Climatic Change (2020) 163:1121–1141
https://doi.org/10.1007/s10584-020-02927-8

EDITORIAL

How evaluation of hydrological models influences results of climate impact assessment—an editorial

Valentina Krysanova[1] · Fred F. Hattermann[1] · Zbigniew W. Kundzewicz[1,2]

Received: 15 October 2020 / Accepted: 29 October 2020 / Published online: 7 December 2020

Abstract

This paper introduces the Special Issue (SI) "How evaluation of hydrological models influences results of climate impact assessment." The main objectives were as follows: (a) to test a comprehensive model calibration/validation procedure, consisting of five steps, for regional-scale hydrological models; (b) to evaluate performance of global-scale hydrological models; and (c) to reveal whether the calibration/validation methods and the model evaluation results influence climate impacts in terms of the magnitude of the change signal and the uncertainty range. Here, we shortly describe the river basins and large regions used as case studies; the hydrological models, data, and climate scenarios used in the studies; and the applied approaches for model evaluation and for analysis of projections for the future. After that, we summarize the main findings. The following general conclusions could be drawn. After successful comprehensive calibration and validation, the regional-scale models are more robust and their projections for the future differ from those of the model versions after the conventional calibration and validation. Therefore, climate impacts based on the former models are more trustworthy than those simulated by the latter models. Regarding the global-scale models, using only models with satisfactory or good performance on historical data and weighting them based on model evaluation results is a more reliable approach for impact assessment compared to the ensemble mean approach that is commonly used. The former method provides impact results with higher credibility and reduced spreads in comparison to the latter approach. The studies for this SI were performed in the framework of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP).

This article is part of a Special Issue on "How evaluation of hydrological models influences results of climate impact assessment," edited by Valentina Krysanova, Fred Hattermann, and Zbigniew Kundzewicz.

✉ Valentina Krysanova
krysanova@pik–potsdam.de

[1] Potsdam Institute for Climate Impact Research, Potsdam, Germany

[2] Institute of Agricultural and Forest Environment of the Polish Academy of Sciences, Poznań, Poland

1 Introduction

Problems related to freshwater are getting increasingly important globally, so that model-based hydrological projections for the future are of considerable social relevance. There are still recognized weaknesses and uncertainties in climate modeling and in large-scale hydrological modeling leading to large spreads of climate impact projections. Therefore, improvement of credibility of projections and reduction of their uncertainty are urgently needed. However, the progress achieved in the last three decades has been huge, even if reduction of projection uncertainty has been limited. Nevertheless, gradual movement from the category of "unknown unknowns" to "unknown knowns" is definitely advantageous.

In order to grasp the size of progress, it is interesting to take recourse to a pioneering paper by Russell and Miller (1990), which showed that general circulation models (GCMs) can be directly used to roughly calculate runoff for the major rivers of the world. This was very important, because in order to gather trust in projections for the future, it is necessary to demonstrate that the models can cope with mimicking historical reference situation. However, comparison of the simulated and observed mean annual runoff for 33 large rivers of the world presented by Russell and Miller (1990) showed very large differences. For example, the modeled mean annual runoff for the Orange was 123 km^3/year, as compared to the observed value of 11 km^3/year; hence, the difference was more than 11-fold. For the Amazon, it was 2332 vs 6300 km^3/year. Smallest differences between the modeled and observed values, below 10%, were reported for the Amur and the St Lawrence rivers. Large errors were present already in modeled precipitation. For instance, the modeled precipitation for the Yellow river basin was 1407 km^3/year, vs the mean observed value of 547 km^3/year. The modeling process involved an amplification of the precipitation error to arrive at a much greater, in relative terms, river-discharge error. Nevertheless, it was just a start that demonstrated that progress is needed in various aspects of the modeling process, including development of basin-wide parameters in the models (Miller et al. 1994). These early papers on the direct use of GCMs were regarded as a benchmark to compare developments.

Now, the GCMs are not directly used for simulation of river runoff but serve as drivers of global-, continental-, and catchment-scale hydrological models, which simulate terrestrial hydrological cycle including river discharge. However, there is a scale mismatch between the large-scale climate models and the catchment-scale hydrological models driven by them, which needs a solution. Water is managed at the catchment scale, and adaptation to changing conditions is being done at the regional or local scales, while global climate models work on large spatial grids of 2–3°, so that climate model outputs are usually downscaled and bias corrected (Krysanova et al. 2016). The use of ensembles of climate models for projecting climate impacts, i.e., doing multiple hydrological model simulations driven by outputs of several climate model runs in order to represent the range of possible futures under the climate scenario of concern, started about 15 years ago (e.g., Milly et al. 2005; Nohara et al. 2006), and now, it is an established approach.

However, until almost a decade ago, studies reported in scientific literature used mostly a single hydrological model to assess climate change impacts on different aspects related to water resources. In the meantime, the progress of hardware and software has made it possible to perform and repeat, within a short time, complex calculations that use large quantities of data and numerous equations, so that the computational barriers largely disappeared. The use of multi-model ensembles of hydrological models (HMs) for projecting impacts along with the ensembles of climate scenarios started in the Water Model Intercomparison Project

(WaterMIP, see Haddeland et al. 2011; Hagemann et al. 2012), and followed in the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP, e.g., Dankers et al. 2014; Prudhomme et al. 2013), and now, it is quite common. The critical barriers are nowadays related to data availability and understanding of processes implemented in models, unlike in the past, when the computational burden used to hamper progress (Krysanova et al. 2016).

In addition to the ensemble approach, another important issue is related to consideration (or lack of consideration) of hydrological model performance in reproducing observed data. Two approaches can be distinguished in recent climate change impact studies regarding issues related to performance of HMs in the historical period: (1) using an unweighted multi-model ensemble disregarding their performance (recommended in Christensen et al. 2010; Gudmundsson et al. 2012), and (2) applying models or ensembles of models after their evaluation and taking into account their performance (advised in Prudhomme et al. 2011; Roudier et al. 2016; Krysanova and Hattermann 2017). Traditionally, the first approach is mostly used in global- and continental-scale studies, and the second one is widely applied at the regional and catchment scales.

Several authors (Coron et al. 2012; Krysanova et al. 2016; Donnelly et al. 2016) noted that calibration and validation of a hydrological model are important before applying it for impact assessment, in order to improve its performance and reduce the uncertainty of impacts. The importance of model performance is closely connected with the necessity of model evaluation, which reveals how well the models perform in the historical reference period against observed river discharge and other variables. The model evaluation could be done for the calibrated and validated models applied at the regional or catchment scales (see Gelfan et al. 2020), and also for uncalibrated models, which are usually applied at the global scale (see overview of papers in Krysanova et al. 2020). For a comparison of the performances of the regional- and global-scale hydrological models, see Hattermann et al. (2017). The term "evaluation" can have different meanings in relation to hydrological models. Sometimes, it is understood as equivalent to "calibration and validation," and sometimes as a step applied independently of calibration (after it or without it).

In the paper on hydrological model performance, Krysanova et al. (2018) discussed both approaches and confirmed the hypothesis that "*a good performance of hydrological models increases confidence of projected climate change impacts, and decreases uncertainty of projections related to hydrological models*" based on analysis of pros and cons presented in the referenced papers and examples from recent impact studies. Besides, they suggested new five-step guidelines for evaluation of catchment- and global-scale hydrological models in the historical period, as well as criteria for model rejection from a multi-model ensemble as a poorly performing outlier.

This Special Issue (SI) is needed to test the suggested model evaluation guidelines and to analyze their effects on climate impact results. The main objectives of this SI are as follows: (a) to test the five-step comprehensive model calibration/validation procedure (Krysanova et al. 2018) for the regional-scale hydrological models, (b) to evaluate performance of the global-scale hydrological models, and (c) to reveal whether the calibration/validation methods and model evaluation results influence climate impacts in terms of the magnitude of the change signal and uncertainty ranges. This was done in several papers using the regional-scale models, by comparing the impacts and projection spreads based on a conventional simplified calibration/validation (only for discharge at the basin outlet) and a comprehensive model evaluation. Then, if the effect was notable, and since the comprehensive evaluation includes special robustness tests for future climate, the model version based on it should be considered

as more suitable for impact assessment. We expect that a comprehensive model evaluation in comparison with the commonly used simplified approach can lead to more credible climate impact results with a lower uncertainty of projections related to HMs. This hypothesis was tested in the papers of this SI.

2 Overview of studies, case study areas, and data

Twelve thematic papers are included in this Special Issue. Eight of them are focused on the main research question posed in the title: whether and how impact model evaluation influences results of assessment of climate impact on water. Seven papers of those eight investigate the influence of model calibration/validation methods on simulated impacts in terms of the long-term mean annual and mean monthly changes in river discharge, as well as effects on projection spreads related to hydrological models. In most cases, a conventional (that is, simple) calibration/validation approach is compared against an enhanced (or comprehensive) approach, based on the five-step model evaluation suggested in Krysanova et al. (2018).

One paper is devoted to a systematic evaluation of global water models in the Arctic region, though impact assessment is not included (Gädeke et al. 2020). The focus of three remaining papers is on other specific research questions, such as testing a new calibration method for a large region in Africa (Chawanda et al. 2020), selection of climate ensemble members based on simulated streamflow (Kiesel et al. 2020), and reviewing sources of uncertainty in climate impact projections (Dankers and Kundzewicz, 2020), but all three have indirect relations to the main topic of this SI.

An overview of the 11 research papers (except for the 12th review paper) listing their case study areas, models applied, calibration/validation approaches, and main foci is given in Table 1.

2.1 Case study areas

All case study areas tackled in the papers of this Special Issue are presented on maps in Figs. 1 and 2. Ten river basins located in Europe, Asia, North and South America, and the Southern African region, including two large drainage basins of the rivers Orange and Limpopo, where various regional-scale models were calibrated/validated and applied for climate impact studies are shown in Fig. 1. The main characteristics of these basins and the region are presented in Table 2.

Figure 2 presents the pan-European domain (about 79% of the European continent, excluding some areas in the eastern part that drain to the Caspian Sea but including Turkey and a small portion of Middle East) where the continental-scale model E-HYPE (Hundecha et al. 2020) was applied, and 58 large river basins on six continents for which the global-scale models were evaluated. These 58 basins are distributed among eight hydrobelts, as defined by Meybeck et al. (2013) (see Fig. 1 in Krysanova et al. 2020). Compared to the Köppen classification of climate zones, the classification of land areas in hydrobelts considers watershed boundaries and other geo-hydrological factors, and therefore, it lends itself well to applications in hydrological modeling. The areas of all 58 basins are larger than 50,000 km^2, matching the crude output resolution of the global models, which is 30′ (0.5° × 0.5° latitude-longitude resolution). The drainage areas of 14 of the 58 basins are less than 100,000 km^2, and 10 basins have drainage areas larger than 1 million square kilometers. The

Table 1 An overview of modeling studies included in the Special Issue (*abbreviations*: *U.* upper, *HM* hydrological model, *LSM* land surface model)

Paper by	Spatial scale	Case study river basin(s) or region	Number and scale of models	Number of calibration/validation approaches	Evaluation of HM performance/including contrast climates	Weighting based on performance?	Influence of model evaluation methods/results on impacts: are the differences notable?	Influence of model evaluation on HM uncertainty: are the differences notable?	Selection of climate models based on HM evaluation
Huang et al.	One to three river basins	3: Rhine (Europe), U. Mississippi (USA), U. Yellow (China)	3 regional HMs: SWAT, SWIM, VIC	2: simple and enhanced	Yes/yes	Yes (with two methods)	Yes (notable in 2 basins of 3)	Yes (notable in 3 basins)	–
Gelfan et al.		2: Lena (Russia), Mackenzie (Canada)	1 regional HM: ECOMAG; 1 global LSM: SWAP	3: no calibration, simple, and multi-site	Yes/yes	No	Yes	Yes	–
Mishra et al.		1: U. Godavari (India)	3 regional HMs: SWAT, SWIM, VIC	2: simple and enhanced	Yes/no	Yes	Yes	Yes	–
Wen et al.		1: U. Yangtze (China)	3 regional HMs: SWAT, HBV-D, VIC	2: simple and enhanced	Yes/yes	Yes	Yes	Yes	–
Ismail et al.		1: U. Indus (China, India, Pakistan)	2 regional HMs: SRM-G and VIC-Glacier	2: simple and enhanced	Yes/yes	No	Yes	Yes	–
Koch et al.		1: Pajeú (Brazil)	1 regional HM: SWIM	2: simple and enhanced	Yes/no	No	Yes	Yes	–
Kiesel et al.		1: U. Danube (Europe)	1 regional HM: COSERO	1: enhanced	Yes/yes	–	–	–	Yes
Hundecha et al.	Large regions and multiple large basins	Pan-European domain	1 continental HM: E-HYPE	2: for major river basins, and for smaller scale catchments	Yes/no	No	Yes (similar large-scale pattern, but notable differences locally)	Yes	–
Chawanda et al.		Southern Africa	1 regional HM: SWAT+	1: Hydrological Mass Balance Calibration	Yes/no	No	Yes	Yes	–
Gaedeke et al.		6 major pan-Arctic river basins	9 global water models	No calibration (except one model)	Yes/no	–	–	–	–
Krysanova et al.		57 large river basins on six continents	6 global water models	No calibration (except one model)	Yes/no	Yes	Yes (notable in 9 of 12 selected basins)	Yes (notable in 9 basins of 12)	–

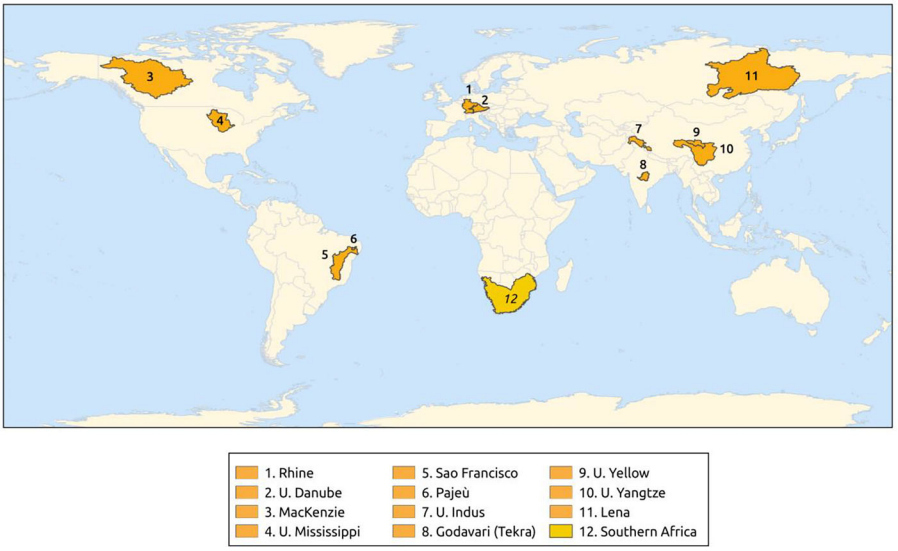

Fig. 1 Case study areas with application of regional-scale models: the drainage basins of the rivers Rhine, Upper Danube, Upper Mississippi, Pajeú (a sub-basin of São Francisco), Mackenzie, Lena, Upper Indus, Godavari (until Tekra), Upper Yellow, and Upper Yangtze and the Southern African region

characteristics of these basins can be found in Table S1 in Krysanova et al. (2020) and Table 1 in Gädeke et al. (2020).

The case study areas where the regional-scale models were applied (Fig. 1) belong to six hydrobelts in total, whereas six basins are located in the northern mid-latitude belt (Table 2).

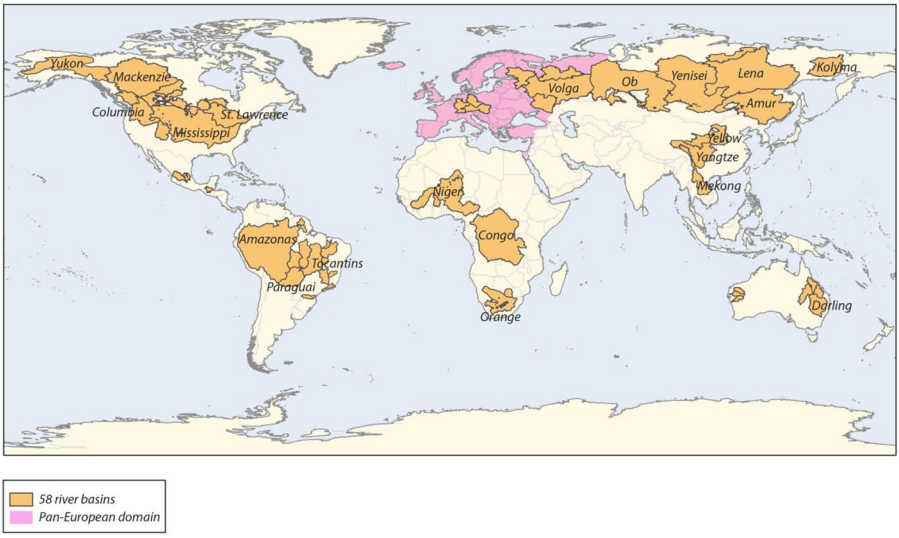

Fig. 2 Case study areas with application of continental- and global-scale models: pan-European domain and 58 large river basins distributed among eight hydrobelts. Names are added only for river basins larger than 470,000 km^2

Table 2 Characteristics of study areas where the regional-scale models were applied. *T* temperature, *P* precipitation, *Q* discharge, *NML* northern mid-latitude, *NST* northern subtropical, *SST* southern subtropical, *SDR* southern dry, *SML* southern mid-latitude hydrobelts

Continent	Basin/region	Last gauge(s)	Drainage area, km^2	Altitude, m a.s.l			T, deg. C 1971–2000 (average)	P, mm 1971–2000 (average)	Q, mm 1971–2000 (average)	Runoff coefficient	Hydrobelt(s)
				Min	Mean	Max					
Europe	Rhine	Lobith	160,800	15	497	4075	8.7	1038	457	0.44	NML
Europe	U. Danube	Vienna	101,810	165	789	4049	6.2	1104	589	0.53	NML
N. America	Mackenzie	Arctic Red River	1,660,000	0	1437	3435	−4.3	435	171	0.39	Boreal
N. America	U. Mississippi	Alton	444,185	124	305	631	7.3	967	257	0.27	NML
S. America	Pajeú	Floresta	12,266	300	732	1170	25.5	655	44	0.07	SST
Asia	Lena	Stolb	2,460,000	1	620	2783	−10.2	384	201	0.52	Boreal
Asia	U. Yellow	Tangnaihai	121,000	2673	4125	6248	−2	506	169	0.33	NML
Asia	U. Yangtze	Cuntan	804,859	139	2903	6575	6.8	768	389	0.51	NML
Asia	U. Indus	Tarbela	173,345	475	4386	8569	1.1*	366*	418*	–	NML
Asia	Godavari	Tekra	119,781	102	573	1125	26.5	1135	312	0.28	NST
Africa	Southern Africa	Vioolsdrif (Orange), Sicacate (Limpopo)	2,179,400	0	1030	3320	19.5**	475**	10**	0.02**	SST, SDR, SML

*For the period 2000–2016

**For the period 1979–2000

The two largest basins for regional-scale applications are the Lena and the Mackenzie (Fig. 1). The area of the Southern African region is also larger than a million of square kilometers. The smallest is the Pajeú catchment, a sub-basin of the São Francisco basin. The mean and maximum elevations are smallest in the Upper Mississippi basin. In contrast, the Upper Yellow river basin is mountainous, with elevations above 2673 m a.s.l.

Climatic conditions are also quite different in the basins where regional-scale models were applied (Table 2). The average annual temperature is above 25 °C in the Godavari and the Pajeú catchments, and below 0 °C in three catchments, of the rivers Mackenzie, Lena, and Upper Yellow. The average annual precipitation ranges from less than 500 mm (in three basins and in Southern Africa) to more than 1000 mm (in three basins). The long-term average runoff ranges from about 10 mm a^{-1} in Southern Africa to 589 mm a^{-1} in the Upper Danube, and the values of runoff coefficient range from 0.02 (Southern Africa) to 0.53 (Upper Danube). Average runoff in the Upper Indus is larger than average precipitation due to significant contribution of melt water from glacial and snow melt: about 70% (Ismail et al. 2020), and therefore, runoff coefficient cannot be defined for this basin. It can be seen that the set of the regional case study basins captures a variety of climatic and hydrological conditions.

2.2 Climate data and scenarios

The following climate data were used in the research reported in the papers of this Special Issue. For evaluation of the regional-scale models in the historical period, three reanalysis datasets were used: EWEMBI (Lange 2018) for the basins of the Upper Indus, Lena, Mackenzie, and Upper Yangtze and for the Southern African region (Ismail et al. Gelfan et al. Wen et al. Chawanda et al. this issue); WATCH (Weedon et al. 2011) for the basins of the Rhine, Upper Mississippi, Upper Yellow, and Pajeú (Huang et al. 2020, Koch et al. 2020); and WFDEI (Weedon et al. 2014) for the Godavari (Mishra et al. 2020).

In turn, global hydrological models (gHMs) were evaluated using model runs driven by WFDEI and GSWP3 (Kim et al. 2014) data for 57 river basins worldwide (Krysanova et al. 2020) and by WATCH, WFDEI, GSWP3, and Princeton (Sheffield et al. 2006) data for six Arctic basins (Gädeke et al. 2020). The continental-scale model E-HYPE applied EFAS-Meteo data (Ntegeka et al. 2013) with resolution of 5 km (Hundecha et al. 2020).

Most regional-scale climate impact studies and the impact assessment for 12 large basins with gHMs were performed using four GCMs from ISIMIP2b: GFDL-ESM2M, HadGEM2-ES, IPSL-CM5A-LR, and MIROC5. Only three studies (Huang et al. Koch et al. and Gelfan et al. this issue) applied five GCMs from ISIMIP2a: HadGEM2-ES, IPSL-CM5A-LR, MIROC280 ESM-CHEM, GFDL-ESM2M, and NorESM1-M. The continental-scale impact assessment (Hundecha et al. 2020) was driven by five climate scenarios from Euro-CORDEX. More details, including the hydrological and geospatial data, can be found in the cited papers.

3 Models and methods

3.1 Models applied

Global hydrological models are designed for the continental and global scales and are usually applied with a crude resolution of 0.5° without calibration. Regional- or catchment-scale hydrological models have much finer resolution, because they are intended for simulation of

catchment characteristics using local input data and applying calibration to observations. Both types of models represent major components of the hydrological cycle, but the level of detail in their description is usually higher in the regional models.

In total, nine regional hydrological models: COSERO, ECOMAG, HBV-D, SRM-G, SWAT, SWAT+, SWIM, VIC, and VIC-Glacier; one continental-scale model: E-HYPE; and ten global-scale models, including four gHMs (H08, MPI-HM, PCR-GLOBWB, and WaterGAP2), five land surface models (LSMs: DBH, JULES-W1, MATSIRO, ORCHIDEE, and SWAP), and one dynamic global vegetation model (LPJmL), were applied in the studies reported in this SI. The LSMs and LPJmL include the full hydrological cycle with water routing, and in that sense, they can be treated as gHMs as well. The references to all these models and papers from the SI, where they were applied, are presented in Table 3. Three of the regional models: SWAT, SWIM, and VIC were applied for 5–6 river basins.

Five catchment-scale hydrological models ECOMAG, HBV, SWAT, SWIM, and VIC are described in Table 2 in Krysanova and Hattermann (2017). The SWAT+ is a restructured version of SWAT, and VIC-Glacier is an extended version of VIC, including glacier processes. The semi-distributed model SRM-G was designed to simulate runoff in the snowmelt-dominated regions. COSERO is a conceptual hydrological model including glacier mass balance and reservoirs. SWAP is a global-scale LSM applied and calibrated at the catchment scale for two large Arctic basins.

Table 3 Hydrological models applied in this Special Issue

Model name	Reference(s)	Applied in this Special Issue in
Catchment-scale models		
COSERO	Kling et al. 2015	Kiesel et al. 2020
ECOMAG	Motovilov et al. 1999	Gelfan et al. 2020
HBV-D	Bergström and Forsman 1973; Krysanova et al. 1999	Wen et al. 2020
SRM-G	Martinec 1975	Ismail et al. 2020
SWAT	Arnold et al. 1998	Huang et al. 2020; Mishra et al. 2020; Wen et al. 2020
SWAT+	Bieger et al. 2017	Chawanda et al. 2020
SWIM	Krysanova et al. 1998	Huang et. 2020; Mishra et al. 2020; Koch et al. 2020
VIC	Liang et al. 1994	Huang et al. 2020; Mishra et al. 2020; Wen et al. 2020
VIC-Glacier	Liang et al. 1994; Naz et al. 2014	Ismail et al. 2020
Continental-scale model		
E-HYPE	Donnelly et al. 2016	Hundecha et al. 2020
Global-scale models		
DBH	Tang et al. 2007	Gädeke et al. 2020; Krysanova et al. 2020
H08	Hanasaki et al. 2008	Gädeke et al. 2020; Krysanova et al. 2020
JULES-W1	Best et al. 2011	Gädeke et al. 2020
LPJmL	Sitch et al. 2003	Gädeke et al. 2020; Krysanova et al. 2020
MATSIRO	Pokhrel et al. 2015	Gädeke et al. 2020; Krysanova et al. 2020
MPI-HM	Stacke and Hagemann 2012	Gädeke et al. 2020
ORCHIDEE	Traore et al. 2014	Gädeke et al. 2020
PCR-GLOBWB	Wada et al. 2014	Gädeke et al. 2020; Krysanova et al. 2020
SWAP	Gusev and Nasonova 1998	Gelfan et al. 2020
WaterGAP2	Müller Schmied et al. 2016	Gädeke et al. 2020; Krysanova et al. 2020

The semi-distributed process-based hydrological model HYPE simulates components of the water cycle and water quality at the catchment scale. It was set-up for the pan-European domain (Donnelly et al. 2016) and is referred to as E-HYPE.

Table 2 in Gädeke et al. (2020) presents nine global water models evaluated for 57 or six Arctic river basins (Fig. 2). Only one of these nine global models, WaterGAP2, was calibrated (Müller Schmied et al. 2014). More details can be found in the papers cited in Table 3.

3.2 Approaches for model calibration/validation

At first glance, calibration and validation of hydrological models seem to be well-established procedure: a differential split-sample test (DSS, first proposed by Klemeš 1986) applied for multiple gauges and two or three variables. However, this procedure is very rarely applied rigorously in climate impact studies, especially for large river basins, where in most cases a simple split-sample test only for discharge at the catchment outlet is used, or the models are applied without any calibration, as it is usually done for gHMs (Dankers et al. 2014; Hattermann et al. 2017).

Besides, the traditional DSS test may be insufficient for checking the preparedness of hydrological models for impact studies. Thus, Refsgaard et al. (2013) suggested a framework for testing models additionally using proxies of future climate conditions, which could be constructed from data in the historical period. Another option is to test models under contrasting historical climate conditions (e.g., Coron et al. 2012; Gelfan and Millionshchikova 2018). Also, Thirel et al. (2015) suggested special protocols for testing models under changing climate conditions. Additionally, Beven and Smith (2015) recommended to include evaluation of observational data quality and take it into account during calibration/validation.

Summarizing the previous recommendations in literature and based on own experience, Krysanova et al. (2018) suggested five steps for a comprehensive calibration/validation of the catchment-scale models intended for impact studies:

1 Evaluate the quality of observational data and take it into account during the model calibration/validation;
2 Apply a differential split-sample test for calibration/validation to optimize the model simultaneously for periods with different climates, or check for proxy climate;
3 Validate model performance at multiple sites and for multiple variables to ensure internal consistency of the simulated processes;
4 Validate whether the model can reproduce the hydrological indicator(s) of interest to be used for impact assessment;
5 Validate for any observed trends (or lack of trends) in discharge, whether they are adequately reproduced by the model.

The evaluation of data quality could be useful for interpretation of calibration results, e.g., some weaker results could be explained by poor data quality. The periods with contrasting climates for the DSS test could be, for example, sub-periods with (i) warmer and drier years and (ii) colder and wetter years than the average climate (see more details in Gelfan et al. and Huang et al. 2020). The global datasets on evapotranspiration, snow cover, etc. could be used for step 3. The validation at steps 4 and 5 could be performed for the total historical period with data. Step 4 could be omitted, if indicators of interest are already included in steps 2 or 3.

For the global-scale models, Krysanova et al. (2018) did not propose calibration, but suggested a simplified five-step model evaluation procedure, where step 2 was substituted by testing the model performance in historical period and steps 3–5 were also weakened. Besides, it was suggested to apply the spatially dependent model performance criteria, assuming that gHMs cannot produce equally plausible results globally.

3.3 Approaches applied in SI for model evaluation and analysis of projections

All applied approaches could be divided in three categories: for regional-scale studies, continental-scale assessment, and multi-basin studies with gHMs, and shortly described as follows:

A) Regional-scale studies

- A simple calibration/validation approach (only for discharge at the basin outlet) and a comprehensive five-step model evaluation method were applied *at the catchment scale* for seven basins in five papers (Huang et al. 2020, Wen et al. 2020, Ismail et al. 2020, Mishra et al. 2020, and Koch et al. 2020). In the first three papers, models were also evaluated for periods with contrasting climates. After that, models with two different parametrizations were applied for climate impact assessment, and the projections were analyzed: whether these two parametrizations influence signals of change in terms of the long-term mean annual and mean monthly flows and (in some cases) extremes. If differences were below 5%, they were considered negligible, and differences higher than 5% (10%) were considered as notable (moderate).
- Three versions of two models were analyzed by Gelfan et al. (2020) for *the Lena and Mackenzie basins*: non-calibrated versions A with a priori parameters, versions B calibrated against daily streamflow at the basin outlets, and versions C calibrated against daily streamflow at multiple gauges. For that, a slightly modified comprehensive evaluation procedure compared to that in Krysanova et al. (2018) was applied. The model robustness was evaluated for the climatically contrasting periods, and effects on future projections were compared.
- Climate model sub-selection methods were assessed for *the Upper Danube basin*, based on the simulated streamflow with COSERO, which was calibrated/validated using the five-step enhanced method and contrasting climate periods (Kiesel et al. 2020).
- A new method of Hydrological Mass Balance Calibration based on global datasets of climate, discharge, evapotranspiration, reservoirs, and irrigation was applied for *the Southern African region*, and its influence on SWAT+ performance and climate projections was analyzed (Chawanda et al. 2020).

B) A continental-scale study

- Three model calibration approaches were applied for *the pan-European domain* using E-HYPE: calibration for major river basins (37 gauges, minimum size 5000 km^2, model version BM), regionalization through calibration at smaller catchments of tributaries (57 gauges, minimum size 1000 km^2, model version M00), and building an ensemble of ten model versions from M00 through parameter sampling (<MXX>). All model versions were applied for projecting climate change impacts, and

differences were analyzed for various indicators on the annual and seasonal basis (Hundecha et al. 2020).

C) Multi-basin studies with the global-scale models

- Evaluation of performance of six global water models in the historical period was done for *57 large river basins on six continents*, considering monthly and long-term mean monthly dynamics of discharge at outlets, based on four common metrics summarized to an aggregated index (AI, ranging from 0 to 1, the higher the better) (Krysanova et al. 2020). Next, a comparison of projected impacts in terms of magnitude of change and spreads was performed for 12 selected river basins using (i) all models, applying the ensemble mean approach (EM), and (ii) only satisfactorily performing models, applying weighting coefficients (WCO) estimated based on the evaluation results.
- A systematic performance evaluation of nine global water models was carried out for *six major Pan-Arctic watersheds* by Gädeke et al. (2020), considering different hydrological indicators: monthly and long-term mean monthly discharge (for multiple gauges), high and low flow extremes (for the outlets), and snow water equivalent. For that, a similar as above aggregated performance index (API, in %, from 0 to 100) based on the usually used criteria was applied.

4 Main findings

Here, results of the comprehensive evaluation of the regional-scale models and evaluation of the continental- and global-scale models, and effects of model evaluation methods and results on future projections and uncertainty ranges are summarized.

4.1 Model evaluation

4.1.1 Enhanced evaluation of the regional models

It was possible to calibrate and validate different catchment-scale models for various basins using the enhanced 5-step method with satisfactory to good results in most cases (see Table 4 with qualitative assessment of results). The evaluation of models ECOMAG and SWAP calibrated at several gauges for the Lena and Mackenzie was also successful. Only in a few cases, results were weak or poor: for the Yellow and Rhine modeled by SWAT, and for the headwater gauge Zhimenda of the Yangtze modeled by VIC and HBV-D.

4.1.2 Evaluation of the continental- and global-scale models

The performance of the benchmark model version BM calibrated for the major river basins in the pan-European domain was slightly worse than that of the other versions (M00 and <MXX>) calibrated for smaller catchments, both in terms of NSE and PBIAS (Hundecha et al. 2020). The median NSE at the calibration stations for BM and M00 are 0.39 and 0.59, respectively. In terms of PBIAS, both versions underestimate the mean flow with median values of − 17% and − 7% for BM and M00, respectively. The model performance in terms of NSE is similar in the validation period, with median NSE of 0.43 and 0.58 for BM and M00, respectively.

Table 4 Performance of the regional-scale models in the case study basins for river discharge in the validation period based on the comprehensive evaluation method: qualitative assessment based on KGE, NSE, and PBIAS criteria

River basin	SWAT	SWIM	VIC VIC-Glacier	HBV-D	SRM+G	ECOMAG version C	SWAP version C	COSERO
U. Mississippi	*** 4/5 * 1/5 cc+	*** 2/5 ** 3/5 cc+	*** 4/5 ** 1/5 cc+					
Rhine	** 2/5 * 1/5 – 2/5 cc–	*** 3/5 ** 2/5 cc+	*** all cc+					
U. Yellow	– all cc–	*** 2/3 ** 1/3 cc+	***1/3 ** 1/3 * 1/3 cc+					
Godavari	*** 1/4 ** 3/4	*** all	*** 2/4 ** 1/4 * 1/4					
U. Yangtze	*** 3/4 ** 1/4 cc+		*** 3/4 – 1/4 cc+	*** 3/4 * 1/4 cc+				
U. Indus			*** 1/3 ** 2/3 cc+		*** 1/3 ** 1/3 * 1/3 cc+			
Lena						*** all cc+	*** all cc+	
Mackenzie						*** all cc+	*** 2/4 ** 2/4 cc+	
Pajeú		* All						
U. Danube								*** all cc+

***, good performance; **, satisfactory performance; *, weak performance; –, poor performance. *n1/n2* in n1 gauges of n2, *all* in all gauges, *cc+/cc−* test in contrasting climate periods passed/not passed

Table 5 An overview of performance of six global hydrological models in terms of the average aggregated index (AI) for 57 large river basins on six continents

average AI	AI≤0.1	0.1<AI≤0.2	0.2<AI≤0.3	0.3<AI≤0.4	0.4<AI≤0.5	0.5<AI≤0.6	0.6<AI≤0.7	0.7<AI≤0.8	0.8<AI≤0.9
performance	poor	poor	poor	poor	weak	satisfactory	good	good	good
	Assniboine	Ob	Winnipeg	Albany	Amur	Lena	Yangtze		Mekong
	Red	Saskachewan	Frazer	N. Dvina	Yenisei	Kolyma	Iguacu		
	Yellow	Churchill	Panuco	Columbia	MacKenzie	Olenek			
	Don	St Lawrence	Congo	Labe	Mississippi	Usumacinta			
	Colorado	Neva	Itapecuru	Rhine	Maroni	Amazonas			
	Niger	Santiago	Sao Francisco	Danube	Tapajos				
	Parnaiba	White Volta	Murchison	Volga	Burdekin				
	Rio Grande	Jari	Gascoyne	Xingu					
	Paraguai	Limpopo	Paraiba Do Sul	Tocantins					
	Orange	Jequitinhonha	Fitzroy	Asuburton					
	Cooper Creek								
	Darling								

The evaluation results of six global models for 57 river basins (Fig. 2) varied between models and basins (Krysanova et al. 2020). The performance, averaged over six gHMs, was satisfactory or good (average AI > 0.5) in eight basins of 57, whereas 42 basins showed poor performance with average AI ≤ 0.4. Table 5 shows an overview of the average performance of six gHMs in terms of the aggregated index for 57 basins. WaterGAP2 was the best performing model (average AI of 0.67), followed by MATSIRO and PCR-GLOBWB with average AIs of 0.28 and 0.26, respectively. The remaining three models showed quite poor performance with significant overestimation of discharge and the amplitude of seasonal dynamics, with median NSE values for monthly discharge being well below zero.

The evaluation results of nine gHMs over six Arctic watersheds (Gädeke et al. 2020) were similarly weak: the average aggregated performance index API exceeded 50% only for one basin of six (Kolyma) for the monthly and seasonal discharge, and for one basin (Ob) for the high and low flows. WaterGAP2 had the highest API (72%) averaged over all basins for the monthly/seasonal discharge, and MATSIRO had an average API > 50% for the monthly/seasonal discharge and extremes. Two more models showed average API > 50% in two cases: MPI-HM for the monthly/seasonal discharge and LPJmL for high flows. Remaining 21 out of 27 cases demonstrated weak or poor performance.

4.2 Influence of model evaluation methods/results on impacts and uncertainties at different scales

The influence of model calibration/validation *methods (simple and comprehensive)* on the projected impacts and uncertainties was investigated in several papers by comparing the impacts and projection spreads. Also, the effect of model *evaluation results* on impacts and uncertainties was analyzed for several differently calibrated model versions (including the non-calibrated models in one study) in two papers: for the Lena, Mackenzie, and pan-European domain. Besides, the influence of model *evaluation results* on the projections simulated by gHMs was analyzed, applying and comparing the EM and WCO methods (see Section 3.3).

4.2.1 Regional-scale studies

Three river basins investigated in Huang et al. (2020) had different sensitivities of projections to two model parametrizations. Comparison of results has shown moderate to strong influences on the ensemble medians and means of discharge for the Upper Mississippi (differences up to

23%), minor to moderate effects for the Upper Yellow (differences up to 16%), and smaller effects for the Rhine (maximum 7%). For the Mississippi, two calibration methods even led to contradictory signals of change (positive and negative) in terms of mean/median. The shares of uncertainty related to HMs decreased for three hydrological indicators in all basins after the enhanced calibration, except for the high and low flows in the Rhine. However, when SWAT was excluded from the ensemble for this basin due to its poor performance, the shares of the HM uncertainty decreased in all cases. Thus, we can conclude that even a single poorly performing model could substantially increase the HM share of uncertainty and the total uncertainty of projections.

In *the Godavari basin*, the influence of calibration/validation methods on the projected mean annual discharge and high flows was minor for three gauge stations, including the outlet, and notable for one gauge, Bamini (43% of the total area): about 10% difference was found for the projected changes in mean annual flow, and 14% for high flows (Mishra et al. 2020). However, for high flow frequency, considerable influence was noticed: up to 35–40% differences at three gauges, including the outlet.

For *the Upper Yangtze basin*, comparison of impacts based on the simple and comprehensive calibration methods was done for three hydrological models (Wen et al. 2020). The simulated increases in mean annual discharge at the end of the twenty-first century related to the reference period were approximately doubled based on the simple calibration, in comparison to those based on the comprehensive method, with the mean annual differences of 7–8% under RCPs 2.6 and 8.5. The same tendency was found for high flow. For low flow, changes in different directions were simulated based on two methods under RCPs 2.6 and 4.5, and differences were up to 15%.

The study for *the Upper Indus basin* with two HMs has shown notable differences in impacts at the annual scale (8–10%) based on two methods for RCPs 2.6 and 8.5 in the mid-century and far future periods (Ismail et al. 2020). The median changes based on two methods differed in sign in all periods under RCP2.6 and in the near future for RCP8.5. At the monthly scale, the largest differences were found in March (− 17%) and October–November (18–19%) in the far future under RCP8.5. The uncertainty contribution from HMs based on the enhanced method was larger in the near future but became negligible and smaller in the far future, in comparison with the conventional method.

Differences in impacts and uncertainties were analyzed for *the Lena and Mackenzie river basins* (Gelfan et al. 2020), based on simulations of three versions of two models (see Section 3.3). Both models simulated increase in the mean annual discharge for the Lena by 43% on average at the end of this century based on the non-calibrated A versions, whereas application of the calibrated versions B and C resulted in 24% and 19% increases, respectively. The corresponding uncertainty spreads were 125%, 63%, and 37% for A, B, and C, respectively. Similar differences were found for the Mackenzie. It was found that A projections differed quite significantly from B and C projections in both basins: by 10–22%, whereas projections based on B and C diverged by 5–6% only.

For *the Pajeú catchment* in a semi-arid area in Brazil, differences between the projected changes in the long-term mean annual/monthly discharges, averaged over all GCMs based on two differently calibrated model versions, were analyzed for the near future 2021–2050 (Koch et al. 2020). The differences were rather low under RCP8.5 and slightly higher (up to 8%) under RCP2.6. The analysis of projections based on two model versions for the Serrinha II Reservoir revealed notable differences between the projected changes in maximum mean monthly volume, up to 9–10% under both RCPs, and in mean reservoir discharge, 10% under RCP2.6.

4.2.2 Continental-scale study

In the study for the pan-European domain, the impacts simulated by the benchmark model, BM, differed distinctly from those of model versions M00 and <MXX> for different indicators, whereas they were similar for the latter two (Hundecha et al. 2020). The median changes projected for mean annual discharge by BM in the whole area were close to zero but biased towards negative changes, whereas M00 projected a moderate increase (median value of 8%) and a wetter pattern. The projected changes in soil moisture had similarly shifted distributions, with a major portion of the distribution being negative for BM, and with the 95th percentile change being positive and about 10% higher for M00. Regarding seasonal changes, all model projections showed a strong increase in discharge in winter: the median increase ranged between 30 and 39% for BM and M00, respectively. In summer, the median increase in aridity projected by BM was nearly twice as much as that of M00 (23% vs 12%). The absolute differences in the projections of the M00 and BM versions ranged between 0 and 55% for the mean annual discharge.

4.2.3 Global-scale models: comparison of two approaches

Comparison of climate change impacts simulated by gHMs was performed for 12 of 18 selected river basins using the EM and WCO approaches (see Section 3.3) (Krysanova et al. 2020), whereas the second approach was based on the model evaluation results. The following results were obtained: (a) impact assessment with WCOs was not possible in six basins (~ 33%) due to poor performance of all models, (b) comparison of impacts resulted in small or negligible differences in four basins (~ 22%), and (c) differences in mean monthly discharge, mean annual discharge, or both were moderate to large in eight basins (~ 44%). A comparison of projection spreads was done for 12 basins, considering 25 to 75% percentiles and full uncertainty ranges. It was found that the spreads were of similar size in four basins; they decreased slightly or moderately (by 15–50%) when the WCO method was applied in five basins, and decreased significantly (by 51–82%) based on the WCO method in three basins.

5 Summary and conclusions

5.1 Model evaluation results

The regional-scale models were successfully calibrated and validated for various basins using the comprehensive 5-step method with satisfactory to good results in most cases. The performance of the benchmark model version of E-HYPE, calibrated for 37 major river basins of the pan-European domain, was only slightly worse than that of the model version calibrated for 57 tributary catchments.

The evaluation of six gHMs for 57 river basins showed satisfactory to good results of WaterGAP2 (acceptable index in 75% of the basins), weaker results of MATSIRO and PCR-GLOBWB (23–26%), and rather poor results of the other three models (only 11–16% of the basins with acceptable indices). The performance, averaged over models, was good or satisfactory only in eight river basins of 57 (14%). Similar results were obtained in another paper, evaluating global water models in the Arctic basins, where WaterGAP2 and MATSIRO showed better results than other models. It was found that the majority of global models

exhibited considerable difficulties in realistically representing observed hydrological processes in many basins.

5.2 Influence of model evaluation methods and results on impacts and uncertainties

The influence of model *evaluation methods* on the projected impacts and uncertainties was investigated in a number of catchments by comparing the impacts and projection spreads. In most cases, notable to moderate differences in projected impacts were found, and in some cases, the differences were stronger. In some basins, application of models with two different parametrizations even led to changes in opposite directions in future periods. The studies, which analyzed projection spreads, concluded that, after comprehensive calibration/validation, models tend to reduce spreads related to HMs.

The influence of *results of comprehensive evaluation* of HMs on impacts and uncertainties was analyzed applying two models in two large river basins, using three differently parametrized versions of each model. The simulated impacts and uncertainty spreads were essentially different between the model versions which successfully passed the evaluation test and the versions which failed to pass it. This allowed authors to conclude (Gelfan et al. 2020) that the successful comprehensive evaluation of model versions increases confidence in their projections in comparison with projections of model versions which did not pass the test.

The influence of *model evaluation results* on impacts and uncertainties was also analyzed for gHMs, applying the traditional ensemble mean method and the approach with weighting coefficients. In most cases, application of the WCO approach based on model evaluation and using only models with satisfactory or good performance resulted in different projections with reduced spreads, compared to the projections using the EM approach, based on all models disregarding their performance. However, the impacts and projection spreads were quite similar for the basins where most gHMs showed acceptable or good performance.

5.3 Updated guidelines for calibration/validation

Based on the experience gained in the practical application of the five evaluation steps in seven papers of this SI, they have been further developed and could be slightly re-formulated as follows:

1. Evaluate the quality of observational data and take it into account during the model evaluation for interpretation of results;
2. Apply a differential split-sample test (or its modification) for discharge at multiple gauges to calibrate/validate the model simultaneously for periods with contrasting climates;
3. Validate model performance for 1–2 additional variables (e.g., evapotranspiration, snow) to ensure internal consistency of the simulated processes; if not successful, return to step 2;
4. Validate whether the model can reproduce hydrological indicator(s) of interest for impact assessment (e.g., extremes); if not successful, return to step 2;
5. Check whether the observed trends (or lack of trends) are reproduced by the model.

The main modifications are as follows: strengthening of point 2 and adding a possibility of iterations if validation at steps 3 or 4 is not successful.

5.4 Grappling with uncertainties

Dankers and Kundzewicz (2020) reviewed the sources of uncertainty in the recent projections of climate change impacts on water resources. Since, in addition to GCM uncertainty, structural uncertainty of impact models can be a significant component of the overall uncertainty, studies that are based on a single hydrological model may well be overconfident and do not adequately sample the uncertainty range, even if they use multiple driving climate models. It may be possible to reduce the spread in multi-model ensemble results by down-weighting or eliminating models that are unable to mimic observed components of the water cycle in a particular catchment. Since uncertainty in applications of large-scale models is greater at the smaller scale of a river basin, this provides a challenge to local adaptation decisions, because greater uncertainty may require costlier protective measures. However, this paper demonstrated that large uncertainties at the local scale do not preclude more robust projections at the global scale (see also Hattermann et al. 2018).

5.5 Conclusions

The comprehensive five-step evaluation of the catchment-scale hydrological models was compared with the commonly used simplified calibration/validation approach in terms of simulated impacts. It was shown that (a) in most cases, impact results for annual and monthly means notably differ between simulations based on two approaches, and (b) uncertainties of projections related to hydrological models are usually reduced after the enhanced model evaluation. As models after successful enhanced evaluation are more robust, and projections based on them differ from the projections simulated by models after simple calibration/validation, we can conclude that the models after the successful enhanced evaluation are more reliable and impacts based on them are more trustworthy.

The evaluation of global hydrological models and their application for impact assessment using the EM and WCO approaches has shown that using only models with satisfactory or good performance on historical data and weighting them based on model evaluation results is a more reliable approach for impact assessment compared to the ensemble mean approach. The obtained results allowed to conclude that, in most cases, the WCO method provides impact results with higher credibility and reduced spreads in comparison to the EM approach. Regarding further gHM applications for climate impact assessment, we could recommend the following: (a) model evaluation should be always done in advance of impact assessment with gHMs; (b) improvement of gHMs performance, also using calibration, is necessary in order to include more suitable models in ensembles for projecting impacts; and (c) inclusion of region-specific processes (e.g., permafrost in the Arctic) in gHMs is also necessary for making impact results more trustworthy.

The value of this SI for the science and the stakeholder community is in the following:

(i) As the comprehensive model evaluation includes special tests of model robustness under changed climate conditions, it improves credibility of simulated impacts for stakeholders, and the reduced spreads of projections related to hydrological models make them more distinct.
(ii) The methodology of comprehensive evaluation of hydrological models developed before and thoroughly tested in the papers in this Special Issue could be useful for a broader

scientific community doing research for sectors beyond hydrology and water resources, where climate change impact assessment is relevant.

(iii) This Special Issue contributes to the improvement of credibility and reduction of uncertainty of climate impact projections that are needed for development of adaptation strategies to climate change.

Acknowledgments The authors are grateful to ISIMIP for providing climate scenarios and gHM simulations for studies in this SI, and to GRDC for providing discharge data. The authors would like to thank Alexander Gelfan for discussion of the paper summary and Rocio Rivas and Iulii Didovets for preparation of maps.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Arnold JG et al (1998) Large-area hydrologic modeling and assessment: part I. Model development. J Am Water Res Assoc 34(1):73–89

Bergström S, Forsman A (1973) Development of a conceptual deterministic rainfall-runoff mode. Nord Hydrol 4: 240–253

Best MJ et al (2011) The Joint UK Land Environment Simulator (JULES), model description - part 1: energy and water fluxes. Geosci Model Dev 4:677–699. https://doi.org/10.5194/gmd-4-677-2011

Beven KJ, Smith PJ (2015) Concepts of information content and likelihood in parameter calibration for hydrological simulation models. In: ASCE Journal of Hydrologic Engineering. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000991

Bieger K et al (2017) Introduction to SWAT+, a completely restructured version of the soil and water assessment tool. J Am Water Resour Assoc 53(1):115–130. https://doi.org/10.1111/1752-1688.12482

Chawanda CJ et al (2020) Mass balance calibration and reservoir representations for large scale hydrological impact studies using SWAT+. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02924-x

Christensen JH et al (2010) Weight assignment in regional climate models. Clim Res 44:179–194. https://doi.org/10.3354/cr00916

Coron L et al (2012) Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. Water Resour Res 48:W05552. https://doi.org/10.1029/2011WR011721

Dankers R et al (2014) First look at changes in flood hazard in the inter-sectoral impact model intercomparison project ensemble. PNAS 111:3257–3261. https://doi.org/10.1073/pnas.1302078110

Dankers R, Kundzewicz ZW (2020) Grappling with uncertainties in climate impact projections of water resources. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02858-4

Donnelly C et al (2016) Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. Hydrol Sci J 61(2):255–273. https://doi.org/10.1080/02626667.2015.1027710

Gädeke A et al (2020) Performance evaluation of global hydrological models in six large Pan-Arctic watersheds. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02892-2

Gelfan A et al (2020) Does a successful comprehensive evaluation increase confidence in a hydrological model intended for climate impact assessment? Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02930-z

Gelfan A, Millionshchikova T (2018) Validation of a hydrological model intended for impact study: problem statement and solution example for Selenga River basin. Water Res 45(S1):90–101. https://doi.org/10.1134/S0097807818050354

Gudmundsson L et al (2012) Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. Water Resour Res 48:11. https://doi.org/10.1029/2011WR010911
Gusev YM, Nasonova ON (1998) The Land Surface Parameterization scheme SWAP: description and partial validation. Glob Planet Chang 19(1–4):63–86
Haddeland I et al (2011) Multimodel estimate of the global terrestrial water balance: setup and first results. J Hydrometeorol 12(5):869–884
Hagemann S et al (2012) Climate change impact on available water resources obtained using multiple global climate and hydrology models. Earth Syst Dynam Discuss 3(3–4):1321–1345
Hanasaki N et al (2008) 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. https://doi.org/10.5194/hess-12-1007-2008
Hattermann FF et al (2017) Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large scale river basins. Clim Chang 141(3):561–576. https://doi.org/10.1007/s10584-016-1829-4
Hattermann FF et al (2018) Sources of uncertainty in hydrological climate impact assessment: a cross-scale study. Environ Res Lett 13(1):015006. https://doi.org/10.1088/1748-9326/aa9938
Huang S et al (2020) Impacts of hydrological model 1 calibration on projected hydrological changes under climate change – a multi-model assessment in three large river basins. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02872-6
Hundecha Y et al (2020) Effect of model calibration strategy on climate projections of hydrological indicators at a continental scale. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02874-4
Ismail MF et al (2020) Comparison of two model calibration approaches and their influence on future projections under climate change in the Upper Indus Basin. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02902-3
Kiesel J et al (2020) Streamflow-based evaluation of climate model sub-selection methods. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02854-8
Kim H et al (2014) Development of a new global dataset for offline terrestrial simulations - for global soil wetness project phase 3. Institute of Industrial Science, The University of Tokyo, Tokyo https://www.isimip.org/gettingstarted/input-data-bias-correction/details/4/
Klemeš V (1986) Operational testing of hydrological simulation models. Hydrol Sci J 31:13–24
Kling H et al (2015) Performance of the COSERO precipitation–runoff model under non-stationary conditions in basins with different climates. Hydrol Sci J 60(7–8):1374–1393. https://doi.org/10.1080/02626667.2014.959956
Koch H et al (2020) Effects of model calibration on hydrological and water resources management simulations under climate change in a semi-arid watershed. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02917-w
Krysanova V et al (1998) Development and test of a spatially distributed hydrological water quality model for mesoscale watersheds. Ecol Model 106(2–3):261–289
Krysanova V et al (1999) Modelling river discharge for large drainage basins: from lumped to distributed approach. Hydrol Sci J 44:313–331
Krysanova V et al (2016) Assessment of climate change impacts on water resources, Chapter 148. In: Singh V (ed) Handbook of applied hydrology, 2nd edn. McGraw-Hill, New York
Krysanova V, Hattermann F (2017) Intercomparison of climate change impacts in 12 large river basins: overview of methods and summary of results. Clim Chang 141(3):363–379. https://doi.org/10.1007/s10584-017-1919-y
Krysanova V et al (2018) How the performance of hydrological models relates to credibility of projections under climate change. Hydrol Sci J 63(5):696–720. https://doi.org/10.1080/02626667.2018.1446214
Krysanova V et al (2020) How evaluation of global hydrological models can help to improve credibility of river discharge projections under climate change. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02840-0
Lange S (2018) Bias correction of surface downwelling longwave and shortwave radiation for the EWEMBI dataset. Earth Syst Dynam 9(2):627–645
Liang X et al (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J Geophys Res 99:14,415–14,428
Martinec J (1975) Snowmelt-runoff model for stream flow forecasts. Nord Hydrol 145–154
Meybeck M et al (2013) Global hydrobelts and hydroregions: improved reporting scale for water-related issues? HESS 17:1093–1111
Miller JR et al (1994) Continental-scale river flow in climate models. J Clim 7(6):914–928
Milly PCD et al (2005) Global pattern of trends in streamflow and water availability in a changing climate. Nature 438:347–350. https://doi.org/10.1038/nature04312
Mishra V et al (2020) Does comprehensive evaluation of hydrological models influence projected changes of mean and high flows in the Godavari River basin? Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02847-7

Motovilov YG et al (1999) Validation of a distributed hydrological model against spatial observations. Agric For Meteorol 98-99:257–277. https://doi.org/10.1016/S0168-1923(99)00102-1
Müller Schmied H et al (2014) Sensitivity of simulated global scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. HESS 18:3511–3538. https://doi.org/10.5194/hess-18-3511-2014
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. https://doi.org/10.5194/hess-20-2877-2016
Naz BS et al (2014) Modeling the effect of glacier recession on streamflow response using a coupled glacio-hydrological model. Hydrol Earth Syst Sci 18:787–802
Nohara D et al (2006) Impact of climate change on river runoff. J Hydrometeorol 7:1076–1089
Ntegeka V et al. (2013) EFAS-Meteo: a European daily high-resolution gridded meteorological data set for 1990–2011. Report EUR, 26408
Pokhrel YN et al (2015) Incorporation of groundwater pumping in a global land surface model with the representation of human impacts. Water Resour Res 51:78–96. https://doi.org/10.1002/2014wr015602
Prudhomme C et al (2011) How well do large-scale models reproduce regional hydrological extremes in Europe? J Hydrometeorol 12(6):1181–1204. https://doi.org/10.1175/2011JHM1387.1
Prudhomme C et al (2013) Hydrological droughts in the 21st century: hotspots and uncertainties from a global multi-model ensemble experiment. PNAS 111(9):3262–3267
Refsgaard JC et al (2013) A framework for testing the ability of models to project climate change and its impacts. Clim Chang 122(1–2):271–282. https://doi.org/10.1007/s10584-013-0990-2
Russell GL, Miller JR (1990) Global river runoff calculated from a global atmospheric general circulation model. J Hydrol 117(241):254
Roudier P et al (2016) Projections of future floods and hydrological droughts in Europe under a +2°C global warming. Clim Chang 135(2):341–355
Sheffield J et al (2006) Development of a 50-yr high-resolution global dataset of meteorological forcings for land surface modeling. J Clim 19(13):3088–3111
Sitch S et al (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. Glob Chang Biol 9:161–185. https://doi.org/10.1046/j.1365-2486.2003.00569.x
Stacke T, Hagemann S (2012) Development and evaluation of a global dynamical wetlands extent scheme. Hydrol Earth Syst Sci 16:2915–2933. https://doi.org/10.5194/hess-16-2915-2012
Tang Q et al (2007) The influence of precipitation variability and partial irrigation within grid cells on a hydrological simulation. J Hydrometeorol 8:499–512. https://doi.org/10.1175/jhm589.1
Thirel G et al (2015) On the need to test hydrological models under changing conditions. Hydrol Sci J 60(7–8): 1165–1173. https://doi.org/10.1080/02626667.2015.1050027
Traore AK et al (2014) Evaluation of the ORCHIDEE ecosystem model over Africa against 25 years of satellite-based water and carbon measurements. J Geophys Res Biogeosci 119:1554–1575. https://doi.org/10.1002/2014JG002638
Wada Y et al (2014) Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. Earth Syst Dynam 5:15–40. https://doi.org/10.5194/esd-5-15-2014
Weedon GP et al (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(5):823–848
Weedon GP et al (2014) The WFDEI meteorological forcing data set: WATCH Forcing data methodology applied to ERA-Interim reanalysis data. Water Resour Res. https://doi.org/10.1002/2014WR015638
Wen S et al (2020) Comprehensive evaluation of hydrological models for climate change impact assessment in the Upper Yangtze River Basin, China. Clim Chang, this issue. https://doi.org/10.1007/s10584-020-02929-6