Climatic Change (2017) 141:381–397
DOI 10.1007/s10584-016-1841-8

Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide

Shaochun Huang[1] · Rohini Kumar[2] · Martina Flörke[3] · Tao Yang[4] · Yeshewatesfa Hundecha[5] · Philipp Kraft[6] · Chao Gao[7] · Alexander Gelfan[8] · Stefan Liersch[9] · Anastasia Lobanova[9] · Michael Strauch[2] · Floris van Ogtrop[10] · Julia Reinhardt[9] · Uwe Haberlandt[11] · Valentina Krysanova[9]

Received: 13 November 2015 / Accepted: 17 October 2016 / Published online: 31 October 2016

Abstract In regional climate impact studies, good performance of regional models under present/historical climate conditions is a prerequisite for reliable future projections. This study aims to investigate the overall performance of 9 hydrological models for 12 large-scale river basins worldwide driven by the reanalysis climate data from the Water and Global Change

This article is part of a Special Issue on "Hydrological Model Intercomparison for Climate Impact Assessment" edited by Valentina Krysanova and Fred Hattermann.

Electronic supplementary material The online version of this article (doi:10.1007/s10584-016-1841-8) contains supplementary material, which is available to authorized users.

✉ Shaochun Huang
shh@nve.no

[1] Norwegian Water Resources and Energy Directorate (NVE), Oslo, Norway

[2] UFZ-Helmholtz Centre for Environmental Research, Leipzig, Germany

[3] Center for Environmental Systems Research, University of Kassel, Kassel, Germany

[4] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Center for Global Change and Water Cycle, Hohai University, Nanjing, China

[5] Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

[6] Justus-Liebig-University Gießen, Gießen, Germany

[7] College of Territorial Resources and Tourism, Anhui Normal University, Wuhu, China

[8] Water Problems Institute, Russian Academy of Sciences, Moscow, Russian Federation

[9] Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany

[10] Centre for Carbon, Water and Food (FAE), The University of Sydney, Sydney, NSW, Australia

[11] Institute of Water Resources Management, Leibniz University of Hannover, Hannover, Germany

(WATCH) project. The results serve as the basis of the application of regional hydrological models for climate impact assessment within the second phase of the Inter-Sectoral Impact Model Intercomparison project (ISI-MIP2). The simulated discharges by each individual hydrological model, as well as the ensemble mean and median series were compared against the observed discharges for the period 1971–2001. In addition to a visual comparison, 12 statistical criteria were selected to assess the fidelity of model simulations for monthly hydrograph, seasonal dynamics, flow duration curves, extreme floods and low flows. The results show that most regional hydrological models reproduce monthly discharge and seasonal dynamics successfully in all basins except the Darling in Australia. The moderate flow and high flows (0.02–0.1 flow exceedance probabilities) are also captured satisfactory in many cases according to the performance ratings defined in this study. In contrast, the simulation of low flow is problematic for most basins. Overall, the ensemble discharge statistics exhibited good agreement with the observed ones except for extremes in particular basins that need further scrutiny to improve representation of hydrological processes. The performances of both the conceptual and process-based models are comparable in all basins.

1 Introduction

In large-scale (continental or global) climate impact studies, it is becoming more common to use an ensemble of global hydrological models for the water sector to investigate changes in runoff characteristics such as annual runoff, magnitude and frequency of floods and droughts, under possible future scenarios (Dankers et al. 2014; Davie et al. 2013; Prudhomme et al. 2014). However, the large-scale hydrological models may not provide an accurate description of the climatological and hydrological system at a given location (Dankers et al. 2014). In contrast, regional hydrological models are often used in climate impact studies for individual river basins, as they require more detailed input data, have higher spatial resolution to represent the modelled processes, and are tuned specifically to represent the observed hydrological processes and discharge dynamics. However, the use of an ensemble of regional hydrological models for multiple regions is less frequent, mainly due to the large effort needed to setup and calibrate the models.

Although regional hydrological models with different levels of complexity may show similar performance for both the total runoff and extremes (i.e., floods and droughts) after calibration (Vansteenkiste et al. 2014), recent studies show that they may produce substantially different climate change impacts (Ludwig et al. 2009; Poulin et al. 2011). For example, Velazquez et al. (2013) applied an ensemble of hydrological models ranging from lumped and conceptual to fully distributed and physically based models. Their results show that the climate change response as identified by hydrological indicators, especially for the low flow, may differ substantially depending on the chosen hydrological model. Vetter et al. (2015) applied three hydrological models in three large-scale catchments and found that the uncertainty related to hydrological model structure can be comparable with the uncertainty related to driving climate models for some specific basins. Hence, the use of an ensemble of hydrological models is of importance to improve reliability of the regional-scale climate impact assessment.

Validation of hydrological models is commonly used to analyze performance of simulation and/or forecasting models (Biondi et al. 2012). It is a prerequisite to evaluate hydrological model performance prior to conducting a climate impact assessment. Such evaluations have been done for global hydrological models to assess performance of predictions of seasonal runoff (Gudmundsson et al. 2012b), runoff percentiles (Gudmundsson et al. 2012a), high and low flows (Prudhomme et al. 2011) and droughts (Van Loon et al. 2012). In general, the global models were found to broadly represent the inter-annual variability of runoff and drought propagation. Large uncertainties and errors were reported for extremes and in smaller catchments. However, in general, all the studies show that the ensemble mean (mean of all model simulations) performs better than the individual models.

Within the second phase of the Inter-Sectoral Impact Model Intercomparison project (ISI-MIP2), 9 regional hydrological models are applied for intercomparison of climate impacts on river discharge and hydrological extremes in 12 large-scale river basins worldwide. The models were built by different groups but all of them were calibrated and validated driven by the reanalysis climate forcing data from the Water and Global Change (WATCH) project (Weedon et al. 2011). The introductory paper of this special issue by Krysanova and Hattermann gives a detailed description on the applied models, basin characteristics, data used, climate scenarios as well as the modelling approach of the project. This information is essential to understand this paper and the whole special issue. Hence, we recommend readers to read this introductory paper first.

A systematic evaluation of the model performances within the ISI-MIP2 project is of particular importance as it provides the basis for the climate impact studies using the same models (see subsequent papers in this special issue). In addition, thanks to the ISI-MIP2 project, we are able to evaluate the model performance for 12 large-scale river basins while past model comparison studies have focused on 1 or 2 basins only (Jiang et al. 2007; Velazquez et al. 2013 and Vansteenkiste et al. 2014). More importantly, we implement a systematic evaluation based on the ability of hydrological models to reproduce monthly discharge, seasonal dynamics, flow duration curves and extremes, while many model comparison studies only considered the overall performance (Gao et al. 2015; Cornelissen et al. 2013; Poulin et al. 2011) or certain processes such as low flow and flow recession (e.g. Staudinger et al. 2011).

The main objectives of this study are: 1) to systematically evaluate the performance of regional hydrological models for each of 12 large-scale river basins in the framework of ISI-MIP2; 2) to analyze the ensemble mean/medians of model outputs in order to test whether they also outperform the individual model results as in global studies and 3) following from 1, to identify which aspects of flow are adequately and poorly simulated in general. Overall, this study provides valuable information for the subsequent climate impact assessment studies in this special issue and suggests potential modelling improvements in the next phase of the ISI-MIP2 project.

2 Method

In this study, 9 regional hydrological models were used: ECOMAG, HBV, HYMOD, HYPE, mHM, SWAT, SWIM, VIC and the regional version of the global model WaterGAP (WaterGAP3) with evaluation being performed in 12 large scale river basins (Table 1). Each model was applied to a different number of basins, and some models were applied by different modelling groups (see Table 4 of the introductory paper in this special issue). All models were calibrated and validated using the WATCH reanalysis climate forcing data. The calibration and validation periods of 8–10 years were selected within the timeframe 1951–2000, but they were different among the basins due to different availability of observed discharge data. The introductory paper of this special issue provides details on the modelled hydrological processes and the calibration and validation procedure.

Table 1 the availability of the observed discharge data at the twelve selected gauges. (Evaluation period = (End year – Start year +1) * (1 – Missing data/100))

Station	River	Drainage area (km^2)	Start year	End year	Missing data (%)	Time step	Evaluation period (year)
Alton	Mississippi	444,185	1971	2001	0.8	Daily	30.8
Almourol	Tagus	67,490	1973	2001	3.7	Daily	27.9
Arctic red river	Mackenzie	1,660,000	1972	2001	3.4	Daily	29.0
Cuntan	Yangtze	866,559	1971	2001	0	Daily	31.0
El Diem	Blue Nile	176,000	1971	1980	0.4	10-day	10.0
Farakka	Ganges	835,000	1965	1973	0	Monthly	9.0
Koulikoro	Niger	120,000	1971	2001	0.06	Daily	31.0
Lobith	Rhine	160,800	1971	2001	0	Daily	31.0
Louth	Darling	489,300	1971	2001	38	Daily	19.2
SP-Olivenca	Amazon	990,781	1973	2001	3.9	Daily	27.9
Stolb	Lena	2,460,000	1971	2001	0	Daily	31.0
Tangnaihai	Yellow	121,000	1971	2001	0	Daily	31.0

In this study, we focused on evaluating the monthly hydrographs and seasonal dynamics for all basins due to the large scale of the studied basins, complex hydrological and anthropogenic conditions in some of the basins, and poor availability of observational data (mostly from global datasets). For some basins with long-term daily observed discharge data, we also analyzed model skill for representing extremes and flow duration curves at the daily time step.

A large number of statistical criteria available for model evaluation can be found in literature, for example, nearly sixty criteria were used by Crochemore et al. (2015). These criteria can be mainly distinguished regarding the applied statistical methods (e.g. residual methods or correlation measures) and regarding their targets (e.g. in evaluating the entire simulation period or single events). Here we used twelve numeric criteria based on different evaluation targets (monthly hydrograph, long-term average seasonal dynamics, flow duration curves and extremes), which were selected based on intensive literature review on model evaluation (Table 2).

To evaluate simulation of the monthly hydrograph, Moriasi et al. (2007) recommend the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe 1970) and percent bias (PBIAS). However, the NSE is strongly influenced by the high flow performance, and Pushpalatha et al. (2012) suggested also using the Nash-Sutcliffe efficiency criterion calculated on inverse flow values (NSEiq) for the low flow evaluation. Recently, the Kling-Gupta efficiency (KGE) was developed to provide diagnostic insights into the model performance by decomposing the NSE into three components: correlation, bias and variability (Gupta et al. 2009). Later the KGE was modified by Kling et al. (2012) to ensure that the bias and variability ratios are not cross-correlated. The KGE is easy to interpret as it gives the lower limit of the three components. In addition, a more general and simple criterion, volumetric efficiency (VE), was proposed by Criss and Winston (2008). This criterion does not overemphasize the high flows as the NSE and allied algorithms, and represents the fractional mismatch of water volume at the proper time. In this paper, we used the NSE, NSEiq, PBIAS, the modified KGE and VE to evaluate the monthly hydrographs.

Table 2 Model performance criteria and their corresponding formulation. (Q_s and Q_o are the simulated and observed discharges, respectively; N is the total number of time steps; and $\overline{Q}$ is the mean value of discharge. The meaning of other parameters is explained as footnotes.)

Criterion	Usual name	Reference	Formula	Range	Ideal value	Time resolution	Objective
NSE	Nash-Sutcliffe Efficiency calculated on flows	(Nash and Sutcliffe 1970)	$1-\frac{\sum_{t=1}^{N}\left(Q_{s,t}-Q_{o,t}\right)^2}{\sum_{t=1}^{N}\left(Q_{o,t}-Q_o\right)}\,2$	(−∞,1)	1	Monthly	Hydrograph
NSEiq	NSE calculated on inverse transformed flows	(Pushpalatha et al. 2012)	$1-\frac{\sum_{t=1}^{N}\left(\frac{1}{Q_{s,t}+\varepsilon}-\frac{1}{Q_{o,t}+\varepsilon}\right)^2}{\sum_{t=1}^{N}\left(\frac{1}{Q_{o,t}+\varepsilon}-\frac{1}{Q_{o,t}+\varepsilon}\right)^2}$	(−∞,1)	1	Monthly	Hydrograph with emphasis on low flows
KGE	Modified Kling-Gupta efficiency	(Kling et al.,2012)	$1-\sqrt{(r-1)^2+(\beta-1)^2+(\gamma-1)^2}$	(−∞,1)	1	Monthly	Hydrograph
VE	Volumetric efficiency	(Criss and Winston 2008)	$1-\frac{\sum_{t=1}^{N}\left\lvert Q_{s,t}-Q_{o,t}\right\rvert}{\sum_{t=1}^{N}Q_{o,t}}$	(0,1)	1	Monthly	Hydrograph
PBIAS	Percent bias in overall flows	(Moriasi et al. 2007)	$\frac{\sum_{t=1}^{N}\left(Q_{s,t}-Q_{o,t}\right)}{\sum_{t=1}^{N}Q_{o,t}}*100$	(−∞,+∞)	0	Monthly	Total water yield
Δσ	Percent bias in standard deviation	(Gudmundsson et al. 2012b)	$\frac{\sigma_s-\sigma_o}{\sigma_o}*100$	(−∞,+∞)	0	Long-term average monthly	Seasonal dynamics
PCC	Pearson's correlation coefficient	(Gudmundsson et al. 2012b)	$\frac{\sum_{t=1}^{12}\left(Q_{s,t,}-\overline{Q}_s\right)\left(Q_{o,t}-\overline{Q}_o\right)}{\sqrt{\sum_{t=1}^{12}\left(Q_{s,t,}-\overline{Q}_s\right)^2}\sqrt{\sum_{t=1}^{12}\left(Q_{o,t}-\overline{Q}_o\right)^2}}$	(−1,1)	1	Long-term average monthly	Seasonal dynamics
ΔFMS	Percent bias in flow duration curve (FDC) midsegment slope	(Yilmaz et al. 2008)	$\frac{\left[\log\left(Q_{s,m1}\right)-\log\left(Q_{s,m2}\right)\right]-\left[\log\left(Q_{o,m1}\right)-\log\left(Q_{o,m2}\right)\right]}{\log\left(Q_{o,m1}\right)-\log\left(Q_{o,m2}\right)}*100$	(−∞,+∞)	0	Daily	Moderate flow (FDC)
ΔFHV	Percent bias in FDC high-segment volume	(Yilmaz et al. 2008)	$\frac{\sum_{h=1}^{H}\left(Q_{s,h}-Q_{o,h}\right)}{\sum_{h=1}^{H}\left(Q_{o,h}\right)}*100$	(−∞,+∞)	0	Daily	High flow (FDC)

Table 2 (continued)

Criterion	Usual name	Reference	Formula	Range	Ideal value	Time resolution	Objective
ΔFLV	Percent bias in FDC low-segment volume	(Yilmaz et al. 2008)	$-1*\frac{\sum_{l=1}^{L}\left[\log\left(Q_{s,l}\right)-\log\left(Q_{s,L}\right)\right]-\sum_{l=1}^{L}\left[\log\left(Q_{s,l}\right)-\log\left(Q_{s,L}\right)\right]}{\sum_{l=1}^{L}\left[\log\left(Q_{o,l}\right)-\log\left(Q_{o,L}\right)\right]}*100$	(−∞,+∞)	0	Daily	Low flow (FDC)
ΔFlood	Percent bias in the 10 and 30-year flood levels		$\frac{(FQ10_s+FQ30_s)-(FQ10_o+FQ30_o)}{(FQ10_o+FQ30_o)}*100$	(−∞,+∞)	0	Daily	Extreme flood
Δ Lowf	Percent bias in the 10 and 30-year low flow levels		$\frac{(LQ10_s+LQ30_s)-(LQ10_o+LQ30_o)}{(LQ10_o+LQ30_o)}*100$	(−∞,+∞)	0	Daily	Extreme low flow

ε: a small constant

r: correlation coefficient between simulated and observed runoff

β: bias ration

γ: the variablity ration

σ: standard deviation of the mean annual cycle

m1 and m2: the 0.2 and 0.7 flow exceedance probabilites.

h: the flow indices for flow exceedance probabilities lower than 0.05.

H: the index of flow exceedance probability of 0.05

l: the index of flow value within the low-flow segment (0.7–1.0 flow exceedance probabilites) of the flow duration curve.

L: the index of the minimum flow

Regarding the seasonal dynamics, the relative difference in standard deviation ($\Delta\sigma$) and the Pearson's correlation coefficient (*PCC*) between the observed and simulated annual mean cycles were used to evaluate the performance of global hydrological models (Gudmundsson et al. 2012b). We also applied these two criteria to evaluate the model skill to reproduce the observed long-term seasonal runoff dynamics.

Yilmaz et al. (2008) developed multiple hydrologically relevant signature measures for comparing flow duration curves (FDC). Three measures included in this study are: a) percent bias in FDC mid-segment slope (ΔFMS, 0.2–0.7 flow exceedance probabilities); b) percent bias in FDC high-segment volume (ΔFHV, 0–0.05 flow exceedance probabilities) (we also analyze the 0–0.02 interval used by Yilmaz et al. (2008) and the 0–0.1 interval to provide additional information on high flows); and c) percent bias in FDC low-segment volume (ΔFLV, 0.7–1.0 flow exceedance probabilities), related to base flow.

For hydrologic extremes, we calculated percent bias for 10 and 30-year flood return intervals (ΔFlood) and the similar methodology was extended for low flow levels (ΔLowf). The 10- and 30-year flood levels were estimated by fitting the Generalized Pareto Distribution (GPD) (Coles 2001) to the peaks over threshold (POT) time series. The approach of the POT threshold was selected to ensure that on average two independent flood events per year were included in the estimation approach (Huang et al. 2014). The 10 and 30 years low flow levels were estimated by fitting the Generalized Extreme Value (GEV) distribution (Coles 2001) to the annual minimum 7-day (AM7) mean flows using the method of L-moments (Huang et al. 2013). Since the 12 basins have different hydrological regimes, we could not use a unique definition of hydrological years to select the AM7 mean flows. Instead, we defined a year starting from a month with the highest monthly flow for each basin (Vetter et al. 2015).

Based on the performance ratings suggested by Dawson et al. (2007), Moriasi et al. (2007), Ritter and Munoz-Carpena (2013) and Crochemore et al. (2015), we chose the rating NSE $\geq$ 0.7, |PBIAS| $\leq$ 15 % and KEG $\geq$0.7 to denote a "good" performance. Considering their similarity to the NSE, the 0.7 threshold was also applied to NSEiq. There is little information on performance ratings for other criteria in literature, partly because they are not used as often as the previous ones and partly because they are data dependent (Dawson et al. 2007). Hence, performance ratings for other criteria were adjusted based on previous applications (Gudmundsson et al. 2012b; Kay et al. 2015) or as analogous to other similar criteria to make the comparison among the basins more straightforward. They are: VE $\geq$ 0.7, $|\Delta\sigma| \leq 15$ %, *PCC* $\geq$ 0.9, and all biases for the FDC segments and extremes should be between −25 % and 25 %.

In addition, the use of the numerical criteria was complemented by visual comparison between the observed and simulated monthly hydrographs, duration curves, seasonal dynamics and hydrologic extremes. The pros and cons of the selected criteria and the performance ratings will be discussed.

3 Study area and data

In the ISI-MIP2 project, 12 large-scale basins located on 6 continents were selected for the regional-scale application of hydrological models. For each basin, 1 or 2 gauge stations were chosen for the model calibration and validation. Here we evaluate the model outputs for only 1 gauge per basin, which has the largest number of model applications. In the following we provide a brief overview about input and forcing data; interested readers may refer to Krysanova and Hattermann (this special issue) for more detailed information on the studied basins and input data.

The morphological input data, such as the Digital Soil Map of the World (FAO) and the Global Land Cover data (GLCF) were recommended to each modelling group. Where available, some groups utilized more accurate locally available data to parameterize their models. Human activities, such as dams/reservoirs, water abstraction for irrigation, should be considered in the basins where their effects are significant, but it was not mandatory. Importantly, all the models were strictly calibrated and validated using the WATCH forcing data available at a grid resolution of 0.5 degrees. The NSE and PBIAS were suggested as the main objective functions within the ISI-MIP2 framework, and, additionally, the high and low flow percentiles had to be compared. The simulations for the entire period were uploaded to a centralized database; and this information was used for the subsequent model evaluation.

Table 1 lists the availability of observed discharge data for each river basin selected for the model evaluation. The observed river discharge data were mainly obtained from the Global Runoff Data Center (GRDC). For some rivers, such as the Yellow and Yangtze, the discharge data were provided by the local authorities. For the gauges Farakka (Ganges) and El Diem (Blue Nile), less than 10 years of the monthly or 10-day discharge data was available within the evaluation time period. Hence, for these cases we only calculated the criteria for monthly discharge and seasonal dynamics. For the Ganges, we specifically asked the modelers to provide the simulation starting in 1965, so that we could evaluate the monthly hydrographs for 9 years. The daily discharge time series at the gauges Louth (Darling) was shorter than 20 years. Therefore, we excluded results of this location from the extreme value analyses due to the small sample size. The other 9 gauges have the relatively long-term daily discharge series (> 26 years) which enabled the model evaluation based on all twelve criteria. In summary, we evaluated the monthly discharge and seasonal dynamics based on the monthly discharge data for all gauges and calculated the criteria related to duration curves and extremes using daily discharge series for 9 gauges.

In addition to the individual model results, we calculated the corresponding ensemble mean and the median outputs for every basin and applied the same evaluation criteria on them.

4 Results

The first glance on the model performance is provided by Fig. 1, which shows the average monthly discharge across 12 river basins. In many cases, the individual models reproduce the seasonal dynamics of streamflow reasonably well. For some basins, such as the Rhine, Yangtze and Yellow, the differences among the model results are not substantial, while for other basins, outliers are distinguishable from the other model results. For example, the VIC model outputs for the Tagus and the WaterGAP3 outputs for the Lena, Niger and Blue Nile, clearly show distinguishable hydrologic simulations than those of the other models. Despite these differences, the ensemble mean and median results show good agreement with the observed values for all basins except the Amazon and Darling.

Figure 2 shows the GPD plots for the flood peaks (upper three rows) and the GEV plots for the low flows (lower three rows). Both the individual model results and the ensemble mean and median are compared to the observed values. All of the applied models underestimated the flood peaks in the Tagus, Yangtze and Lena basins, and most applied models overestimate the floods for the Niger. There is a large spread of flood estimations for the Rhine, indicating the

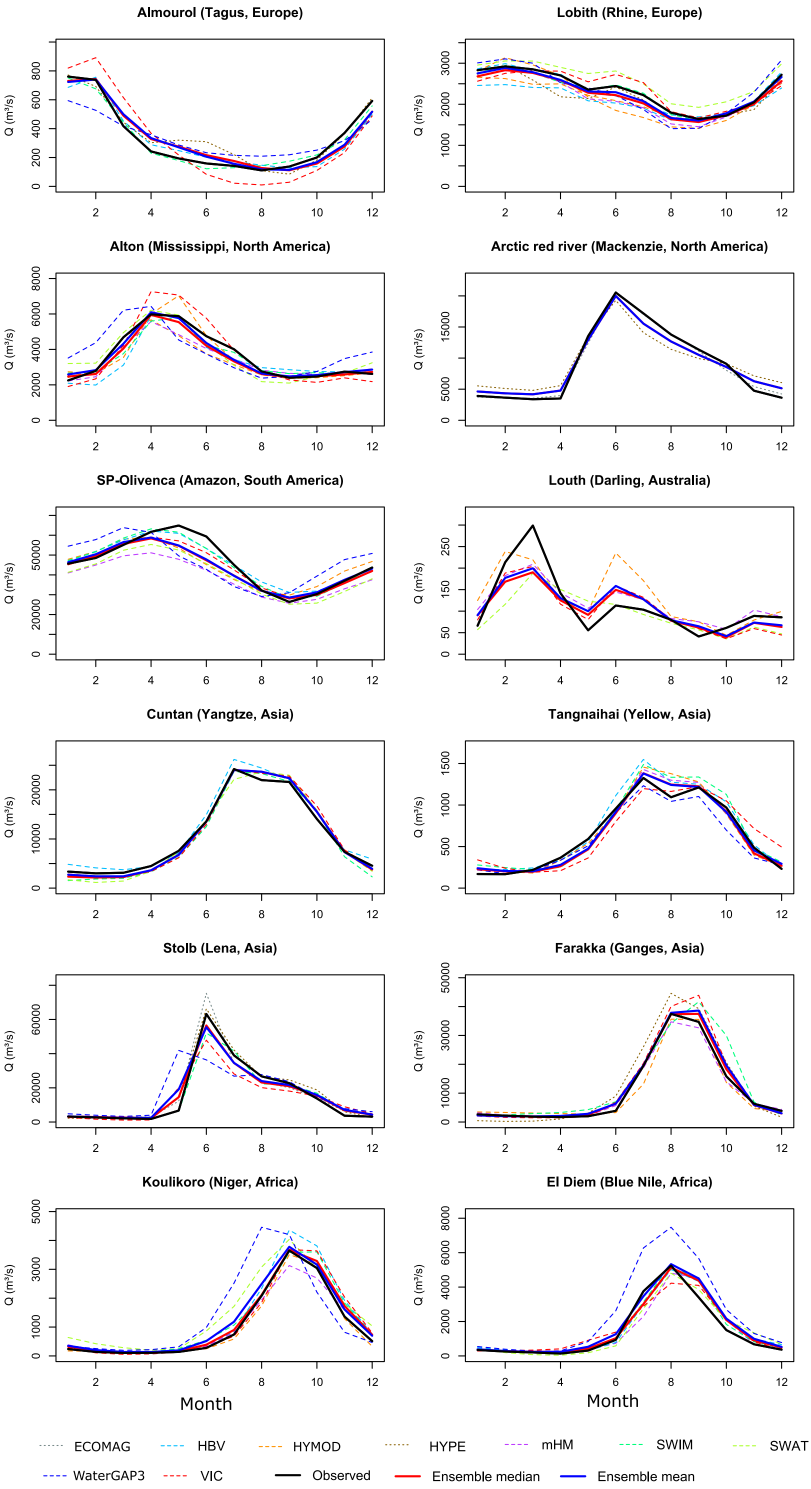

Fig. 1 Comparison of observed and simulated average monthly discharges at twelve gauges

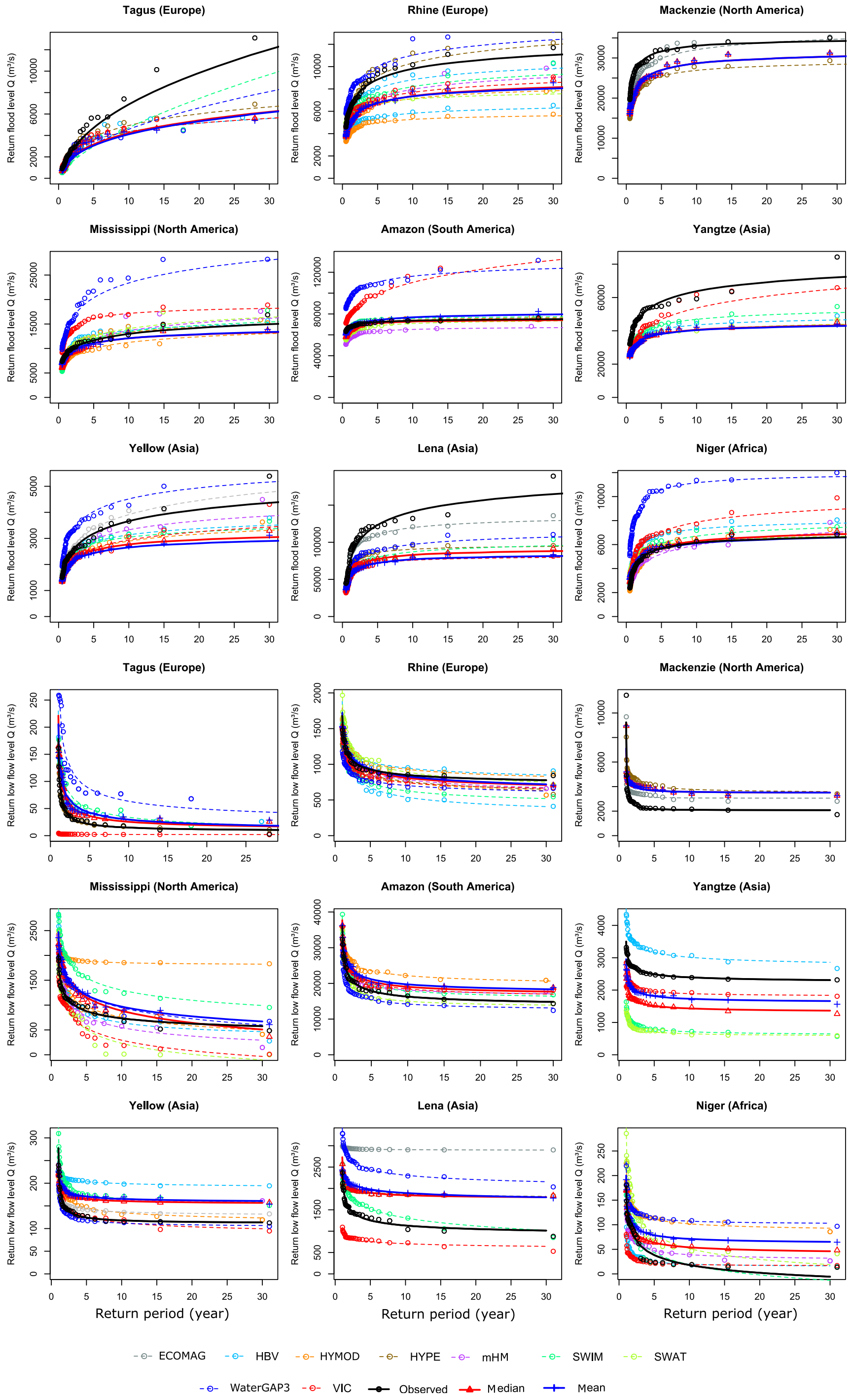

Fig. 2 The GPD plots for observed and simulated POT series (upper three rows) and the GEV plots for observed and simulated AM7 mean flows (the lower three rows) by the individual models and the ensemble mean and median at 9 gauges. All extremes were extracted from daily time series

large uncertainty of simulating flood peaks for this river. The flood peaks in the ensemble mean and median series are generally lower than the peaks from most individual models as different models projected slightly different flood timing. Hence, they did not provide a better estimate than most of the individual models for all studied basins except for the Amazon and Niger.

The GEV curves for low flows from most individual models differ significantly for the Yangtze, Lena, Mississippi and Niger basins, and they are better comparable for the Rhine basin. The ensemble mean and median did not improve the results either because most individual models either overestimated or underestimated the low flows.

Figures A and B (supplementary material) show parts of the FDC for high and low flows separately. Figure A shows that the largest bias in high flows is found within the 0–0.02 flow exceedance probabilities for almost all basins, while the lowest bias being observed for the 0.05–0.1 flow exceedance probability. These results suggest that all models have difficulties to simulate the extremes for most basins but a more robust assessment of projected future high flows can be made analyzing the flows at the 95th or 90th percentiles. For low flows, Figure B clearly shows large differences between the model results and the observations particularly for the Tagus, Niger, Lena, Yangtze and Darling basins.

Figure 3 summarizes all of the evaluated criteria scores for the individual models and the ensemble mean (the ensemble mean and median show very similar results in Figs. 1-2). The "good" performance values are highlighted by the rose coloured background. In addition, all criteria values of all models and basins are provided in Table A (supplementary material).

Most models perform well according to the NSE, KGE and VE criteria. All models in all basins generate low PBIAS, except some models for the two African basins. The model WaterGAP3 did not perform as well as other models for five basins. For the gauge Louth (Darling) more than a half of the applied models could not reproduce the observed discharge satisfactorily based on the analyzed criteria.

Despite the generally satisfactory performance indicated by the NSE, KGE, VE and PBIAS; the NSEiq values are generally lower than the NSE and the KGE indicating that many models have difficulties in adequately reproducing the low flows in most of the analyzed basins. None of the applied models could provide satisfactory results in terms of NSEiq using the threshold 0.7 in the Tagus, Darling, Blue Nile and Niger basins. However, the ensemble mean generally shows a good performance in terms of NSE, KGE, NSEiq and VE, and in some cases it performed better than the best single model result (e.g., in the upper Mississippi river basin).

Most regional hydrological models reproduced the shape and the timing of the mean annual cycle quite well for all the study basins, as indicated by the high *PCC*, with an exception to the performance noted for the Darling basin (Fig. 3). Most $\Delta\sigma$ values (Fig. 3) are within the "good" range for all basins except the Darling and Amazon basins. There are also some large differences in the standard deviation of the mean annual cycle for other basins. For example, the $\Delta\sigma$ ranges from −29 % to 36 % for the Rhine and from −47 % to 32 % for the Tagus. However, the ensemble mean and median always show a good agreement with the observed one for all the rivers except the Amazon and Darling, where deviations are higher.

The values of the bias for different FDC segments also show that the high flow and the mid-segment are better simulated than the low flow. For the high flow simulations, only one or two models show inadequate results for specific basins. However, Table A (supplementary material) shows that the FDC segments for the 0–0.02 exceedance probabilities have higher bias than FDC for the 0–0.05 exceedance probabilities in most cases. The simulation results for the

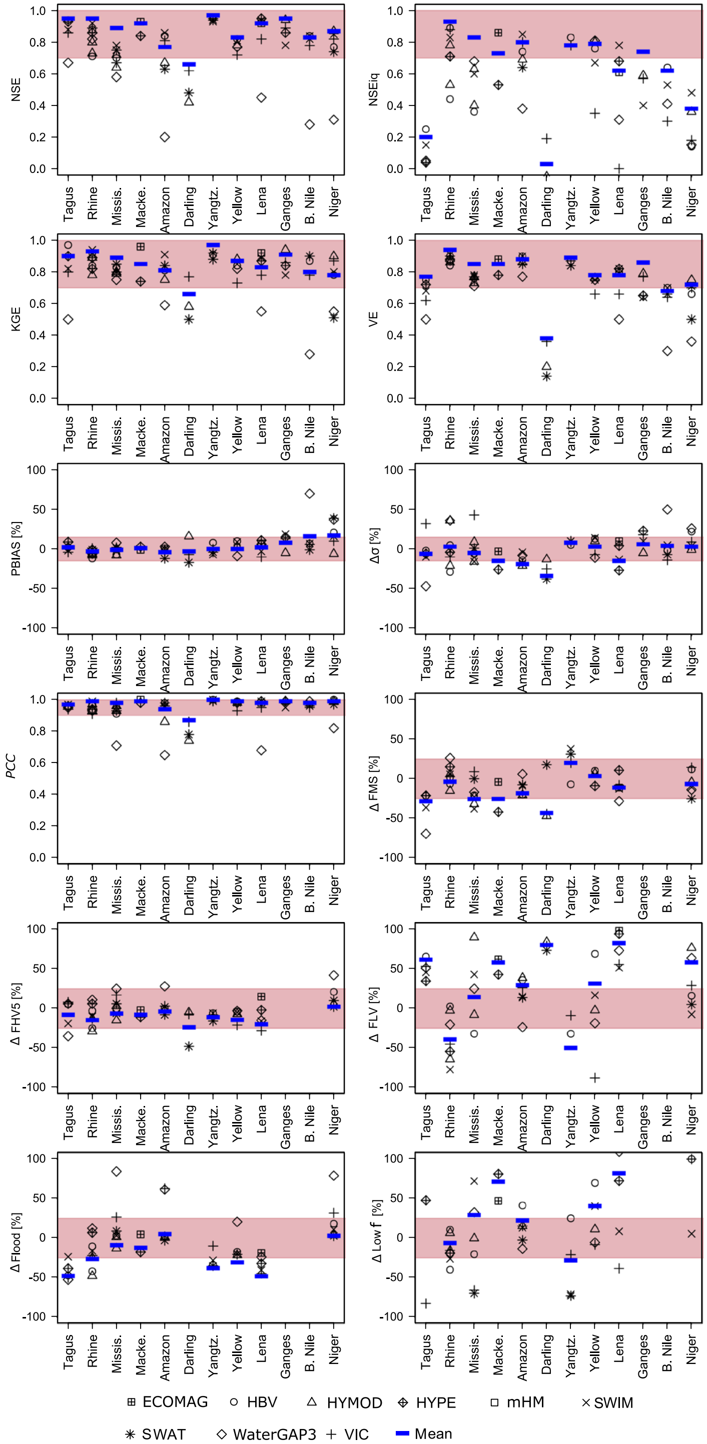

Fig. 3 The summary of the twelve criteria values for all 12 basins. The rose background indicates the "Good" performance range for each criterion

moderate flow by all models are good for the Yellow, Amazon, Rhine and Niger. For the Tagus, Darling and Yangtze, about half of the applied models could not provide sufficiently good simulation results. There are large biases in the low flow simulations for most of the basins, especially for the Yangtze basin. Two out of four applied models have a ΔFLV higher than 100 %, so their results are not visible in Fig. 3. Also the ensemble mean series, which generally provide good results indicated by other criteria, have a large bias for the low flow in all rivers except the Mississippi.

Compared to the high flows evaluated with ΔFHV, the extreme floods evaluated with ΔFlood are not simulated well by most models in several basins. The flood peaks are underestimated in the Tagus, Yangtze and Lena while they are overestimated in the Niger basin by most applied models (Figs. 2 and 3). The WaterGAP3 and VIC models simulate distinguishably higher flood levels than other models for the Mississippi, Amazon and Niger rivers.

The simulated extreme low flows have larger bias compared to floods. Good results could be achieved by most of the models for the Rhine and Amazon. For other rivers, especially for the Tagus and Niger, some models generate very large biases (out of the +/− 100 % range), and therefore are not visible in Fig. 3. The ensemble mean and median simulations tend to match around the performance of the individual models results.

Finally, since the 9 hydrological models were applied to different number of basins, we could not rank the performances of individual regional models consistently and quantitatively. The model performances could be only roughly compared in Fig. C (supplementary material). In general, both types of regional models, conceptual and process-based, have similar performances for the 12 large-scale basins, and HBV and SWIM perform slightly better than other models in terms of monthly hydrographs, seasonal dynamics and high flows for most basins. The regional version of the global hydrological model WaterGAP used in this study did not perform as well as other regional models. However, all models had difficulties to reproduce low flows for most basins. The ensemble mean and median show good results for almost all basins regarding monthly hydrographs, seasonal dynamics, moderate and high flows, but they do not outperform the individual models regarding low flows and floods.

5 Discussion

In this study, we selected 12 numeric criteria for the model evaluation from literature. Some are widely used in the field of hydrological modelling, such as NSE and PBIAS, others focus on certain aspects of the hydrograph, such as biases for different FDC segments, or specifically investigate extreme events.

Each single criterion has its own peculiarities and problems. For example, the advantages and weaknesses of KGE and NSE are discussed by Gupta et al. (2009). Pushpalatha et al. (2012) also mentioned that the NSE values calculated on the transformed flows (i.e. NSEiq in this study) emphasize different model errors depending on regime characteristics, flow variability and model performances. In addition, Schaefli and Gupta (2007) pointed out that high NSE for catchments with greater seasonality may not reflect higher model skills, and they suggested using the mean annual cycle as a benchmark in the NSE equation. The negative values of the benchmark NSE in Table A (supplementary material) show that some models do not perform better than the long-term annual cycles for the Blue Nile, Lena, Amazon and Niger. Hence, the use of benchmark NSE can be suggested for the future calibrations of these basins.

Another example is the ΔFLV, which measures the shape of the FDC low-segment but does not consider the bias between the observed and simulated low flow segments. This weakness leads, for example, to a larger ΔFLV of the ECOMAG model compared to that of the HYPE model for the Mackenzie river, but the ECOMAG simulation output is visually closer to the observations than that from HYPE (Fig. B, supplementary material).

We also noticed that the twelve criteria are not all independent. A correlation analysis was carried out between each pair of criteria using the Spearmans rank-correlation coefficient. High correlations (absolute values >0.7) were found between ΔFlood and ΔFHV, ΔLowf and ΔFLV, and NSE and KGE. The correlation between ΔFlood and ΔFHV is due to their focus on the high flows. The difference is that ΔFlood uses only the independent peaks, and ΔFHV considers all daily discharges within the high FDC segment. However, ΔFlood has larger biases than ΔFHV (Fig.3), indicating that it is generally more difficult to simulate the very extreme events than high flows. This difference also applies to the two low flow criteria. Hence, these criteria still provide some additional information; even though some of them are correlated.

The poorer model performance for the Amazon and Darling basins can be partly explained by poor climate data. For example, the WATCH data was found to be not reliable in terms of precipitation amounts in the Amazon basin due to undercatch of fog/mist in tropical montane cloud forests and improperly resolved precipitation gradients along the Andes mountain range (Strauch et al. 2016). In addition, the complex water management, semi-arid climate, and the flat landscape contribute to the difficulties of modelling the Darling basin.

The systematic underestimation of flow peaks in the Tagus, Yangtze and Lena basins could be a result of underestimating rainfall intensity before and during the flood events. The unsatisfactory results of the three low flow criteria (NESiq, ΔFLV and ΔLowf) can be partly due to a high sensitivity of low flow to water management, which was not considered in simulations. Moreover, the low flow observations may be inaccurate for specific rivers, for example the Lena River due to the river ice effect. However, in some cases, the relative large biases for the small value of low flow do not necessarily mean poor performance. For example, the absolute biases for the Tagus and Niger rivers are much lower than for other basins but the relative biases are worst among all basins. Hence, the visual comparison and use of absolute values should also be considered in the evaluation of low flow simulations.

Another reason for the inability of models to reproduce low flows well may be as a result of the choice of objective functions for calibrating the models. The NSE and PBIAS values are sensitive to high flow, and especially for the rivers with high coefficient of variation (CV) of discharges. For example, a large absolute bias of extreme low flows was found for the Yangtze and Lena rivers, where the CV of discharges is 0.8 and 1.3, respectively. In contrast, relatively good low flow results were found for the Rhine and Amazon rivers, which have the lowest CV (0.5 and 0.3, respectively). Therefore, it is recommended that additional criteria, such as NSEiq, specifically for the evaluation of low flow simulations, should be included in the calibration procedure in the future.

The WaterGAP3 model did not perform as well as other models for five basins, which is to some extent related to the fact that WaterGAP3 was calibrated using only two parameters. The other models used 5–7 parameters and in some cases up to 13–17 parameters for calibration. However, it is not possible to investigate each individual model performance for different climate/hydrological regimes in this study, mainly due to the unequal number of hydrological models applied for each basin. Furthermore, in some cases, the modelers applied their own local data instead of the recommended global data.

Finally, it was not possible to evaluate the transferability of model parameters under different climate conditions in this study, especially for the gauges which have short discharge records. Coron et al. (2012) showed that the parameters for the calibration period may introduce significant errors in other periods with contrast climate. In addition to the model performance for the historical period, the climate impact studies should take the potential influence of this shortcoming into account.

6 Conclusions

This paper presents results of the model performance evaluation of 9 hydrological models for 12 large-scale basins using 12 numerical criteria and visual comparison. These results are essential and provide the basis for the follow-up climate change impact studies using these hydrological models. The results show that most regional hydrological models can adequately reproduce the monthly discharges, seasonal dynamics, moderate flows and high flows (0.05–0.1 flow exceedance probabilities) for most of the basins. The Darling basin needs further attention of the modelers to improve the simulation results.

The flood peaks are underestimated in the Tagus, Yangtze and Lena, while they are overestimated in the Niger basin by most applied models. The simulated low flows are more problematic for most basins and models. Results from scenario studies should be wary of conclusions drawn from simulated extreme floods and even more so with simulated extreme low flows.

The ensemble median and mean of the simulated discharges have generally a good agreement with the observed values; hence they can be used to analyze the average and seasonal discharges under the climate scenario conditions. However, caution should be taken with the assessment of hydrological extremes. The ensemble mean and median did not always provide better performance than most individual models. Both the conceptual and process-based models provide similar simulation results in terms of all twelve criteria used in this study.

This study evaluated the model performance regarding river discharge only. Further evaluation including other components of the water balance (evaporation, soil moisture and groundwater recharge) could provide more insights into the model performance, particularly for the representation of different hydrologic processes. In addition, we could not inter-compare the individual model performances for all basins or certain hydrological regimes due to the unequal number of model applications for each basin. We could not directly compare the individual models in terms of model structures either, because the current results may be affected by several other factors, e.g. neglecting water management in many simulations, different calibration procedures and uncertainty/errors in input forcing data. We plan to fill these modelling gaps and improve the general model performance by re-calibration with multiple objective functions at multiple outlets and inclusion of water management information for basins where its role is significant in the next phase of the ISI-MIP2 project.

Acknowledgments The authors would like to thank all project partners who contributed to this study in the ISI-MIP2 project. The Chinese partner was funded by National Natural Science Foundation of China (41571018).

References

Biondi D, Freni G, Iacobellis V, Mascaro G, Montanari A (2012) Validation of hydrological models: conceptual basis, methodological approaches and a proposal for a code of practice. Phys Chem Earth 42-44:70–76

Coles S (2001) An introduction to statistical modeling of extreme values. Springer-Verlag, London, UK

Cornelissen T, Diekkruger B, Giertz S (2013) A comparison of hydrological models for assessing the impact of land use and climate change on discharge in a tropical catchment. J Hydrol 498:221–236

Coron L, Andreassian V, Perrin C, Lerat J, Vaze J, Bourqui M, Hendrickx F (2012) Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. Water Resour Res 48.

Criss RE, Winston WE (2008) Do Nash values have value? Discussion and alternate proposals. Hydrol Process 22:2723–2725

Crochemore L, Perrin C, Andreassian V, Ehret U, Seibert SP, Grimaldi S, Gupta H, Paturel J (2015) Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrol Sci J 60:402–423

Dankers R, Arnell NW, Clark DB, Falloon PD, Fekete BM, Gosling SN, Heinke J, Kim H, Masaki Y, Satoh Y, Stacke T, Wada Y, Wisser D (2014) First look at changes in flood hazard in the inter-sectoral impact model intercomparison project ensemble. P Natl Acad Sci USA 111:3257–3261

Davie JCS, Falloon PD, Kahana R, Dankers R, Betts R, Portmann FT, Wisser D, Clark DB, Ito A, Masaki Y, Nishina K, Fekete B, Tessler Z, Wada Y, Liu X, Tang Q, Hagemann S, Stacke T, Pavlick R, Schaphoff S, Gosling SN, Franssen W, Arnell N (2013) Comparing projections of future changes in runoff from hydrological and biome models in ISI-MIP. Earth Syst Dynam 4:359–374

Dawson CW, Abrahart RJ, See LM (2007) HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ Model Softw 22:1034–1052

Gao C, Yao MT, Wang YJ, Zhai JQ, Buda S, Fischer T, Zeng XF, Wang WP (2015) Hydrological model comparison and assessment: criteria from catchment scales and temporal resolution. Hydrol Sci J

Gudmundsson L, Tallaksen LM, Stahl K, Clark DB, Dumont E, Hagemann S, Bertrand N, Gerten D, Heinke J, Hanasaki N, Voss F, Koirala S (2012a) Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. J Hydrometeorol 13:604–620

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

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J Hydrol 377:80–91

Huang SC, Krysanova V, Hattermann FF (2013) Projection of low flow conditions in Germany under climate change by combining three RCMs and a regional hydrological model. Acta Geophys 61:151–193

Huang SC, Krysanova V, Hattermann FF (2014) Does bias correction increase reliability of flood projections under climate change? A case study of large rivers in Germany. Int J Climatol 34:3780–3800

Jiang T, Chen YQD, Xu CYY, Chen XH, Chen X, Singh VP (2007) Comparison of hydrological impacts of climate change simulated by six hydrological models in the Dongjiang Basin, South China. J Hydrol 336: 316–333

Kay AL, Rudd AC, Davies HN, Kendon EJ, Jones RG (2015) Use of very high resolution climate model data for hydrological modelling: baseline performance and future flood changes. Clim Chang 133:193–208

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

Ludwig R, May I, Turcotte R, Vescovi L, Braun M, Cyr JF, Fortin LG, Chaumont D, Biner S, Chartier I, Caya D, Mauser W (2009) The role of hydrological model complexity and uncertainty in climate change impact assessment. Adv Geosci 21:63–71

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. T Asabe 50:885–900

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I -a discussion of principles. J Hydrol 10:282–290

Poulin A, Brissette F, Leconte R, Arsenault R, Malo JS (2011) Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. J Hydrol 409:626–636

Prudhomme C, Parry S, Hannaford J, Clark DB, Hagemann S, Voss F (2011) How well do large-scale models reproduce regional hydrological extremes in Europe? J Hydrometeorol 12:1181–1204

Prudhomme C, Giuntoli I, Robinson EL, Clark DB, Arnell NW, Dankers R, Fekete BM, Franssen W, Gerten D, Gosling SN, Hagemann S, Hannah DM, Kim H, Masaki Y, Satoh Y, Stacke T, Wada Y, Wisser D (2014) Hydrological droughts in the twenty-first century, hotspots and uncertainties from a global multimodel ensemble experiment. P Natl Acad Sci USA 111:3262–3267

Pushpalatha R, Perrin C, Le Moine N, Andreassian V (2012) A review of efficiency criteria suitable for evaluating low-flow simulations. J Hydrol 420:171–182

Ritter A, Munoz-Carpena R (2013) Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J Hydrol 480:33–45

Schaefli B, Gupta HV (2007) Do Nash values have value? Hydrol Process 21:2075–2080

Staudinger M, Stahl K, Seibert J, Clark MP, Tallaksen LM (2011) Comparison of hydrological model structures based on recession and low flow simulations. Hydrol Earth Syst Sci 15:3447–3459

Strauch M, Kumar R, Eisner S, Mulligan M, Reinhardt J, Santini W, Vetter T, Friesen J (2016) Adjustment of global precipitation data for enhanced hydrologic modeling of tropical Andean watersheds. Clim Chang. doi:10.1007/s10584-016-1706-1

Van Loon AF, Van Huijgevoort MHJ, Van Lanen HAJ (2012) Evaluation of drought propagation in an ensemble mean of large-scale hydrological models. Hydrol Earth Syst Sci 16:4057–4078

Vansteenkiste T, Tavakoli M, Van Steenbergen N, De Smedt F, Batelaan O, Pereira F, Willems P (2014) Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation. J Hydrol 511:335–349

Velazquez JA, Schmid J, Ricard S, Muerth MJ, St-Denis BG, Minville M, Chaumont D, Caya D, Ludwig R, Turcotte R (2013) An ensemble approach to assess hydrological models' contribution to uncertainties in the analysis of climate change impact on water resources. Hydrol Earth Syst Sci 17:565–578

Vetter T, Huang S, Aich V, Yang T, Wang X, Krysanova V, Hattermann F (2015) Multi-model climate impact assessment and intercomparison for three large-scale river basins on three continents. Earth Syst Dynam 6: 17–43

Weedon GP, Gomes S, Viterbo P, Shuttleworth WJ, Blyth E, Osterle H, Adam JC, Bellouin N, Boucher O, Best M (2011) Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century. J Hydrometeorol 12:823–848

Yilmaz KK, Gupta HV, Wagener T (2008) A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resour Res 44