Veldkamp, Ted Isis Elize and Zhao, Fang and Ward, Philip J. and Moel, Hans de and Aerts, Jeroen C.J.H. and Müller Schmied, Hannes and Portmann, Felix T. and Masaki, Yoshimitsu and Pokhrel, Yadu and Liu, Xingcai and Satoh, Yusuke and Gerten, Dieter and Gosling, Simon N. and Zaherpour, Jamal and Wada, Y. (2018) Human impact parameterizations in global hydrological models improves estimates of monthly discharges and hydrological extremes: a multi-model validation study. Environmental Research Letters . ISSN 1748-9326

Environmental Research Letters

ACCEPTED MANUSCRIPT • OPEN ACCESS

Human impact parameterizations in global hydrological models improves estimates of monthly discharges and hydrological extremes: a multi-model validation study

To cite this article before publication: Ted Isis Elize Veldkamp *et al* 2018 *Environ. Res. Lett.* in press https://doi.org/10.1088/1748-9326/aab96f

Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study

T.I.E. Veldkamp[1,9], F. Zhao[2], P.J. Ward[1], H. de Moel[1], J.C.J.H., Aerts[1,3], H. Müller Schmied[4,5], F.T. Portmann[4], Y. Masaki[6], Y. Pokhrel[7], X. Liu[8], Y. Satoh[9], D. Gerten[2,10], S.N. Gosling[11], J. Zaherpour[11], Y. Wada[9,12]

[1] Institute for Environmental Studies (IVM), VU Amsterdam, the Netherlands
[2] Potsdam Institute for Climate Impact Research, Potsdam, Germany
[3] Department of Geography, University of California, Santa Barbara, Santa Barbara, USA
[4] Institute of Physical Geography, Goethe-University Frankfurt, Frankfurt, Germany
[5] Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt, Germany
[6] National Institute for Environmental Studies, Tsukuba, Japan
[7] Department of Civil and Environmental Engineering, Michigan State University, Michigan, USA
[8] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[9] International Institute for Applied Systems Analysis, Laxenburg, Austria
[10] Department of Geography, Humboldt-Universität zu Berlin, Berlin, Germany.
[11] School of Geography, University of Nottingham, Nottingham, United Kingdom
[12] Department of Physical Geography, Utrecht University, Netherlands

E-mail: ted.veldkamp@vu.nl

Abstract

Human activities have a profound influence on river discharge, hydrological extremes, and water-related hazards. In this study, we compare the results of five state-of-the-art global hydrological models (GHMs) with observations to examine the role of human impact parameterizations (HIP) in the simulation of the mean, high, and low flows. The analysis is performed for 471 gauging stations across the globe and for the period 1971-2010. We find that the inclusion of HIP improves the performance of GHMs, both in managed and near-natural catchments. For near-natural catchments, the improvement in performance results from improvements in incoming discharges from upstream managed catchments. This finding is robust across GHMs, although the level of improvement and reasons for improvement vary greatly by GHM. The inclusion of HIP leads to a significant decrease in the bias of long-term mean monthly discharge in 36-73% of the studied catchments, and an improvement in modelled hydrological variability in 31-74% of the studied catchments. Including HIP in the GHMs also leads to an improvement in the simulation of hydrological extremes, compared to when HIP is excluded. Whilst the inclusion of HIP leads to decreases in simulated high-flows, it can lead to either increases or decreases in low-flows. This is due to the relative importance of the timing of return flows and reservoir operations and their associated uncertainties. Even with the inclusion of HIP, we find that model performance still not optimal. This highlights the need for further research linking the human management and hydrological domains, especially in those areas with a dominant human impact. The large variation in performance between GHMs, regions, and performance indicators, calls for a careful selection of GHMs, model components, and evaluation metrics in future model applications.

1. Introduction

Human activities have a profound influence on river discharge, hydrological extremes, and water-related hazards, like flooding, droughts, water scarcity, and water quality issues (Van Loon et al., 2016; Liu et al,. 2017; Padowski et al., 2015; Veldkamp et al., 2017; Wada et al., 2011; Winsemius et al., 2016). As a result, research efforts have been made to parameterize human activities in global hydrological models (hereafter: GHMs, a full list of abbreviations is presented in **supplementary table 2**) (Bierkens, 2015; Pokhrel et al., 2016). These model parameterizations include: the incorporation of dam and reservoir operations; the representation of human water use and return flows; and the representations of land use, land management, and land cover change (Pokhrel et al., 2016; Wada et al., 2016a, 2017).

GHMs are widely used in scientific studies. For example, they have been used to assess the historical and future impacts of socioeconomic developments and/or hydro-climatic variability and change, on freshwater resources, droughts, and water scarcity (Biemans et al., 2011; Döll et al., 2009; Döll and Müller Schmied, 2012; Fujimori et al., 2017; Gosling et al., 2017; Haddeland et al., 2006, 2007, 2014; Hanasaki et al., 2013; Van Huijgevoort et al., 2013; Kummu et al., 2016; Müller Schmied et al., 2016; Munia et al., 2016; Rost et al., 2008; Veldkamp et al., 2015a,b, 2016, 2017, Wada et al., 2011, 2013a,b, 2014a, Wanders et al., 2015). They are also increasingly used in practice. Global institutions increasingly rely on GHMs to conduct first-order assessments of water-related hazards because data, time, or resources are in short-supply for setting-up and executing multiple in-depth local studies. For example, GHMs have provided input into a multitude of high-level policy documents, such as: UN World Water Development Reports (e.g. Alcamo and Gallopin, 2009); Global Environmental Outlooks (UNEP, 2007); World Bank series on climate change and development (Hallegatte et al., 2016, 2017); and IPCC assessment reports (IPCC, 2007, 2013).

As GHMs continue to improve in terms of detail, granularity, and speed, their importance for global, regional, and local applications is likely to increase further (Bierkens, 2015). Therefore, it is essential to have a thorough understanding of how well these GHMs represent real-world hydrological conditions. However, most GHM validation studies are limited to near-natural river catchments and make use of naturalized discharge data (Beck et al., 2016; Gudmundsson et al., 2011, 2012). Studies that have validated GHM simulations where human activities included have either focused on a single GHM and/or few selected river catchments (Biemans et al., 2011; Döll et al., 2003; 2009; De Graaf et al., 2014; Haddeland et al., 2006; Masaki et al., 2017; Müller Schmied et al., 2014; Pokhrel et al. 2012; Wada et al., 2011, 2013a, 2014a).

To date, a comprehensive validation of the ability of multiple GHMs to represent the influence of human activities on discharge and hydrological extremes in near-natural and managed catchments is missing. As a result, there is a limited understanding of whether (and where) the parameterizations of human activities in GHMs leads to an increase (or decrease) in model performance. To address this issue, the main objectives of this study are: (a) to evaluate the performance of five state-of-the-art GHMs that include the parameterizations of human activities in their modelling scheme; and (b) to compare the performance of these GHMs when run with and without human impact parameterizations.

2. Data and Methods

The overall methodological framework used in this study is shown in **figure 1**. In brief, the method involves three main steps: (1) obtaining river discharge from GHMs with human impact parameterizations (HIP) and without human impact parameterizations (NOHIP); (2) selecting observed river discharge data; and (3) evaluating model performance. Each of these steps is explained in the following subsections.

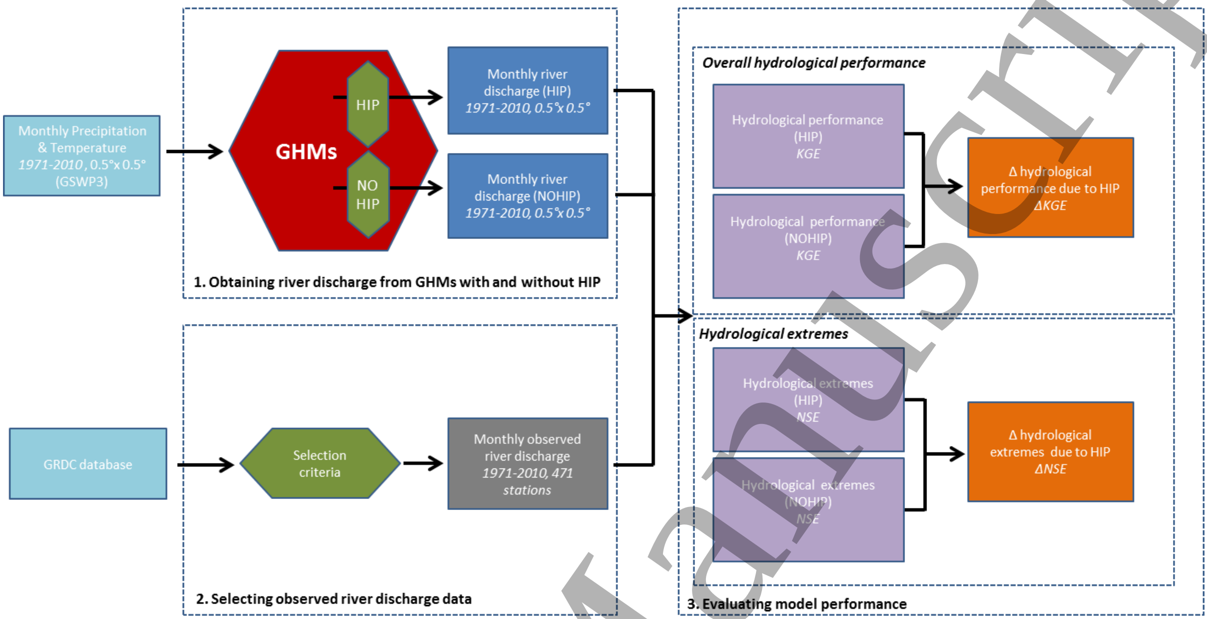

Figure 1: Flowchart of the methodological steps taken in this study. Steps 1, 2, and 3 correspond to paragraphs 2.1, 2.2 and 2.3.

2.1 Obtaining river discharge from GHMs with and without HIP

We used modelled monthly discharge (0.5° x 0.5° spatial resolution) for the period 1971–2010 from five GHMs: H08 (Hanasaki et al., 2008a,b), LPJmL (Bondeau et al., 2007; Rost et al., 2008; Schaphoff, et al., 2013), MATSIRO (Pokhrel, et al., 2012, 2015;Takata et al., 2003), PCR-GLOBWB (van Beek et al., 2011; Wada et al., 2011, 2014b), and WaterGAP2 (Müller Schmied et al., 2016). All simulations were carried out under the modelling framework of phase 2a of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2a: https://www.isimip.org/protocol/#isimip2a). For each GHM, we used two simulations: (1) HIP: a model run including time-varying land use and land cover change, historical dam construction and operation, irrigation, and upstream consumptive water abstractions; and (2) NOHIP: a ‘naturalized’ model run without HIP.

An overview of the model characteristics of each of the GHMs, and the methods used to parameterize hydrological processes and human impacts, can be found in **supplementary table 1**, and details on each GHM can be found in the individual model references provided therein. In the following subsections, we briefly outline the most important characteristics of the hydrological and human impacts parameterizations.

2.1.1 Parameterizations of hydrological processes

Each GHM in this study is forced with daily (MATSIRO: three-hourly) inputs from the GSWP3 historical climate data-set (http://hydro.iis.u-tokyo.ac.jp/GSWP3). The GHMs applied in this study differ in hydrological representation and parameterizations (**supplementary table 1.A**). H08 and MATSIRO model the energy balance explicitly and use the bulk formula in the evaporation scheme (Hanasaki et al., 2008a,b; Pokhrel, et al., 2012, 2015;Takata et al., 2003). LPJmL, PCR-GLOBWB, and WaterGAP2 do not include the energy balance explicitly and use the Priestley-Taylor and Hammon formulas in their evapotranspiration schemes (van Beek et al., 2011; Bondeau et al., 2007; Müller Schmied et al., 2014,2016; Schaphoff et al., 2013; Verzano et al., 2012; Wada et al., 2011).

To generate runoff, all GHMs use a saturation excess formula, although the formula is integrated differently in the various GHMs. Snow accumulation and melt are integrated in the modelling framework via the energy balance (H08, MATSIRO) or by means of a degree-day calculation method (LPJmL, PCR-GLOBWB, WaterGAP2). All GHMs use a linear reservoir method in their routing scheme. Whilst H08, LPJmL, and MATSIRO route with a constant flow velocity (based on Manning's Strickler), PCR-GLOBWB and WaterGAP2 use variable flow velocities. The number of soil layers and their depths vary significantly between GHMs, from one layer with varying depth (e.g. WaterGAP2, H08) to 12 fully resolved layers.

2.1.2 Parameterizations of human impacts

All GHMs use a combination of socioeconomic and hydro-climatological parameters to estimate sectoral water demands (Hanasaki et al., 2008a,b; Müller Schmied et al., 2016; Pokhrel, et al., 2015; Rost et al., 2008; Schaphoff, et al., 2013; Takata et al., 2003; Van Beek et al., 2011; Wada et al., 2014b). Livestock water needs (**supplementary 1.B**) are estimated by combining historical gridded livestock density maps with their species-specific water demands. Domestic water demands (**supplementary table 1.C**) are derived by applying a time-series regression at the country-scale, accounting for drivers like population and per capita GDP, and in some cases (PCR-GLOBWB) total electricity production, energy consumption, and temperature. Industrial water demands (**supplementary table 1.D**) are based on historical country-scale estimates from the WWDR-II dataset (Shiklomanov, 1997; Vorosmarty et al., 2005; WRI, 1998) and the FAO-AQUASTAT database (http://www.fao.org/nr/water/aquastat/dbase/index.stm), for PCR-GLOBWB and H08 respectively. WaterGAP2 simulates global thermoelectric water use using spatially explicit information on the location of power plants. Manufacturing water demand is simulated in WaterGAP2 for each country using its yearly Gross Value Added (GVA), and factors representing technological change and water use intensity. The models estimate irrigation water use (**supplementary table 1.E**) by multiplying the area equipped for irrigation with its utilization intensity, the total crop-specific water requirements – determined by the hydro-climatic conditions (temperature, precipitation, potential evapotranspiration, soil moisture, crop-growth curves, length and timing of the crop-growth season), and a parameter that accounts for the irrigation water use efficiency.

LPJmL, H08, and MATSIRO use surface water (first) to accommodate the sectoral water needs (**supplementary table 1.F**). WaterGAP2 uses the groundwater to fulfil water demands, and surface water is only used if enough is available. PCR-GLOBWB applies a share of readily available groundwater reserves, based on the ratio between simulated daily base-flow and long-term mean river discharge, to be used for consumptive water needs. The remainder of the water needs are fulfilled in PCR-GLOBWB by means of surface water. Whilst all GHMs deal consistently with return flows (**supplementary table 1.G**) for industry (surface water, same day), domestic (surface water, same day), and livestock (no return flow), returns from irrigation water use are incorporated differently. PCR-GLOBWB and H08 allow excess irrigation water return to the soil and groundwater layers by means of infiltration and additional recharge. LPJmL and MATSIRO return directly to the rivers, for which LPJmL uses a fixed ratio of 50%. Excess irrigation water in WaterGAP2 is returned to the surface waters using a cell-specific artificial drainage fraction, while the rest of the excess water is returned to groundwater.

All GHMs include either irrigation and/or non-irrigation purposes in their reservoirs schemes (**supplementary table 1.H**), and PCR-GLOBWB also includes flood control and navigation. The retrospective operation schemes of Hanasaki et al. (2006), Biemans et al. (2011), and Haddeland et al. (2006) form the basis of the reservoir operation schemes in most models. PCR-GLOBWB uses a prospective reservoir operation scheme that integrates efforts of Haddeland et al. (2006) and Adam et al. (2007). H08 is the only model that does not account for increased evapotranspiration over reservoirs.

2.2 Selecting observed river discharge data

Observed monthly river discharge data were taken from the Global Runoff Data Centre (GRDC, 56068 Koblenz, Germany). From the 9,051 gauging stations in the GRDC database, we selected stations that meet the following criteria: (1) a minimum of 5-year coverage (not necessarily consecutive) during the period 1971–2010 with a completeness of observations of ≥95%; and (2) a minimum catchment area of 9,000 km^2, to omit catchments whose hydrological processes cannot be adequately represented by models operating at 0.5° x 0.5° (Hunger and Döll, 2008). Finally, we discarded the stations for which the difference in catchment area in GRDC database and that estimated by using the DDM30 river routing network (Döll and Lehner, 2002) is >25%.

We then made a distinction between near-natural and managed catchments. Following Beck et al. (2016), a catchment is classified as near-natural if the share of land-area subject to irrigation is <2% and the total reservoir capacity is <10% of its long-term mean annual discharge. If these conditions are not met the catchment was classified as managed. The classification was based on the HYDE 3/MIRCA land cover dataset (Fader et al., 2010; Klein Goldewijk and Van Drecht, 2006; Portmann et al., 2010; Ramankutty et al., 2008) together with the Global Reservoir and Dam database (Lehner et al., 2011). Two stations shifted from near-natural to human impacted conditions between 1971 and 2010, and were discarded from further analysis.

The aforementioned steps resulted in 471 stations with a total catchment area covering 19.8% of the global land (**figure 2**), of which 92 are located at the outlet of a catchment area. The mean length of observations is 32.8 years for all stations. Of all stations, 226 are located in managed catchments and 245 in near-natural catchments. Of the stations located at the outlet of a catchment, 45 are managed (4.8% of the global land area), and 47 are near-natural (15.1% of the global land area).

Figure 2 shows that the majority of selected stations (blue) are located in Northern and Latin-America, Europe, Southern Africa, and Australia. The number of stations in Northern and Central Africa and Asia is relatively small. We selected 12 stations in river basins located in different geographic regions (green circles in **figure 2**: Amazonas, Amur, Colorado, Congo, Guadiana, Mackenzie, Murray, Ob, Rhine, Tocantins, Volga, and the Zambezi) for which a detailed analysis is provided in the **Supplementary results** section **(Supplementary)**.

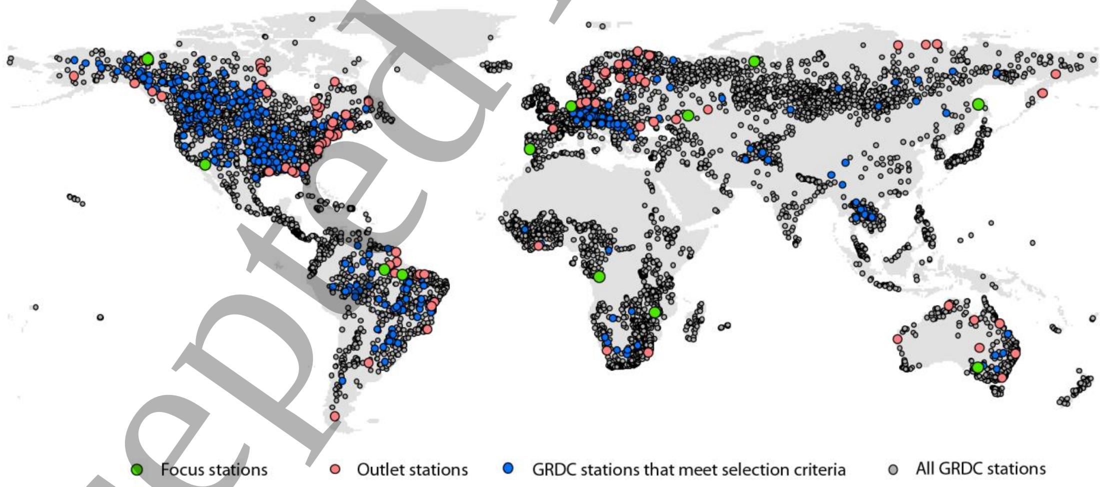

Figure 2: Spatial distribution of GRDC stations used for this study.
Each dot shows a GRDC station (n = 9,051) from the station catalogue. Blue dots indicate all GRDC stations (n = 471) that meet the selection criteria, whereas the red dots refer to the stations (n = 92) that are located at the outlet of a catchment. The green dots indicate those stations (n = 12) that were selected for detailed analyses.

2.3 Evaluating model performance

To evaluate the GHMs' simulation of monthly discharge and hydrological extremes under HIP and NOHIP conditions, we compared modelled results with observed river discharge data using several evaluation metrics described below. To ensure a consistent comparison between modelled and observed data, we only used modelled data for the same years for which observations were available. We also corrected modelled discharges for potential over-/underestimations caused by the difference in catchment size between model and GRDC. To do this, we used a multiplier that represents the difference in upstream area as reported by the GRDC and as estimated from the DDM30 network.

First, we applied the modified Kling-Gupta Efficiency index (KGE) with its sub-components: the linear correlation coefficient (rKGE); the bias ratio (βKGE); and the variability ratio (γKGE) (Gupta et al., 2009; Kling et al., 2012). The KGE is a widely applied indicator for the validation of hydrological performance in modelling studies at the global and regional scale and provides a good representation of the "closeness" of simulated discharges to observations (Huang et al. 2017, Kuentz et al., 2013; Nicolle et al., 2014; Revilla-Romero et al. ,2015; Thiemig et al., 2013, 2015; Thirel et al., 2015; Wöhling et al., 2013). Moreover, use of its three sub-components enables the identification of reasons for sub-optimal model performance (Gupta et al., 2009; Kling et al., 2012; Thiemig et al., 2013). This was achieved by estimating for each sub-parameter its distance to optimal performance, and by subsequently comparing these distances across the different sub-parameters. Statistical significance of the change in KGE outcomes due to the inclusion of HIP was tested by means of regular bootstrapping ($n = 1,000$, $p \leq 0.05$ (two-tailed)), following the method of Livezey and Chen (1982) and Wilks (2006).

Second, we applied the Nash-Sutcliffe Efficiency test (NSE, Nash and Sutcliffe, 1970) to evaluate the representation of Q_1 (high-flow) and Q_{99} (low-flow) conditions (e.g. Beck et al., 2017a; Blösch et al., 2013; Hejazi and Moglen, 2008; Mohamoud, 2008), obtained under fixed threshold level settings (van Loon, 2015). By means of a two-sample Kolmogorov-Smirnov (KS) test (Massey, 1951; $p \leq 0.05$) we tested how often HIP leads to significant changes in the fit of the full modelled exceedance probability curve for hydrological extremes compared to the full observed exceedance probability curve.

Table 1: The performance metrics used in this study and their calculation procedure.
Here, s_i and o_i are simulated and observed monthly discharge at station i; μ_s and μ_o are simulated and observed mean monthly discharge at station i; σ_s and σ_o are the standard deviation of the simulated and observed discharge at station i, respectively; Q_s and Q_o are the simulated and observed hydrological extremes.

Abbreviation	Name	Calculation procedure	Range and ideal value
KGE	Modified Kling-Gupta Efficiency Index	$KGE = 1 - \sqrt{(\mathrm{r}KGE^* - 1)^2 + (\beta KGE^* - 1)^2 + (\gamma KGE^* - 1)^2}$	-∞ - 1 (ideal value: 1)
rKGE	KGE correlation coefficient	$\mathrm{rKGE} = \frac{\sum_{i=1}^{n}(s_i - \mu_{s,i})(o_i - \mu_{o,i})}{\sqrt{\sum_{i=1}^{n}(s_i - \mu_{s,i})^2}\sqrt{\sum_{i=1}^{n}(o_i - \mu_{o,i})^2}}$	-1 - 1 (ideal value: 1)

	(Pearson)		
βKGE	KGE bias ratio	$\beta\text{KGE} = \mu_{s,i} / \mu_{o,i}$	0 - ∞ (ideal value: 1)
γKGE	KGE variability ratio	$\gamma\text{KGE} = \frac{\sigma_{s,i}/\mu_{s,i}}{\sigma_{o,i}/\mu_{o,i}}$	0 - ∞ (ideal value: 1)
NSE	Nash-Sutcliffe Model Efficiency	$NSE = 1 - \frac{\sum(Q_s - Q_o)^2}{\sum(Q_o - \overline{Q_o})^2}$	-∞ - 1 (ideal value: 1)
Q_1	High-flow indicator	Monthly discharge (m^3/s) that is exceeded on average in 1 out of 100 months	
Q_{99}	Low-flow indicator	Monthly discharge (m^3/s) that is exceeded on average in 99 out of 100 months	
KS	Two sample Kolmogorov-Smirnov test	[h, p] = kstest2(cdf(Q_s,), cdf(Q_o), 'Alpha',0.05)*	For p > 0.05 H_0 (the two cdfs come from the same distribution) is not rejected.

** Calculation procedure for the two-sample Kolmogorov-Smirnov test presented in the table is the Matlab function for the KS-test.*

3. Results

3.1 Validation and influence of human impact parameterizations on overall model performance

Including the parameterizations of human impacts in the GHMs leads to a large improvement in overall model performance. Hydrological performance under the HIP simulations shows a significant improvement compared to the NOHIP simulations for between 40.8% and 72.3% of the land area studied, depending on the GHM (**figure 3a**). For most GHMs, the positive effects of including HIP in the simulations outweigh the negative effects. This is the case for both near-natural and managed catchments, although the positive effects are more pronounced for the managed catchments (**figure 3a-d**). Near-natural catchments are only indirectly impacted by HIP, for example by receiving improved or altered water simulations from upstream managed catchments. The KGE sub-components show significant improvement in performance in large shares of the land area studied, especially for the bias and variability ratio. The bias ratio improves significantly for 36.1-73.0% of the total land area for all catchments, compared to 64.8-90.6% and 24.3-70.4% in managed and near-natural catchments respectively (**figure 3b**). For the variability ratio, improvements were found for 31.4-74.4% of land area for all catchments (48.9-92.6% for managed / 23.0-73.2% for near-natural) (**figure 3c**). The lowest improvements are found for the correlation coefficient, with improvements for 15.9-58.1% of total land area for all catchments (22.1-75.1% for managed /13.9-61.4% for near-natural) (**figure 3d**).

Results are shown for each station in **figure 4** for the overall model performance (KGE), and in **supplementary figure 1** for the KGE sub-parameters. The results show particularly strong improvements in overall performance in Latin America, Southern Africa, and Northwest U.S.. There are only a limited number of stations for which the inclusion of HIP leads to a significant decrease in overall hydrological performance for the majority of GHMs or where no to limited changes occur, for example in near-natural areas (e.g. the Amazonas).

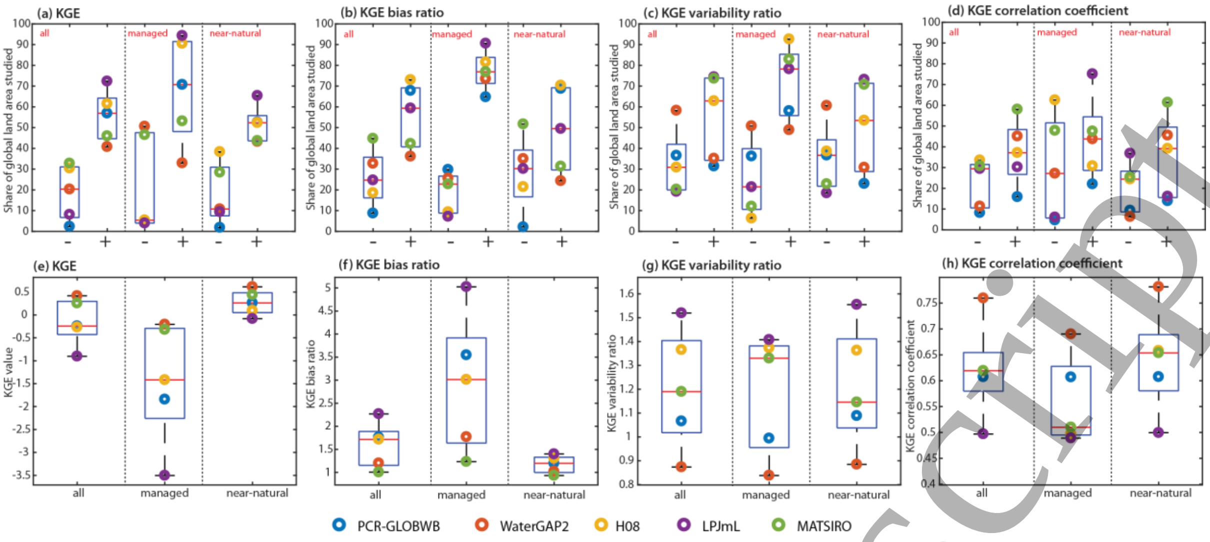

Figure 3: Global weighted-mean (improvement ('+') or deterioration ('-') in the) representation of hydrological performance due to HIP for all catchments, managed catchments, and near-natural catchments.
Figures 3a-d visualize for each GHM the share of land area with a significant change in overall hydrological performance due to the inclusion of HIP. Figures 3e-h indicate the globally weighted-mean hydrological performance after inclusion of HIP. On each box, the red mark indicates the median. The bottom and top edges of the box indicate the 25th and 75th percentiles of the model ensemble, respectively

When considering overall hydrological performance for each GHM under HIP conditions (**figure 3e**), WaterGAP2 and MATSIRO show the best performance globally. Even though the simulations with HIP include human impact parameterizations by definition, all GHMs still show better performance in near-natural catchments than in managed catchments (**figure 3e-h).** The KGE bias ratio values >1 indicate that all models systematically overestimate long-term mean monthly discharge (**figure 3f**), up to 5-fold for LPJmL in managed catchments. For the variability ratio (**figure 3g**), WaterGAP2 is the only GHM that tends to slightly underestimate variability (variability ratio <1) in monthly discharge, in both the managed and near-natural catchments. All other GHMs show overestimations, up to 1.55-fold for LPJmL for near-natural catchments. All GHMs show a reasonable correlation with observed monthly discharge estimates (**figure 3h**), with values ranging between 0.49 to 0.69 in the managed catchments and 0.50 to 0.79 in the near-natural catchments. The highest correlation coefficients including HIP are found for WaterGAP2, with a global mean value across all catchments of 0.76 (0.69 for managed catchments / 0.78 for near-natural catchments).

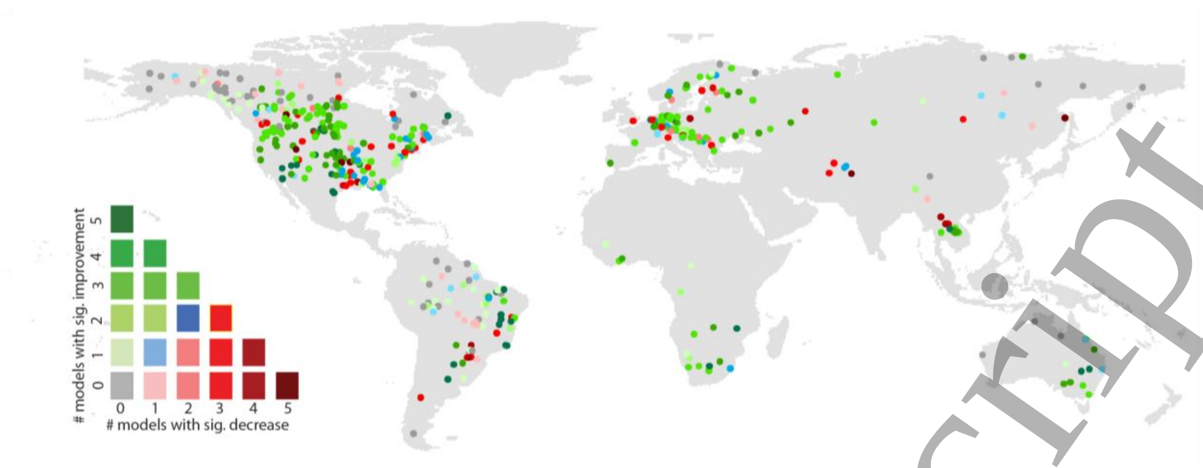

Figure 4: Number of GHMs with a significant improvement or deterioration in overall hydrological performance (KGE) due to inclusion of HIP.
Figures for the underlying KGE sub-parameters (bias ratio, variability ratio, correlation coefficient) are presented in supplementary figure 1. Supplementary figure 2 shows the KGE performance values per GHM under HIP conditions.

For each catchment (and therefore its associated land area), it is possible to distinguish which of the KGE sub-parameters contributes most to sub-optimal performance. These results are summarised in **figure 5**. The results show that under HIP conditions, the bias ratio contributes most to sub-optimal performance in managed catchments for most GHMs, except WaterGAP2 (for which the correlation coefficient contributes most). For near-natural catchments, sub-optimal performance is most often caused by the variability ratio for H08, LPJmL and WaterGAP2, by the bias ratio for MATSIRO, and by the correlation coefficient for PCR-GLOBWB.

Spatially explicit results vary per GHM and are shown in **supplementary figure 3**. The distribution of dominant contributors to the sub-optimal overall hydrological performance is similar for H08, LPJmL, and PCR-GLOBWB. For these GHMs, we find a dominant contribution of the bias ratio in Southern Africa, Australia, and inland U.S. Dominant contributions of the variability ratio and the correlation coefficient for these GHMs are found in Latin America, and at higher latitude and altitude regions. For Europe, the dominant contributions for H08, LPJmL, and PCR-GLOBWB are the variability ratio, the correlation coefficient, and the bias ratio respectively. The dominant contributors that cause sub-optimal overall hydrological performance for MATSIRO and WaterGAP2 are more equally distributed across the globe. While sub-components contribute to sub-optimal overall hydrological model performance for MATSIRO, it is predominantly the correlation coefficient and the variability ratio that determines the sub-optimal performance in WaterGAP2.

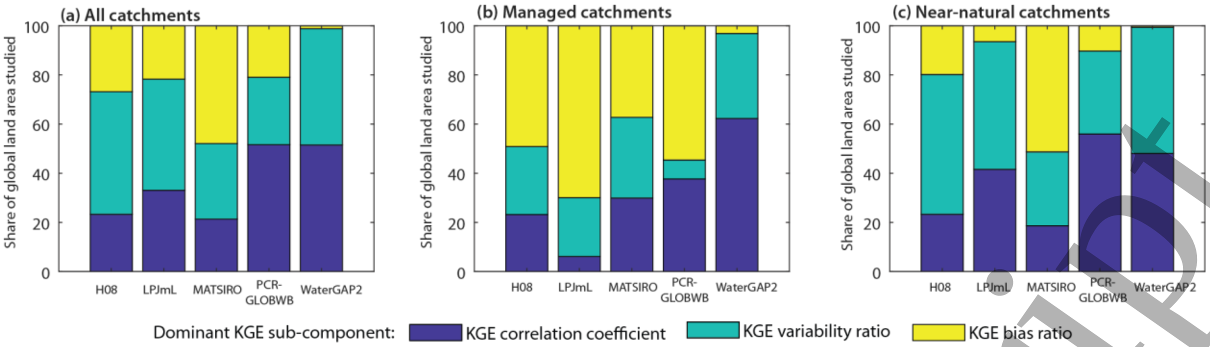

Figure 5: Share of land area with dominant contribution of the different KGE sub-components (KGE correlation coefficient, KGE variability ratio, KGE bias ratio) to sub-optimal overall hydrological performance under HIP conditions.

Supplementary figure 3 shows per model the spatial distribution of dominant KGE sub-components.

3.2 Validation and influence of human impact parameterizations on the simulation of hydrological extremes

The inclusion of HIP in the simulations affects the ability of GHMs to estimate hydrological extremes correctly in the majority of the land area studied (**figure 6**). The inclusion of HIP leads to better model performance for all GHMs, across a substantial share of the land area studied (**figure 6a-b**). For high-flows, HIP improves model performance significantly across 34.6-77.0% of the land area for all catchments (36.4-94.7% for managed / 24.1-79.2% for near-natural). For low-flows, HIP improves model performance significantly across 39.4-80.4% of the land area for all catchments (29.3-81.8% for managed / 42.7-90.3% for near-natural). The KS-test results (**supplementary figure 4**) show that HIP only leads to significant changes in the representation of the exceedance probability curve in a limited number of cases for H08 and LPJmL (up to 14.1% of the land area studied), predominantly in managed catchments.

Overall, hydrological extremes are represented reasonably well under HIP conditions, with globally weighted-mean NSE values ranging between 0.80-0.98 for high-flows, and 0.84-0.98 for low-flows (**figure 6c-d**). However, there is a significant difference in the ability of the GHMs to represent hydrological extremes between managed and near-natural catchments.

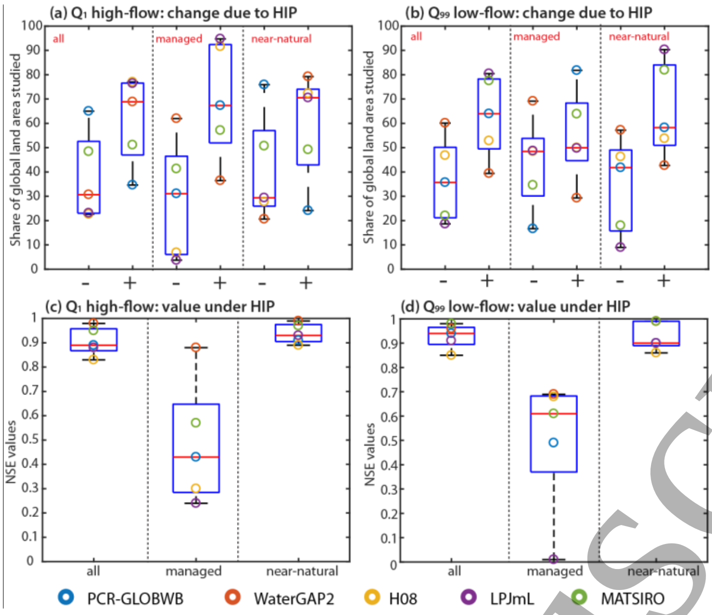

Figure 6: Global weighted-mean (improvement ('+') or deterioration ('-') in the) representation of hydrological extremes (Q_1 high-flow and Q_{99} low-flows) due to HIP, for all catchments, managed catchments, and near-natural catchments respectively.
On each box, the red mark indicates the median. The bottom and top edges of the box indicate the 25th and 75th percentiles of the model ensemble, respectively

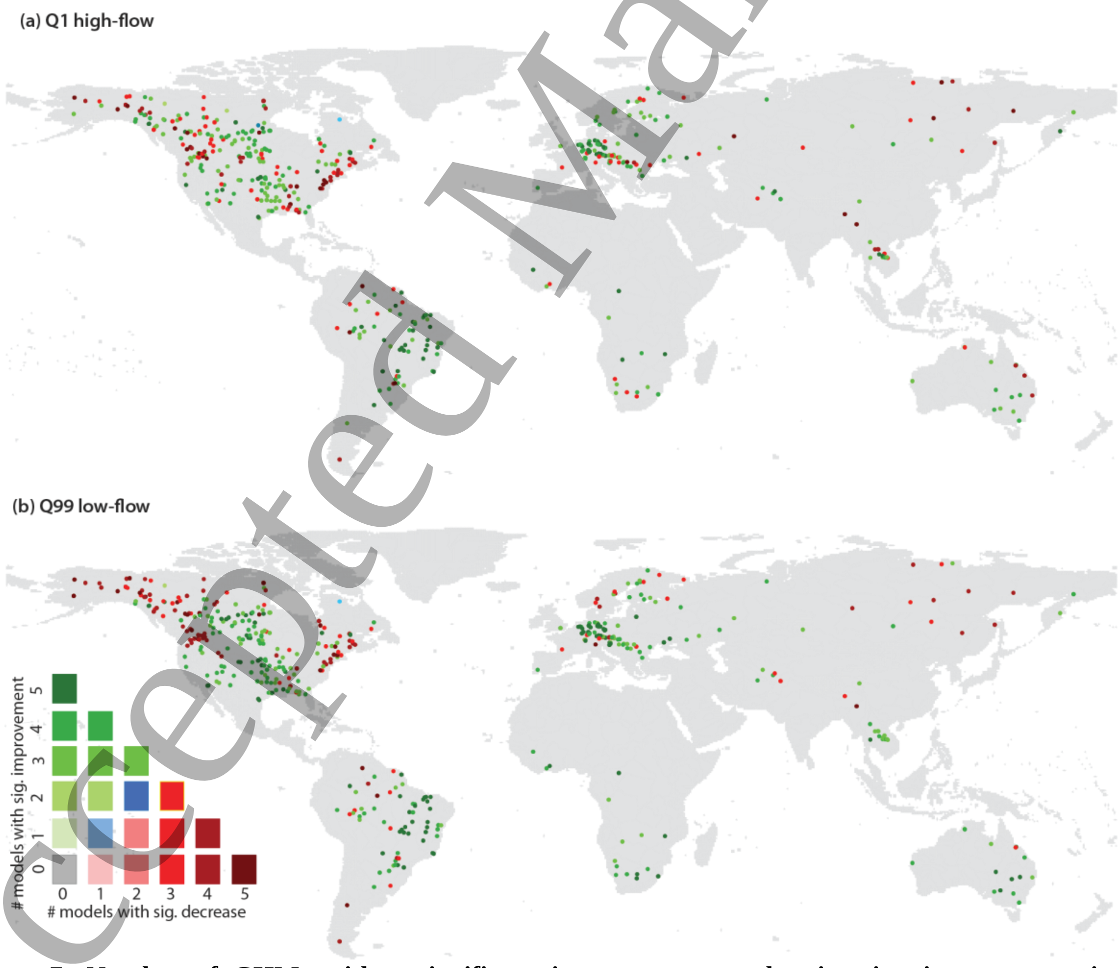

Figure 7: Number of GHMs with a significant improvement or deterioration in representation of hydrological extremes due to inclusion of HIP.

Figure 7 indicates that for the majority of stations, the inclusion of HIP leads to an improvement in the representation of hydrological extremes, for most GHMs. A deterioration in the representation of hydrological extremes across the majority of GHMs as a result of the inclusion of HIP was only found in selected areas, for example at higher latitudes and along the east-coast of the U.S.. When comparing the results for the Q_1 high-flows with the Q_{99} low-flows, no large differences in the spatial distribution of the number of GHMs with a significant improvement or deterioration are found.

The effects of HIP on the magnitude of extreme discharge differ for low-flows and high-flows (**supplementary figure 5**). Whilst the magnitude of high-flows mostly decreases with the inclusion of HIP, the effects on the magnitude of low-flows are both positive and negative. The convergence of results towards higher observed discharges, in both high- and low-flow estimates (as identified for all models in **supplementary figure 5)**, indicates that HIP becomes less important for the correct representation of hydrological extremes with increasing discharge volumes.

4. Discussion

Our results show that including HIP in GHMs generally improves the overall hydrological performance of the models, as well as their representation of hydrological extremes. However, we also show that further improvements are needed. In this section, we discuss: (1) possible reasons for the improved model performance due to HIP; (2) the main limitations of the current modelling frameworks and their representation of HIP, and potential ways to improve them; and we reflect on (3) general limitations in the current study design and provide suggestions for further research.

4.1 Improvements in model performance due to HIP and challenges ahead

Whilst the inclusion of HIP predominantly leads to the largest improvements in simulated discharge in the managed catchments, simulated discharge is also improved in a large share of the near-natural catchments. Improvements in model performance associated with the inclusion of HIP can be attributed to improvements in the different KGE sub-components, and in turn to different model components parameterizing the hydrological and human processes. In addition, insights into those factors bounding the optimal hydrological model performance under HIP conditions may help to identify priorities for further model improvement.

4.1.1 Representation of long-term mean discharges (bias ratio)

Our study shows that the representation of long-term mean discharges significantly improved with the inclusion of HIP, especially in managed catchments. Inclusion of HIP generally results in lower simulated discharges. As most GHMs systematically overestimate river discharges in the NOHIP simulation, this results in an improved performance. When HIP is included, we only find a deterioration in the bias ratio in selected higher latitude/altitude regions, where discharges are underestimated; this finding is in line with outcomes of single-model studies performed by Döll et al. (2009), De Graaf et al. (2014), and Haddeland et al. (2006). Improvements in bias ratios due to the inclusion of HIP can be attributed to the inclusion of water abstractions and return flows (**supplementary table 1.B-G**), and the incorporation of irrigated areas and irrigation rules, which influence evapotranspiration rates and the generation of runoff (**supplementary table 1.E**).

However, despite improvement in the bias ratio with the inclusion of HIP, this KGE sub-indicator contributes most to sub-optimal performance in managed catchments for H08, LPJmL, MATSIRO, and PCR-GLOBWB under HIP conditions. As the GHMs continue to overestimate long-term mean discharges in most cases under HIP conditions, future model improvements should be targeted to correcting this bias in these locations. This may be achieved by critically revisiting the methods used to represent evapotranspiration rates (**supplementary table 1.A**), runoff generation processes (**supplementary table 1.A**) and the level of water abstractions in managed catchments (**supplementary table 1.B-E**). The relatively good performance of WaterGAP2, in which biases in long-term mean annual discharge are adjusted using a parameter that determines the portion of effective precipitation that becomes surface runoff (Müller Schmied et al., 2014), highlights the potential importance of including a calibration routine (**supplementary table 1.I**). Calibration is also performed for H08, but this calibration aims to minimize runoff bias by modifying two parameters of subsurface flow for four climatic groups (Hanasaki et al., 2008a,b); it is therefore less effective in minimizing the bias ratio under HIP conditions.

4.1.2 Representation of hydrological variability (variability ratio)

The inclusion of HIP leads to mixed results regarding the representation of hydrological variability. Whilst HIP improved the representation of variability in some catchments and for some GHMs, it deteriorated the representation of variability for others. For example, it led to improvements in west-coast U.S., Southern Africa, and Australia, but a deterioration for most GHMs in Europe and inland U.S.. Similar results were found by Biemans et al. (2011), De Graaf et al. (2014), and Masaki et al. (2017) for a selection of catchments. Changes in the variability ratio due to the inclusion of HIP are predominantly driven by the timing of water abstractions and return flows, as well as by reservoir operation rules (**supplementary table 1.F-H**). These human activities influence the relative size of high- and low-flows compared to their long-term mean discharge values.

The variability ratio is the KGE sub-parameter that contributes most to the sub-optimal performance in near-natural catchments with the inclusion of HIP, for H08, LPJmL, and WaterGAP2. These GHMs significantly overestimate hydrological variability in near-natural catchments (except WaterGAP2, which underestimates variability in managed and near-natural catchments), and model improvement should therefore focus on better representing the speed of hydrological response, e.g.

through an improved representation of the soil moisture storage capacity or the ratio between surface and sub-surface runoff (**supplementary table 1.A**). In those cases where the variability ratio is also the KGE sub-parameter that contributes most to sub-optimal performance in managed catchments, model improvement should target the timing of water abstractions, return flows, and reservoir management (**supplementary table 1.F-H**).

4.1.3 Representation of the goodness-of-fit (correlation coefficient)

The inclusion of HIP only led to improved correlation coefficients in limited cases, and often resulted in a deterioration, even in managed catchments. Correlation coefficients between observed and modelled discharges, which are predominantly determined by the hydro-meteorological forcing data (Döll et al., 2016; Beck et al., 2016), were found to be generally high under both HIP and NOHIP conditions. Perturbations of the hydrological cycle due to human activitiesleading to changes in the timing of discharges and in the shape of the hydrograph, like return flows and reservoir operations, explain the observed decrease in the correlation coefficient in a substantial share of catchments and models globally (**supplementary table 1.F-H**).

Under HIP conditions, the correlation coefficient is the KGE sub-parameter that contributes most to sub-optimal performance only in PCR-GLOBWB for near-natural catchments and WaterGAP2 for managed catchments. It should be acknowledged, though, that correlation coefficients for PCR-GLOBWB and WaterGAP2 are relatively high, especially compared to the other GHMs. The relatively low correlation coefficients in near-natural catchments found at higher latitudes in all models may be addressed by critically reviewing the snow accumulation and melt processes in the GHMs (**supplementary table 1.A**). Higher correlation coefficients in the managed catchments may be established by improving the timing and quantification of return flow estimates and the representativeness of reservoir operations (**supplementary table 1.F-H**).

4.1.4 Representation of hydrological extremes

The inclusion of HIP also led to significant changes in the ability of most GHMs to represent hydrological extremes (both high- and low-flows), although the strength of this change is very much dependent on the location and GHM in question. Whilst the magnitude of high-flow estimates mainly decreased due to the inclusion of HIP, low-flow estimates showed mixed results. This is because the impacts of human activities tend to be greater for lower discharges, as the relative 'size' of human perturbations (such as water abstractions, return flows, or delayed releases of water via reservoir operations) is higher as a percentage of overall discharge when flows are low. Both De Graaf et al. (2014) and Wada et al. (2013a) found similar results when investigating hydro-climatic extremes. However, even with inclusion of HIP, the representation of hydrological extremes is sub-optimal. Future model improvements should aim to better characterize these extremes and to improve the representation of human activities during extreme hydrological conditions.

4.2 Limitations and further research

As the GHMs have very different parameterizations of hydrological and human processes, the current study does not allow a systematic assessment of specific cause-effect relations between HIP and the observed improvements in performance (Döll et al., 2016; Haddeland et al., 2014; Hagemann et al., 2013; Schewe et al., 2014; Beck et al., 2016). To do this, a substantial Monte-Carlo analysis would be required, whereby individual parameters and combinations of parameters are systematically modified for all GHMs (Döll et al., 2016). Undertaking such an analysis in parallel for the different GHMs incorporated is computationally expensive and requires a strict modelling-protocol. It may provide, however, additional information on how to adapt and improve the individual models and would be a valuable addition to the results presented in this study.

When interpreting the results of this study one must take into account that we only evaluated the GHMs with respect to monthly discharge. Whilst monthly discharge may be sufficient for the assessment and management of low-flows, droughts, and freshwater resource availability, flood risk assessment and management require information on daily peak discharge. Further research should therefore attempt to validate GHMs using daily peak discharge and assess how daily peak discharge is affected by the inclusion of HIP.

The spatial resolution of the GHMs applied in this study is 0.5° x 0.5° (~50 km x 50 km at the equator), dictated by the resolution of the GSWP3 input dataset. At a 0.5° spatial resolution hydrological processes are often represented by GHMs in a simplified or generalized form not fit for local applications (Bierkens, 2015). To account for this, we applied a minimum catchment size of 9,000 km^2, thereby omitting catchments too small to be adequately represented by GHMs (Hunger and Döll, 2008). Newer versions of several of the GHMs now operate at higher resolutions; for example WaterGAP and PCR-GLOBWB have recently published 5-min/6-min versions respectively (Verzano et al., 2012; Wada et al., 2016b). Future research could investigate whether the inclusion of these high-resolution model-runs improves the representation of discharges and hydrological extremes in the selected catchments and whether these high-resolution runs also allow for the inclusion of smaller catchments.

In this study, a relatively simple distinction was made between managed and near-natural catchments using two parameters: irrigated agriculture and reservoirs. These parameters were chosen as they have been reported to be the most significant human parameters on river hydrology (Beck et al., 2016, 2017a). However, to make a more detailed distinction between catchments that are impacted by human activities and those that are not, future studies could consider incorporating additional criteria, such as the share of sectoral water abstractions and return flows, and the share of built-up land area.

Additional catchment descriptors (Eisner, 2016), like climate conditions and physiographic properties of the drainage area, could also be applied to further assess the important controls on modelled discharges.

When evaluating the impact of HIP on hydrological extremes we only incorporated results for the Q_1 high-flow and Q_{99} low-flow. In this study we did not consider other ranges of the extreme value distribution explicitly. Although the inclusion of HIP shows influences these hydrological extremes substantially, we found very few instances in which this led to a significant change in the full exceedance probability curve . Future research should therefore also incorporate other ranges of the probability exceedance curve in order to do a full assessment of the influence of HIP on high- and low-flow extremes.

Next to the parameterizations and representation of hydrological processes and human impacts, other sources contribute to the uncertainty in the modelling of discharges and hydrological extremes. , These include the quality of, and uncertainties in, input data and observation datasets, and the calibration/validation strategy (Döll et al., 2016; Sood and Smakhtin, 2015). The quality of the selected forcing data, for example, may limit the representation of monthly discharges and hydrological extremes significantly (Döll et al., 2016; Beck et al., 2016), but has not been evaluated explicitly in this study. However, climate forcing uncertainty is probably a dominant driver for model outputs (Müller Schmied et al 2014, 2016). A benchmarking of the GSWP3 dataset against historical observations of precipitation and temperature, or against other forcing datasets (e.g. similar to Beck et al., 2017b; Sun et al., 2017), may therefore be of added value.

Differences in the quality and trustworthiness of the historical discharge observations (e.g. due to sampling, measurement, and interpretation errors), may potentially result in artificial biases in the validation results (Renard et al., 2010). The spatial representativeness of our results is limited by the availability of consistent publicly available in situ observations of sufficient quality. Future research should therefore consider extending the GRDC data-points with regional repositories of observed discharges, such as recently attempted by Beck et al. (2016), Do et al. (2017), and Gudmundsson et al. (2017). However, increasing the spatial representation comes at the cost of consistency, and special attention should be paid to the harmonization of these different databases. The use of remotely sensed data could also provide a valuable way of carrying out calibration and validation in ungauged regions (Döll et al., 2014a,b; Scanlon, et al. 2018). Remotely sensed data can also be of added value in: the assessment of the water consumed by agricultural irrigation (Peña-Arancibia et al., 2016), operational drought monitoring and early warning (Ahmadalipour et al., 2017); and the estimation of terrestrial water budgets (Zhang et al., 2017). Moreover, a clear potential exists for the assimilation of remotely sensed data into models (Eicker et al., 2014).

Calibration and validation are essential for compensating for factors such as the impossibility to measure all required model parameters at the applied scale, the lack of process understanding, the simplistic process representation in GHMs, and errors in forcing data (Beck et al., 2016; Bierkens, 2015; Döll et al., 2016; Liu et al., 2017). Hence, calibration/validation is key for realistic model performance. It should be acknowledged, though, that the representation of hydrological and/or human processes is artificially altered by means of calibration/validation processes and that a limited calibration may introduce uncertainties to the model output (Sood and Smakhtin, 2015). Before using any calibrated/validated model-data one should therefore critically reflect on whether the calibration/validation procedure executed, together with their optimization objectives, are fit for the specific application in-mind.

5. Summary and conclusions

This study shows that the inclusion of human activities in GHMs can significantly improve the simulation of monthly discharges and hydrological extremes, for the majority of catchments studied. The finding is robust across both managed and near-natural catchments. The global and spatially distributed results presented in this study indicate that the inclusion of human impact parameterizations is associated with improvements in the bias ratio and the variability ratio. Whilst the biases in long-term mean monthly discharge decrease significantly in 36.1-73.0% of the studied catchments due to the inclusion of HIP, the modelling of hydrological variability improves significantly in 31.4-74.4% of the catchments. Estimates of hydrological extremes are also significantly influenced by the inclusion of HIP, although the influence is highly dependent on the location and GHM in question. While HIP generally leads to a decrease (and thus improvement) in the absolute magnitude of simulated high-flows, its impact on low-flows is mixed.

Even when human activities are included in GHMs, their performance is still limited; this is particularly the case in managed catchments Moreover, the systematic misrepresentation of hydrological extremes across all GHMs calls for a careful interpretation of risk assessments based on their results, and further study into the overarching research theme of water resources, hydrological extremes, human interventions, and feedback linkages. The large variation in performance between GHMs, regions, and performance indicators, highlights the importance of a careful selection of models, model components, and evaluation metrics in future model applications. For example, for a study of droughts it is essential to correctly represent hydrological variability, whilst to study water scarcity it is crucial to minimize biases.

Sub-KGE results, which were presented in this study for each GHM, allow for the attribution of different hydrological and human impact model-components limiting optimal hydrological performance. In most GHMs model performance is limited due to the overestimation of long-term mean discharges. The correlation coefficient is the limiting factor for optimal model performance for WaterGAP2, despite the high correlation coefficients that were found for this model relative to the other GHMs studied. A better understanding of these factors, as provided by this study, may assist in the identification of priorities for further model improvement.

Acknowledgements

The Global Runoff Data Centre (GRDC, 56068 Koblenz, Germany) are thanked for providing the observed discharge data. This work has been conducted under the framework of phase 2 of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2a: www.isimip.org) and the authors want to thank the coordination team responsible for bringing together the different global hydrological modelling groups and for coordinating the research agenda, which resulted in this manuscript. The research leading to this article is partly funded by the EU 7th Framework Programme through the project Earth2Observe (grant agreement no. 603608). JZ was funded by the Islamic Development Bank. PJW received additional funding from the Netherlands Organisation for Scientific Research (NWO) in the form of a Vidi grant (016.161.324). J.C.J.H.A. received funding from the Netherlands Organisation for Scientific Research (NWO) VICI (grant no. 453-14-006). YM was supported by the Environment Research and Technology Development Fund (S-10) of the Ministry of the Environment.

References

Adam, J. C., Haddeland, I., Su, F., & Lettenmaier, D. P. (2007). Simulation of reservoir influences on annual and seasonal streamflow changes for the Lena, Yenisei, and Ob' rivers. *Journal of Geophysical Research Atmospheres*, *112*, 1–22.

Ahmadalipour, A., Moradkhani, H., Hongxiang, Y., & Zarekarizi, M. (2017). Remote Sensing of Drought: Vegetation, Soil Moisture, and Data Assimilation. In V. Lakshmi (Ed.), *Remote Sensing of Hydrological Extremes* (pp. 121–149). Switzerland: Springer Remote Sensing/Photogrammetry.

Alcamo, J. & Gallopin, G. (2009) .The United Nations World Water Assessment Programme: Building a second generation of world water scenarios, Paris, France.

Beck, H. E., Van Dijk, A. I. J., De Roo, A., Miralles, D. G., McVicar, T. M., Schellekens, J. & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resour. Res.* 52, 3599–3622.

Beck, H. E., Van Dijk, A. I. J., De Roo, A., Dutra, E., Fink, G., Orth, R., Schellekens, J. (2017a). Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrol. Earth Syst. Sci.*, 21, 2881-2903.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I. J., Weedom, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., Wood, E. F. (2017b). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modelling. *Hydrol. Earth Syst. Sci.*, 21, 6201-6217.

van Beek, L.P.H., Wada, Y. & Bierkens, M.F.P. (2011). Global monthly water stress: I. Water balance and water availability. *Water Resour. Res.* 47, W07517.

Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R.W.A., Heinke, J., Von Bloh, W. & Gerten, D. (2011). Impact of reservoirs on river discharge and irrigation water supply during the 20th century. *Wat. Resour. Res.* 47, W03509.

Bierkens, M.F.P. (2015). Global hydrology 2015: State, trends, and directions. *Water Resour. Res*, 51, 4923-4947.

Blösch, G., Sivapalan, M., Wagener, T., Viglione, A., Savenije, H. (2013). Runoff predictions in ungauged basins: synthesis across processes, places and scales. Cambridge University Press, UK.

Bondeau, A. *et al.* (2007). Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology,* 13, 679–706.

Do, H. X., Gudmundsson, L., Leonard, M., Westra, S. & Seneviratne, S.I. (2017). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of daily streamflow archive and metadata. *Earth Syst. Sci. Data Discuss*.

Döll, P. & Lehner, B. (2002). Validation of a new global 30-min drainage direction map. *J. Hydrol.*, 258, 214-231.

Döll, P. & Siebert, S. (2002). Global modelling of irrigation water requirements. *Water Resour. Res*, 38, W1037.

Döll, P., Kaspar, F. & Lehner, B. (2003). A global hydrological model for deriving water availability indicators: model tuning and validation. *J. Hydrol.*, 270, 105-134.

Döll, P., Fiedler, K. & Zhang, J. (2009). Global-scale analysis of river flow alterations due to water withdrawals and reservoirs. *Hydrol. Earth Syst. Sci.* 13, 2413-2432.

Döll, P., & Müller Schmied, H. (2012). How is the impact of climate change on river flow regimes related to the impact on mean annual runoff? A global-scale analysis. *Environ. Res. Lett.*, 7, 014037.

Döll, P., Fritsche, M., Eicker, A. & Müller Schmied, H. (2014a). Seasonal Water Storage Variations as Impacted by Water Abstractions: Comparing the Output of a Global Hydrological Model with GRACE and GPS Observations. *Surv. Geophys*., 35(6), 1311-1331.

Döll, P., Müller Schmied, H., Schuh, C., Portmann, F.T. & Eicker, A. (2014b). Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Water Resour. Res*., 50, 5698-5720.

Döll, P., Douville, H., Güntner, A. , Müller Schmied, H. & Wada, Y. (2016). Modelling Freshwater Resources at the Global Scale: Challenges and Prospects. *Surv. Geophys*., 37(2), 195-221.

Eicker, A., Schumacher, M., Kusche, J., Döll, P. & Müller Schmied, H. (2014). Calibration/data assimilation approach for integrating GRACE data into the WaterGAP Global Hydrology Model (WGHM) using an Ensemble Kalman Filter: First results. *Surveys in Geophysics* 35 (6), 1285-1309

Eisner, S. (2016). Comprehensive evaluation of the WaterGAP3 Model across climatic, physiographic, and anthropogenic gradients. PhD thesis (http://nbn-resolving.de/urn:nbn:de:hebis:34-2016031450014)

Fader, M., Rost, S., Muller, C., Bondeau, A., Gerten, D. (2010). Virtual water content of temperate cereals and maize: Present and potential future patterns. *J. Hydrol.*,384 (3-4), 218-231.

Flörke, M., Kynast, E., Bärlund, I., Eisner, S., Wimmer, F. & Alcamo, J. (2013). Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study. *Global Environ. Change*, 23, 144–156.

Fujimori, S., Hanasaki, N., Masui, T. (2017). Projections of industrial water withdrawal under shared socioeconomic pathways and climate mitigation scenarios. *Sust. Sci.*, 12, 275-292.

Gosling, S.N. *et al.* (2017). A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 C, 2 C, and 3 C. *Clim. Change*, 141, 577-595.

Gupta, H.V., Kling, H., Yilmaz, K.K. & Martinez, G.F. (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *J. Hydrol.*, 377 (1–2), 80–91.

de Graaf, I.E.M., Van Beek, L.P.H., Wada, Y. & Bierkens, M.F.P. (2014). Dynamic attribution of global water demand to surface water and groundwater resources: Effects of abstractions and return flows on river discharges. *Adv. Water Resourc.*, 64, 21-33.

Gudmundsson, L., *et al.* (2011). Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. *J. Hydrometeor.*, 13, 604 – 620.

Gudmundsson, L., Wagener, T., Tallaksen, L.M. & Engeland, K. (2012). Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. *Wat. Resour. Res.*, 48, W11504.

Gudmundsson, L., Do, H.X., Leonard, M., Westra, S. & Seneviratne, S. I. (2017). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment. *Earth Syst. Sci. Data Discuss*.

Haddeland, I., Skaugen, T. & Lettenmaier , D. P. (2006). Anthropogenic impacts on continental surface water fluxes. *Geophys. Res. Lett.*, 33, L08406.

Haddeland, I., Skaugen, T., & Lettenmaier, D. P. (2007). Hydrologic effects of land and water management in North America and Asia: 1700–1992. *Hydrol. Earth Syst. Sci.* 11, 1035–1045.

Haddeland, I. *et al.* (2014). Global water resources affected by human interventions and climate change. *Proc. Natl. Acad. Sci. U. S. A.*, 111, 3251–3256.

Hagemann, S. *et al*. (2013). Climate change impact on available water resources obtained using multiple global climate and hydrology models. *Earth Syst. Dynam.* 4, 129-144.

Hallegatte, S., Bangalore, M., Bonzanigo, L., Fay, M., Kane, T., Narloch, U., Rozenberg, J., Treguer, D., Vogt-Schilb, A. (2016). Shock Waves: Managing the Impacts of Climate Change on Poverty. Climate *Change and Development Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/22787*

Hallegatte, S., Vogt-Schilb, A., Bangalore, M. & Rozenberg, J. (2017). Unbreakable: Building the Resilience of the Poor in the Face of Natural Disasters. *Climate Change and Development;.Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/25335*

Hanasaki, N., Kanaeo, S., Oki, T. (2006). A reservoir operation scheme for global river routing models. *J. Hydrol.*, 327, 1-2, 22-41.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y. & Tanaka, K. (2008a). An integrated model for the assessment of global water resources – Part 1: Model description and input meteorological forcing. *Hydrol. Earth Syst. Sci.*, 12, 1007–1025.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y. & Tanaka, K. (2008b). An integrated model for the assessment of global water resources – Part 2: Applications and assessments. *Hydrol. Earth Syst. Sci.*, 12, 1027–1037.

Hanasaki, N. *et al*. (2013). A global water scarcity assessment under Shared Socio-economic Pathways – Part 2: Water availability and scarcity. *Hydrol. Earth Syst. Sci.*, 17, 2393-2413.

Hejazi, M. I., Moglen, G.E. (2008). The effect of climate and land use change on flow duration in the Maryland Piedmont region. *Hydrol. Proc.*, 22, 4710-4722.

Huang, S., Kumar, R., Flörke, M., Yang, T., Hundecha, Y., Kraft, P., Gao, C., Gelfan, A., Liersch, S., Lobanova, A., Strauch, M., Van Ogtrop, F., Reinhardt, J., Haberlandt, U., Krysanova, V. (2017). Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide. *Clim. Change*, 141, 381-397.

Hunger, M. & Döll, P. (2008) Value of river discharge data for global-scale hydrological modelling. *Hydrol. Earth Syst. Sci.*, 12, 841-861.

Van Huijgevoort, M. H. J., Hazenberg, P., Van Lanen, H. A. J., Teuling, A. J., Clark, D. B., Folwell, S., Gosling, S. N., Hanasaki, N., Heinke, J., Koirala, S., Stacke, T., Voss, F., Sheffield, J. & Uijlenhoet, R. (2013). Global multimodel analysis of drought in runoff for the second half of the twentieth century. *J. Hydrometeor.*,14, 1535-1552

IPCC (2007). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland.

IPCC (2013). Climate change 2013: the physical science basis. IPCC, Geneva, Switzerland.

Klein Goldewijk, K. & Van Drecht, G.(2006). HYDE 3: Current and historical population and land cover. MNP (2006) (Edited by A.F. Bouwman, T. Kram and K. Klein Goldewijk), Integrated modelling of global environmental change. An overview of IMAGE 2.4. Netherlands Environmental Assessment Agency (MNP), Bilthoven, The Netherlands.

Kling, H., Fuchs, M. & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J. Hydrol.*, 424–425, 264–277.

Kuentz, A., Mathevet, T., Gailhard, J., Perret, C., Andreassian, V. (2013). Over 100 years of climatic and hydrologic variability of a Mediterranean and mountainous watershed: the Durance river. In: *Cold and Mountain Region Hydrological Systems Under Climate Change: Towards Improved Projections* Proceedings of H02, IAHS-IAPSO-IASPEI Assembly, Gothenburg, Sweden. IAHS publication, 360.

Kummu, M., Guillaume, J. H. A., De Moel, H., Eisner, S., Flörke, M., Porkka, M., Siebert, S., Veldkamp, T. I. E., Ward, P.J. (2016) The world's road to water scarcity: shortage and stress in the 20th century and pathways towards sustainability. *Sci. Rep.*, 6, 38495.

Lehner, B., *et al.* (2011). High resolution mapping of the world's reservoirs and dams for sustainable river flow management. *Front. Ecol. Environ.*, 9, 494-502.

van Loon, A.F. (2015). Hydrological drought explained. *WIREs Water*, 2, 359-392.

van Loon, A.F., *et al.* (2016). Drought in the Anthropocene. *Nat. Geoscience*, 9, 89-91.

Livezey, R. E., & Chen, W. Y. (1982). Statistical field significance and its determination by monte carlo techniques. *Mon. Weather Rev.*, 111, 46–59.

Liu, X., Tang, Q., Cui, H., Mu, M., Gerten, D., Gosling, S., Masaki, Y., Wada, Y. & Satoh, Y. (2017). Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations. *Environ. Res. Lett.*, 12, 025009.

Masaki, Y., *et al.* (2017). Intercomparison of regulated river discharge among multiple global hydrological models under multiple forcings – Part II: Multiple models in two case-study river basins, Missouri-Mississippi and Green-Colorado. *Environ. Res. Lett.*, 12, 055002.

Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46, 68–78.

Mohamoud, Y.M. (2008). Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. *Hydrol. Sci. J.*, 53, 706-724.

Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F.T., Flörke, M. & Döll, P. (2014). Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. *Hydrol. Earth. Syst. Sci*., 18, 3511-3538.

Müller Schmied, H. *et al.* (2016). 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.

Munia, H., Guillaume, J.H.A., Mirumachi, N., Porkka M., Wada,Y. & Kummu, M. (2016). Water stress in global transboundary river basins: significance of upstream water use on downstream stress. *Environ. Res. Lett.*,11, 014002.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.*,*10*(3), 282-290.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J.-M., Viel. C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., Morice, E. (2014). Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. *Hydrol. Earth Syst. Sci*., 18, 2829-2857.

Padowski, J.C., Gorelick, S.M., Thompson, B.H., Rozelle, S. & Fendorf, S. (2015). Assessment of human-natural system characteristics influencing global freshwater supply vulnerability. *Environ. Res. Lett*., 10, 104014.

Peña-Arancibia, J. L., Mainuddin, M., Kirby, J. M., Chiew, F. H. S., McVicar, T. R., & Vaze, J. (2016). Assessing irrigated agriculture's surface water and groundwater consumption by combining satellite remote sensing and hydrologic modelling. *Sci. Total Environ.*, *542*, 372–382.

Pokhrel, Y., Hanasaki, N., Koirala, S., Cho, J., Yeh, P. J.-F., Kim, H., Kanae, S. & Oki, T. (2012). Incorporating anthropogenic water regulation modules into a land surface model. *J. Hydrometeor,* 13(1), 255–269.

Pokhrel, Y. N., Koirala, S., Yeh, P.J.-F., Hanasaki, N., Longuevergne, L., Kanae, S. & Oki, T. (2015). Incorporation of groundwater pumping in a global land surface model with the representation of human impacts, *Water Resour. Res*., 51, 7896.

Pokhrel, Y. N., Hanasaki, N., Wada, Y., & Kim, H. (2016). Recent progresses in incorporating human land-water management into global land surface models toward their integration into Earth system models. *Wiley Interdiscip. Rev. Water,* 3, 548-574.

Portmann, F.T., Siebert, S. & Döll., P (2010). MIRCA2000 – Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modelling. *Global Biogeochem. Cycles*, 24, GB1011.

Ramankutty, N., Evan, A.T., Monfreda, C. & Foley, J.A. (2008). Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochem. Cycles*, 22, GB1003

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors.*Water Resources Research*,*46*, W05521.

Revilla-Romero, B., Beck, H.E., Burek, P., Salamon, P., De Roo, A., Thielen, J. (2015). Filling the gaps: Calibrating a rainfall-runoff model using satellite-derived surface water extent. *Remote Sensing of Env.*, 171, 118-131.

Rost, S., Gerten, D., Bondeau, A., Lucht, W., Rohwer, J. & Schaphoff, S. (2008). Agricultural green and blue water consumption and its influence on the global water system. *Water Resour. Res*. 44, W09405.

Scanlon, B. R., et al. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proc. Nat. Acad. Sci.,* 201704665.

Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J. & Lucht, W. (2013). Contribution of permafrost soils to the global carbon budget. *Env. Res. Lett*., 8, 014026.

Schewe, J., et al. (2014). Multimodel assessment of water scarcity under climate change. *Proc. Nat. Acad. Sci. U.S.A.*, 111, 3245-3250.

Shiklomanov, I. A. (1997). Assessment of water resources and water availability in the world, Comprehensive assessment of the freshwater resources of the world. Stockholm, Sweden.

Sood, A. & Smakhtin, V. (2015) Global hydrological models: a review. *Hydrol. Sci. J*.,60, 549-565.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S. & Hsu, K.-L. (2017). A review of global precipitation data sets: data sources, estimation, and intercomparisons, *Reviews of Geophysics*, 56.

Takata, K., Emori, S. & Watanabe, T. (2003). Development of minimal advanced treatments of surface interaction and runoff. *Global Planet. Change*, 38, 209–222.

Thiemig, V., Bisselink, B., Pappenberg, F., Thielen, J. (2015). A pan-African medium-range ensemble flood forecast system. *Hydr. Earth. Syst. Sci*., 19, 3365-3385.

Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., De Roo, A. (2013). Hydrological evaluation of satellite-based rainfall estimates over the Volta and Baro-Akobo basin. *J. of Hydrol*., 499, 324-338.

Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P, Folton, N., Furusho, C., Kuentz,A., Lerat, J., Lindström, G., Martin, E., Mathevet, T., Merz, R., Parajka, J., Ruelland, D., Vaze, J. (2015). Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. *Hydrol. Sci. J*., 60, 1184-1199.

UNEP (2007). Global Environment Outlook 4: environment for development.

Veldkamp, T. I. E., Eisner, S., Wada, Y., Aerts, J. C. J. H., & Ward, P. J. (2015). Sensitivity of water scarcity events to ENSO-driven climate variability at the global scale. *Hydrol. Earth Syst. Sci*., *19*, 4081-4098.

Veldkamp, T. I. E., Eisner, S., Wada, Y., Aerts, J. C. J. H., & Ward, P. J. (2015). Sensitivity of water scarcity events to ENSO-driven climate variability at the global scale. *Hydrol. Earth Syst. Sci*., *19*, 4081-4098.

Veldkamp, T. I. E., Wada, Y., Aerts, J. C. J. H., Döll, P., Gosling, S. N., Liu, J., Masaki, Y., Oki, T., Ostberg, S., Pokhrel, Y., Satoh, Y., Kim, H., & Ward, P. J. (2017). Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century. *Nat. Comm.*, *8*, 15697.

Veldkamp, T. I. E., Wada, Y., Aerts, J. C. J. H., & Ward, P. J. (2016). Towards a global water scarcity risk assessment framework: incorporation of probability distributions and hydro-climatic variability. *Environ. Res. Lett.*, *11*, 024006.

Veldkamp, T. I. E., Wada, Y., de Moel, H., Kummu, M., Eisner, S., Aerts, J. C. J. H., & Ward, P. J. (2015b). Changing mechanisms of global water scarcity events: Impacts of socioeconomic changes and inter-annual hydro-climatic variability. *Glob. Env. Change*, *32*, 18-29.

Verzano, K., Barlund, I., Flörke, M., Lehner, B., Kynast, E. & Voß, F. (2012). Modeling variable river flow velocity on continental scale: Current situation and climate change impacts in Europe, *J. Hydrol.*, 424–425, 238–251.

Vörösmarty, C. J., Leveque, C., & Revenga, C. (2005). Freshwater ecosystems. In *Millennium Ecosystem Assessment Volume 1: Conditions and Trends* (pp. 165–207).

Wada, Y., Van Beek, L. P. H. & Bierkens, M. F. P. (2011). Modelling global water stress of the recent past: on the relative importance of trends in water demand and climate variability. *Hydrol. Earth Syst. Sci.*, 15, 3785–3808.

Wada, Y., Van Beek, L. P. H., Wanders, N. & Bierkens, M. F. P. (2013a). Human water consumption intensifies hydrological drought worldwide. *Environ. Res. Lett.*, 8, 034036.

Wada, Y., et al. (2013b). Multimodel projections and uncertainties of irrigation water demand under climate change. *Geophys. Res. Lett.*, 40, 4626-4632.

Wada, Y., Gleeson, T. & Esnault, L. (2014a). Wedge approach to water stress. *Nat. Geosc.*, 7, 615-617.

Wada, Y., Wisser, D. & Bierkens, M. F. P. (2014b). Global modelling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth Syst. Dyn.*, 5, 15–40.

Wada, Y., *et al.* (2016a). Modeling global water use for the 21st century: the Water Futures and Solutions (WFaS) initiative and its approaches. *Geosci. Model Dev.*, 9, 175-222.

Wada, Y., De Graaf, I. E. M. & Van Beek, L. P. H. (2016b) High-resolution modeling of human and climate impacts on global water resources.J. *Adv. Model. Earth Syst.*, 8, 735-763

Wada, Y., *et al.* (2017). Human–water interface in hydrological modelling: current status and future directions, *Hydrol. Earth Syst. Sci.*, 21, 4169-4193

Wanders, N., Wada, Y. & Van Lanen, H. A. J. (2015). Global hydrological droughts in the 21st century under a changing hydrological regime. *Earth Syst. Dynam*., 6, 1-15.

Wilks, D.S. (2006). On 'Field significance' and the false discovery rate. *J. Appl. Meteorol. Climatol.*, 45, 1181-1189.

Winsemius, H.C., *et al.* (2016). Global drivers of future flood risk. *Nat. Climate Change*, 6, 381-385.

Wöhling, T., Samaniego, L., Kumar, R. (2013). Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment. *Env. Earth. Sci.*, 59, 453-468.

WRI, UNEP, UNDP, & World Bank (1998). World Resources 1998-99: A guide to the global environment -environmental change and human health. Washington D.C., U.S.A

Yoshikawa, S., Cho, J., Yamada, H.G., Hansaki, N. & Kanae, S. (2014). An assessment of global net irrigation water requirements from various water supply sources to sustain irrigation: rivers and reservoirs (1960-2050). *Hydrol. Earth Syst. Sci*., 18, 4289-4310.

974 Zhang, L., Dobslaw, H., Stacke, T., Güntner, A., Dill, R., & Thomas, M. (2017). Validation of
975 Terrestrial Water Storage Variations as Simulated by Different Global Numerical Models with
976 GRACE Satellite Observations. *Hydrol. Earth Syst. Sci*., 821–837.
977
978 Zhao, F., *et al.* (2017). The critical role of the routing scheme in simulating peak river discharge in
979 global hydrological models. *Environ. Res. Lett.*, *12*, 075003.
980
981