacerbates European soil moisture droughts. *Nat. Clim. Change* **8**, 421–426 (2018).
4. Schewe, J. et al. Multimodel assessment of water scarcity under climate change. *Proc. Natl Acad. Sci.* **111**, 3245–3250 (2014).
5. Pokhrel, Y. et al. Global terrestrial water storage and drought severity under climate change. *Nat. Clim. Change* **11**, 226–233 (2021).

6. Reinecke, R. et al. Uncertainty of simulated groundwater recharge at different global warming levels: a global-scale multi-model ensemble study. *Hydrol. Earth Syst. Sci.* **25**, 787–810 (2021).
7. IGRAC *Global Groundwater Information System* https://www.un-igrac.org/global-groundwater-information-system-ggis (2022).
8. Sheffield, J. et al. A drought monitoring and forecasting system for sub-Sahara African water resources and food security. *Bull. Am. Meteorol. Soc.* **95**, 861–882 (2014).
9. Wada, Y. et al. Global depletion of groundwater resources. *Geophys. Res. Lett.* **37**, L20402 (2010).
10. Richey, A. S. et al. Quantifying renewable groundwater stress with GRACE. *Water Resour. Res.* **51**, 5217–5238 (2015).
11. Bierkens, M. F. P. Global hydrology 2015: state, trends, and directions. *Water Resour. Res.* **51**, 4923–4947 (2015).
12. Giuntoli, I., Vidal, J.-P., Prudhomme, C. & Hannah, D. M. Future hydrological extremes: the uncertainty from multiple global climate and global hydrological models. *Earth Syst. Dyn.* **6**, 267–285 (2015).
13. Beck, H. E. et al. Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrol. Earth Syst. Sci.* **21**, 2881–2903 (2017).
14. Wartenburger, R. et al. Evapotranspiration simulations in ISIMIP2a—evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets. *Environ. Res. Lett.* **13**, 075001 (2018).
15. Gleeson, T. et al. GMD perspective: the quest to improve the evaluation of groundwater representation in continental- to global-scale models. *Geosci. Model Dev.* **14**, 7545–7571 (2021).
16. Hrachowitz, M. et al. A decade of predictions in ungauged basins (PUB)—a review. *Hydrol. Sci. J.* **58**, 1198–1255 (2013).
17. Peel, M. C. & Blöschl, G. Hydrological modelling in a changing world. *Prog. Phys. Geogr.: Earth Environ.* **35**, 249–261 (2011).
18. Wagener, T., Reinecke, R. & Pianosi, F. On the evaluation of climate change impact models. *WIREs Clim. Change* **13**, e772 (2022).
19. Dooge, J. C. I. Looking for hydrologic laws. *Water Resour. Res.* **22**, 46S–58S (1986).
20. Budyko, M. I. *Climate and Life* (Academic Press, 1974).
21. Němec, J. & Schaake, J. Sensitivity of water resource systems to climate variation. *Hydrol. Sci. J.* **27**, 327–343 (1982).
22. Zhang, Y. et al. Future global streamflow declines are probably more severe than previously estimated. *Nat. Water* **1**, 261–271 (2023).
23. Peterson, T. J., Saft, M., Peel, M. C. & John, A. Watersheds may not recover from drought. *Science* **372**, 745–749 (2021).
24. Wagener, T., Sivapalan, M., Troch, P. & Woods, R. Catchment classification and hydrologic similarity. *Geogr. Compass* **1**, 901–931 (2007).
25. Black, P. E. Watershed functions. *JAWRA J. Am. Water Resour. Assoc.* **33**, 1–11 (1997).
26. Betts, A. K. Understanding hydrometeorology using global models. *Bull. Am. Meteorol. Soc.* **85**, 1673–1688 (2004).
27. Dirmeyer, P. A., Koster, R. D. & Guo, Z. Do global models properly represent the feedback between land and atmosphere? *J. Hydrometeorol.* **7**, 1177–1198 (2006).
28. Koster, R. D. & Milly, P. The Interplay between transpiration and runoff formulations in land surface schemes used with atmospheric models. *J. Clim.* **10** (1997).
29. Koster, R. D. & Mahanama, S. P. P. Land surface controls on hydroclimatic means and variability. *J. Hydrometeorol.* **13**, 1604–1620 (2012).
30. Randerson, J. T. et al. Systematic assessment of terrestrial biogeochemistry in coupled climate–carbon models. *Glob. Change Biol.* **15**, 2462–2484 (2009).
31. Swart, N. C. et al. The Canadian Earth System Model version 5 (CanESM5.0.3). *Geosci. Model Dev.* **12**, 4823–4873 (2019).
32. Mahnken, M. et al. Accuracy, realism and general applicability of european forest models. *Glob. Change Biol.* **28**, 6921–6943 (2022).
33. Kapangaziwiri, E., Hughes, D. & Wagener, T. Incorporating uncertainty in hydrological predictions for gauged and ungauged basins in southern Africa. *Hydrol. Sci. J.* **57**, 1000–1019 (2012).
34. Troy, T. J., Wood, E. F. & Sheffield, J. An efficient calibration method for continental-scale land surface modeling. *Water Resour. Res.* **44**, W09411 (2008).
35. Greve, P., Burek, P. & Wada, Y. Using the Budyko framework for calibrating a global hydrological model. *Water Resour. Res.* **56**, e2019WR026280 (2020).
36. Gupta, H. V., Wagener, T. & Liu, Y. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrol. Processes* **22**, 3802–3813 (2008).
37. Eyring, V. et al. Taking climate model evaluation to the next level. *Nat. Clim. Change* **9**, 102–110 (2019).
38. L'vovich, M. I. *World Water Resources and Their Future* (American Geophysical Union, 1979).
39. Moeck, C. et al. A global-scale dataset of direct natural groundwater recharge rates: a review of variables, processes and relationships. *Sci. Total Environ.* **717**, 137042 (2020).
40. MacDonald, A. M. et al. Mapping groundwater recharge in Africa from ground observations and implications for water security. *Environ. Res. Lett.* **16**, 034012 (2021).
41. Frieler, K. et al. Assessing the impacts of 1.5 °C global warming—simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b). *Geosci. Model Dev.* **10**, 4321–4345 (2017).
42. Telteu, C.-E. et al. Understanding each other's models: an introduction and a standard representation of 16 global water models to support intercomparison, improvement, and communication. *Geosci. Model Dev.* **14**, 3843–3878 (2021).
43. Jung, M. et al. The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Sci. Data* **6**, 74 (2019).
44. Elnashar, A., Wang, L., Wu, B., Zhu, W. & Zeng, H. Synthesis of global actual evapotranspiration from 1982 to 2019. *Earth Syst. Sci. Data* **13**, 447–480 (2021).
45. Berghuijs, W. R., Luijendijk, E., Moeck, C., van der Velde, Y. & Allen, S. T. Global recharge data set indicates strengthened groundwater connection to surface fluxes. *Geophys. Res. Lett.* **49**, e2022GL099010 (2022).
46. Zoccatelli, D. et al. Contrasting rainfall–runoff characteristics of floods in desert and Mediterranean basins. *Hydrol. Earth Syst. Sci.* **23**, 2665–2678 (2019).
47. Do, H. X., Gudmundsson, L., Leonard, M. & Westra, S. The Global Streamflow Indices and Metadata Archive (GSIM)–part 1: the production of a daily streamflow archive and metadata. *Earth Syst. Sci. Data* **10**, 765–785 (2018).
48. Gudmundsson, L., Do, H. X., Leonard, M. & Westra, S. The Global Streamflow Indices and Metadata Archive (GSIM)–part 2: quality control, time-series indices and homogeneity assessment. *Earth Syst. Sci. Data* **10**, 787–804 (2018).
49. 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).
50. Zaherpour, J. et al. Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts. *Environ. Res. Lett.* **13**, 065015 (2018).
51. Berghuijs, W. R., Woods, R. A. & Hrachowitz, M. A precipitation shift from snow towards rain leads to a decrease in streamflow. *Nat. Clim. Change* **4**, 583–586 (2014).
52. Milly, P. C. D. Climate, soil water storage, and the average annual water balance. *Water Resour. Res.* **30**, 2143–2156 (1994).

53. Zhang, L., Dawes, W. R. & Walker, G. R. Response of mean annual evapotranspiration to vegetation changes at catchment scale. *Water Resour. Res.* **37**, 701–708 (2001).
54. Liu, Y., Wagener, T., Beck, H. E. & Hartmann, A. What is the hydrologically effective area of a catchment? *Environ. Res. Lett.* **15**, 104024 (2020).
55. Wang, D. & Hejazi, M. Quantifying the relative contribution of the climate and direct human impacts on mean annual streamflow in the contiguous United States. *Water Resour. Res.* **47**, W00J12 (2011).
56. Haddeland, I. et al. Multimodel estimate of the global terrestrial water balance: setup and first results. *J. Hydrometeorol.* **12**, 869–884 (2011).
57. Milly, P. C. D. & Dunne, K. A. Potential evapotranspiration and continental drying. *Nat. Clim. Change* **6**, 946–949 (2016).
58. Milly, P. C. D. & Dunne, K. A. Colorado River flow dwindles as warming-driven loss of reflective snow energizes evaporation. *Science* **367**, 1252–1255 (2020).
59. Meira Neto, A. A., Niu, G.-Y., Roy, T., Tyler, S. & Troch, P. A. Interactions between snow cover and evaporation lead to higher sensitivity of streamflow to temperature. *Commun. Earth Environ.* **1**, 56 (2020).
60. Bowling, L. C., Pomeroy, J. W. & Lettenmaier, D. P. Parameterization of blowing-snow sublimation in a macroscale hydrology model. *J. Hydrometeorol.* **5**, 745–762 (2004).
61. Tabari, H. & Talaee, P. H. Local calibration of the Hargreaves and Priestley–Taylor equations for estimating reference evapotranspiration in arid and cold climates of Iran based on the Penman–Monteith model. *J. Hydrol. Eng.* **16**, 837–845 (2011).
62. Scanlon, B. R., Healy, R. W. & Cook, P. G. Choosing appropriate techniques for quantifying groundwater recharge. *Hydrogeol. J.* 22 (2002).
63. Cuthbert, M. O. et al. Observed controls on resilience of groundwater to climate variability in sub-Saharan Africa. *Nature* **572**, 230–234 (2019).
64. West, C. et al. Ground truthing global-scale model estimates of groundwater recharge across Africa. *Sci. Total Environ.* **858**, 159765 (2023).
65. Reinecke, R. et al. Challenges in developing a global gradient-based groundwater model (G^3M v1.0) for the integration into a global hydrological model. *Geosci. Model Dev.* **12**, 2401–2418 (2019).
66. Pastorello, G. et al. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci. Data* **7**, 225 (2020).
67. Krabbenhoft, C. A. et al. Assessing placement bias of the global river gauge network. *Nat. Sustain.* https://doi.org/10.1038/s41893-022-00873-0 (2022).
68. Veldkamp, T. I. E. et al. Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study. *Environ. Res. Lett.* **13**, 055008 (2018).
69. Puy, A., Borgonovo, E., Lo Piano, S., Levin, S. A. & Saltelli, A. Irrigated areas drive irrigation water withdrawals. *Nat. Commun.* **12**, 4525 (2021).
70. Massmann, A., Gentine, P. & Runge, J. Causal inference for process understanding in Earth sciences. Preprint at https://arxiv.org/abs/2105.00912 (2021).
71. Reichstein, M. et al. Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
72. Meybeck, M., Kummu, M. & Dürr, H. H. Global hydrobelts and hydroregions: improved reporting scale for water-related issues? *Hydrol. Earth Syst. Sci.* **17**, 1093–1111 (2013).
73. Wagener, T. et al. On doing hydrology with dragons: realizing the value of perceptual models and knowledge accumulation. *WIREs Water* **8**, e1550 (2021).
74. Pastor, A. V. et al. The global nexus of food–trade–water sustaining environmental flows by 2050. *Nat. Sustain.* **2**, 499–507 (2019).
75. Zhao, M. et al. Ecological restoration impact on total terrestrial water storage. *Nat. Sustain.* **4**, 56–62 (2021).
76. Denissen, J. M. C. et al. Widespread shift from ecosystem energy to water limitation with climate change. *Nat. Clim. Change* **12**, 677–684 (2022).
77. Bonetti, S., Wei, Z. & Or, D. A framework for quantifying hydrologic effects of soil structure across scales. *Commun. Earth Environ.* **2**, 1–10 (2021).
78. Eagleson, P. S. The emergence of global-scale hydrology. *Water Resour. Res.* **22**, 6S–14S (1986).
79. Thiery, W. et al. Present-day irrigation mitigates heat extremes. *J. Geophys. Res. Atmos.* **122**, 1403–1422 (2017).
80. Burek, P. et al. Development of the Community Water Model (CWatM v1.04)—a high-resolution hydrological model for global and regional assessment of integrated water resources management. *Geosci. Model Dev.* **13**, 3267–3298 (2020).
81. Hanasaki, N., Yoshikawa, S., Pokhrel, Y. & Kanae, S. A global hydrological simulation to specify the sources of water used by humans. *Hydrol. Earth Syst. Sci.* **22**, 789–817 (2018).
82. Best, M. J. et al. The Joint UK Land Environment Simulator (JULES), model description–part 1: energy and water fluxes. *Geosci. Model Dev.* **4**, 677–699 (2011).
83. Jägermeyr, J. et al. Water savings potentials of irrigation systems: global simulation of processes and linkages. *Hydrol. Earth Syst. Sci.* **19**, 3073–3091 (2015).
84. Takata, K., Emori, S. & Watanabe, T. Development of the minimal advanced treatments of surface interaction and runoff. *Glob. Planet. Change* **38**, 209–222 (2003).
85. Sutanudjaja, E. H. et al. PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geosci. Model Dev.* **11**, 2429–2453 (2018).
86. Müller Schmied, H. et al. Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use. *Hydrol. Earth Syst. Sci.* **20**, 2877–2898 (2016).
87. Dirmeyer, P. A. et al. GSWP-2: multimodel analysis and implications for our perception of the land surface. *Bull. Am. Meteorol. Soc.* **87**, 1381–1398 (2006).

Acknowledgements

S.G., R.R., L.S. and T.W. acknowledge support from the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research (BMBF). Y.S. was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (2021H1D3A2A03097768). Y.P. acknowledges the support from the National Science Foundation (grant number 1752729). A.K. and M.G. have received support from REACT4MED (GA 2122) PRIMA funded project, supported by Horizon 2020. N.H. is financially supported by JSPS KAKENHI grant number 21H05002. This publication is based upon work from COST Action CA19139—PROCLIAS, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu). We thank P. Döll for providing helpful comments on the manuscript. We also thank the ISIMIP team for their continued efforts within the ISIMIP project.

Author contributions

S.G., R.R., L.S. and T.W. designed the study. Y.W., W.T., H.M.S., Y.S., Y.P., S.O., A.K., N.H., M.G. and P.B. conducted hydrological simulations

under the ISIMIP 2b project, and S.N.G. and H.M.S. coordinated the ISIMIP global water sector. S.G. and R.R. processed the simulation results and conducted the analyses, and S.G., R.R. and L.S. prepared the graphics. S.G. wrote the first paper draft together with R.R., L.S. and T.W. Y.W., W.T., H.M.S., Y.S., Y.P., S.O., A.K., N.H., M.G., S.N.G., P.B. and M.F.P.B. contributed to discussions and interpretations of the results and edited the paper.

Funding

Open access funding provided by Universität Potsdam.

Competing interests

The authors declare no competing interests.

Additional information

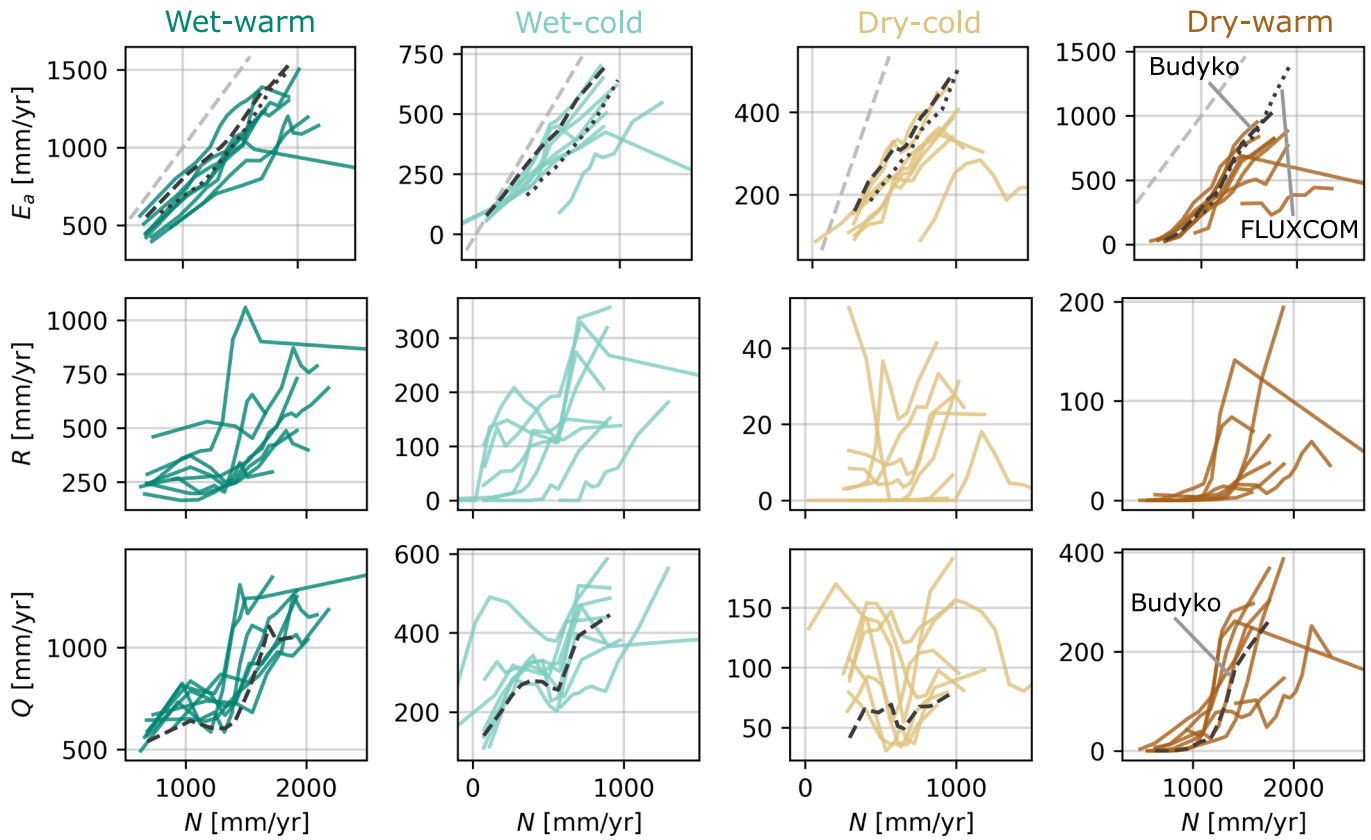
Extended data is available for this paper at https://doi.org/10.1038/s44221-023-00160-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44221-023-00160-y.

Correspondence and requests for materials should be addressed to Sebastian Gnann.

Peer review information *Nature Water* thanks Yongqiang Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s44221-023-00160-y



Extended Data Fig. 1 | Average functional relationships between net radiation and three key water fluxes. Average functional relationships based on models and benchmark datasets between net radiation N and actual evapotranspiration E_a, groundwater recharge R and total runoff Q, respectively. The colored lines represent one model each, the grey-black lines represent different observational datasets, labeled on the outer-right panels. The lines connect binned medians (10 bins along the x-axis with equal amount of points per bin) for each climate region. The grey dashed line shows the 1:1 line, indicating the water limit assuming all water is supplied by precipitation. Note that the graphs do not show the full range for some curves to better illustrate the model differences.

nature portfolio

Corresponding author(s): Sebastian Gnann

Last updated by author(s): 10.03.2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was downloaded from the respective data providers, as specified in data availability statement.
Data analysis	Python and R codes used to perform the analysis are available at https://github.com/HydroSysPotsdam/GHM_Comparison.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The long-term averages created and used in this study are deposited at https://zenodo.org/record/7714885. Correlations and other statistics are available in the Supporting Information. Data used in this study can be downloaded from the following links. ISIMIP 2b data (model outputs and GSWP3 precipitation data) are available from https://www.isimip.org/. FLUXCOM data are available from http://www.fluxcom.org/. MacDonald et al. recharge data are available from https://www2.bgs.ac.uk/nationalgeosciencedatacentre/citedData/catalogue/45d2b71c-d413-44d4-8b4b-6190527912ff.html. Contains data supplied by permission of the

Natural Environment Research Council [2022]. Moeck et al. recharge data are available from https://opendata.eawag.ch/dataset/globalscale_groundwater_moeck. GSIM data are available from https://doi.pangaea.de/10.1594/PANGAEA.887477 and https://doi.pangaea.de/10.1594/PANGAEA.887470. MSWEP data can be requested for research purposes from http://www.gloh2o.org/mswep/.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study evaluates 8 global water models using functional relationships that capture the spatial co-variability of forcing (precipitation, net radiation) and response variables (actual evapotranspiration, groundwater recharge, total runoff). To diagnose how models represent key aspects of the global water cycle, the study compares functional relationships between models and between models and several quasi-global observational or observation-driven datasets.
Research sample	We used inputs to and outputs from 8 global water models, and several observational datasets of the following meteorological and hydrological variables: precipitation, temperature, net radiation, actual evapotranspiration, groundwater recharge, total runoff.
Sampling strategy	n/a
Data collection	Data are based on previously published observational datasets, global gridded data products, or model outputs, all available from the respective links in the data availability statement.
Timing and spatial scale	Model inputs and outputs are global (0.5° resolution) from the years 1975-2004. Details on observational data can be found in the paper.
Data exclusions	We excluded evidently unrealistic values (e.g. negative fluxes) and only used streamflow from catchments with areas from 250-25000 km² with minimum 10y of data. Details can be found in the Methods section of the paper.
Reproducibility	All analyses are based on publicly available data and can be reproduced. Link to the code used is given in the code availability statement.
Randomization	n/a
Blinding	n/a

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging