Evaluation and machine learning improvement of global hydrological model-based flood simulations

Tao Yang1,2,3, Fubao Sun1,3,4,5, Pierre Gentine2,6, Wenbin Liu1,4,6, Hong Wang1, Jiabo Yin2,7, Muye Du1,8 and Changming Liu1

1 Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, People’s Republic of China
2 Department of Earth and Environmental Engineering, Columbia University, New York, United States of America
3 University of Chinese Academy of Sciences, Beijing, People’s Republic of China
4 State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, People’s Republic of China
5 Akesu National Station of Observation and Research for Oasis Agro-ecosystem, Akesu, Xinjiang, People’s Republic of China
6 Earth Institute, Columbia University, New York, United States of America
7 State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, People’s Republic of China
8 Authors to whom any correspondence should be addressed.

E-mail: sunfb@igsnrr.ac.cn and liuwb@igsnrr.ac.cn

Keywords: flood simulation, machine learning, global hydrological model, long short-term memory

Supplementary material for this article is available online

Abstract

A warmer climate is expected to accelerate global hydrological cycle, causing more intense precipitation and floods. Despite recent progress in global flood risk assessment, the accuracy and improvement of global hydrological models (GHMs)-based flood simulation is insufficient for most applications. Here we compared flood simulations from five GHMs under the Inter-Sectoral Impact Model Intercomparison Project 2a (ISIMIP2a) protocol, against those calculated from 1032 gauging stations in the Global Streamflow Indices and Metadata Archive for the historical period 1971–2010. A machine learning approach, namely the long short-term memory units (LSTM) was adopted to improve the GHMs-based flood simulations within a hybrid physics-machine learning approach (using basin-averaged daily mean air temperature, precipitation, wind speed and the simulated daily discharge from GHMs-CaMa-Flood model chain as the inputs of LSTM, and observed daily discharge as the output value). We found that the GHMs perform reasonably well in terms of amplitude of peak discharge but are relatively poor in terms of their timing. The performance indicated great discrepancy under different climate zones. The large difference in performance between GHMs and observations reflected that those simulations require improvements. The LSTM used in combination with those GHMs was then shown to drastically improve the performance of global flood simulations (especially in terms of amplitude of peak discharge), suggesting that the combination of classical flood simulation and machine learning techniques might be a way forward for more robust and confident flood risk assessment.

1. Introduction

Flooding is one of the costliest natural disasters in the world [1], which has been projected to occur more frequently under global warming conditions [2, 3]. Reasonable and reliable projection of maximum discharge is critical for flood risk assessment and climate change adaptation. Global hydrological models (GHMs) are broadly adopted to assess the impacts of climate change on water availability and scarcity, floods, hazard and risk, water resources management and planning [4–6], on the response of the global hydrological cycle to climate change, and for forecasting and examining the role of water in agriculture production [2, 7–12] at both global and regional scales. Global institutions have been utilizing more GHMs to make first
order global policy decisions on water-relate hazards in recent decades because of the lack of applicable input data to support in-depth regional studies. These examples include the World Bank series on climate change and development [13], IPCC assessment reports [14] and Global Environmental Outlooks [15].

Nonetheless, the accuracies of flood simulations are less frequently evaluated across GHMs under different socio-econ-scenarios and climate zones nowadays. Some existing studies merely focused on typically large basins, or selected stations in accordance to the performance of discharge simulations [5, 10, 12, 16]. To our knowledge, a general evaluation of flood simulations globally is still missing due to the lack of observed global daily streamflow used to calculate flood-related indices. The recently launched Global Streamflow Indices and Metadata Archive data make it possible.

Machine learning has shown remarkable potential in geosciences in recent years for various applications such as land use change detection, precipitation prediction and bias-correcting forecast [17–19]. Deep learning methods have made revolutionary advances in the sequential data-modeling domain [20–24] or for streamflow forecasting [25]. Recently, several studies have used the so-called hybrid modeling approach, coupling physical-models with machine learning [20, 21, 26]. Conceptually, those models aim at leveraging the predictive power of physical-models (generalization) while harvesting existing data to correct biases. Moreover, in term of discharge simulation, the GHMs without a specific calibration to river basins generally do not perform well enough when compared to the specifically calibrated regional models [10]. In this study, we implemented a novel scheme through the combination of GHMs, other existing data and a state-of-the-art machine learning approach accounting for temporal correlations, the long short-term memory units (LSTM), to improve the performance of flood simulations at the global-scale.

The objectives of this study are to: (1) evaluate the performance of GHMs simulations in terms of flood-relevant indices under different hydro-climatological conditions over the historical period; (2) highlight the current strengths and limitations of GHMs-based global-scale flood simulations and (3) detect the potential of LSTM combined with those GHMs to improve global-scale flood modeling. This paper is organized as follows: section 2 introduces the datasets and methods used in this study and the input datasets of the LSTM. Section 3 describes the detailed evaluation and LSTM improvement for GHMs-based flood simulations, followed by conclusions in section 4.

## 2. Data and methods

To evaluate the performance of GHMs forced with observational climate data (1971–2010) within the Inter-Sectoral Impact Model Intercomparison Project 2a (ISIMIP2a) project on flood simulation, we processed 15 routing simulations with the Catchment-based Macro-scale Floodplain (CaMa-Flood) model [27, 28] driven by daily runoff data simulated from five GHMs (section 2.1) under three socio-economic conditions, i.e. naturalized conditions (NONSOC), present-social conditions (PENSSOC) and dynamic-social economic scenarios (VARSOC). We then conducted a comparison in terms of five flood-relevant indices (describing the amplitude and timing of peak discharge, section 2.2) offline-simulated by GHMs and CaMa-Flood against those calculated from GSIM dataset.

### 2.1. Evaluating GHMs-based flood simulation against GSIM data

We used daily runoff simulations (0.5° × 0.5°) for the period 1971–2010 from five GHMs: DBH [29], H08 [30, 31], LPJML [32], PCR-GLOBWB [33, 34] and VIC [35] (please find the general characteristics and configurations of those models in supplementary table S1, available online at stacks.iop.org/ERL/14/114027/ mmedia and www.isimip.org/impactmodels/ under the ISIMIP2a protocol (www.isimip.org/)). Briefly, the ISIMIP2a offers a framework for projecting the impacts of climate change consistently across different impact sectors and spatial scales, and serves as a basis for model evaluation and improvement, allowing for better estimates of the biophysical and socio-economic impacts of climate change at different levels of worlding. The GHMs used here were run under the VARSOC conditions, which account for time-varying (historical) dam constructions, water abstraction and land use changes. We then compared the outputs with GHM-simulations under the NONSOC (no human impacts, such as dams and water abstractions on river flow) and the PENSSOC (with socio-economic conditions fixed at 2005 levels) conditions to test the sensitivity of our main findings to the implementation of direct human influences.

After obtaining daily runoff data from the GHMs (through https://esg.pik-potsdam.de/search/isimip/), we applied the CaMa-Flood model as the river routing approach to simulate basin-wide daily streamflow. The CaMa-Flood is currently the only open-source global river model available (http://hydro.iis.u-tokyo.ac.jp/~yamada/cama-flood/index.html) that is capable of simulating backwater effect within a reasonable computational time and providing flooded area, flooded volume and flood depth. It has been widely used in flood assessment and future projection of global flood risk [12, 16, 36].

It should be noted that none of these GHMs-CaMa-Flood model chains have been calibrated with observed discharge, allowing us to use GSIM as an independent data benchmark to evaluate the performance of these model chains in flood simulations. The
GSIM dataset (available from https://doi.pangaea.de/10.1594/PANGAEA.887470) is a collection of calculated discharge indices (i.e. mean daily streamflow, standard deviation, coefficient of variation, percentiles, the inter quartile range of streamflow at monthly, seasonal and yearly scale), based on daily streamflow observations at more than 35 000 stations globally [37, 38], which is useful for evaluating the performance of flood simulations. To do this, we selected 1032 catchments from the GSIM dataset according to the following criteria:

1. To avoid under-sampling, only catchments with at least 10-year complete and continuous observations during the period of 1971–2010 were chosen.

2. A minimum catchment size of 9000 km² (approximately four grid cells) and a distance between stations and grid line being less than 5% grid edge length were selected, to omit catchments whose hydrological processes could not be correctly represented by the GHMs operating at 0.5°.

To assess the main properties of flood amplitude and timing, five indices (i.e. annual maximum discharge, annual maximum 7 d averaged discharge and their corresponding days of year, the 90-percentile maximum discharge from the yearly part of GSIM dataset) were selected in this study.

2.2. Evaluation metrics

To evaluate the GHMs-based flood simulation, three evaluation metrics, i.e. relative error (RE), Nash–Sutcliffe efficiency (NSE) and Pearson’s correlation coefficient (PCC), which have been widely used in evaluating performance of hydrological models on streamflow simulation [10, 39], were selected. Moreover, RE = 0 indicates a perfect correlation between the observed and simulated streamflow. NSE (ranges from −∞ to 1) is a parameter that determines the relative importance of residual variance (noise) compared with the variance in the measured data, where NSE = 1 indicates a perfect correlation. PCC has a value between +1 and −1, where 1 indicates total positive linear correlation; 0 indicates no linear correlation; and −1 indicates total negative linear correlation.

2.3. Improving GHMs-based flood simulations through LSTM

To future improve the GHMs-based flood simulations, we explored the potential of using a combination of GHMs and a state-of-the-art machine learning approach (LSTM). The feedforward neural network is usually comprised of an input layer, a few hidden layers and an output layer. A physics-guided neural network (PGNN) leverages the output of physics-based model simulations along with observational features to generate predictions using neural network architecture [20, 26]. Here, we put forward a general PGNN method which leverages the advantages of these two approaches by combining knowledge-based GHMs and a novel technique, the LSTM, to build a hybrid simulation scheme. Briefly, LSTM is a special recurrent neural network, which feeds not only input data but also remember the states of the hidden neurons of the previous time steps, and simplifies solving the temporal lag and autocorrelation (these two properties exist in discharge simulations) of the data and avoids vanishing gradients [40]. Thus, we combined GHMs-CaMa-Flood with LSTM to improve the performance of flood simulations in this study.

While noticing the large spatial heterogeneity among basins, we applied LSTM at each selected station (at daily time scale) during 1979–2010 (according to the validation of meteorological data). We used basin-averaged daily mean air temperature, precipitation, wind speed and the simulated daily discharge (corresponding to the observed gauging stations selected) from the simulations of GHMs-CaMa-Flood model chain as the inputs of LSTM, and used observed daily discharge as the output value. Specifically, we chose simulations from VIC-CaMa-Flood model chain as the input of LSTM in order to simplify the calculations after noticing the similar performance of the five GHMs (in section 3.2). Considering a typical e-folding time scale (recession time) of streamflow, we set the time lag at 30 d (both forward and backward). After that, we selected the first half of the time series at each station as training samples (including several complete years) (as those need to be continuous unlike regular neural networks). The last half time series was then used as validation samples. For example, suppose the station has complete observations in 1981–2000, the observations in 1981–1990 would be used as training samples, while the observations in 1991–2000 would be used as testing samples.

Moreover, observed daily streamflow data were downloaded from the Global Runoff Data Center (GRDC, 56068 Koblenz, Germany) and U.S. Geological Survey (USGS, https://waterdata.usgs.gov/nwis) streamflow data. Following the criteria used for GSIM in section 2.1, 2110 stations were selected. For a better spatial coverage, a minimum of 5-year coverage instead of 10-year coverage was adopted in this section. The basins’ daily areal precipitation was derived from Multi-Source Weighted-Ensemble Precipitation (MSWEP, [41]), and the basin average daily mean air temperature and wind speed were derived from ERA-interim dataset [42], available from www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim).
3. Results and discussion

3.1. The performance of GHMs-based flood simulations across basins

To conduct a comprehensive evaluation of GHMs-based flood simulations, we analyzed the performances of these GHMs through four dimensions: the five flood-relevant indices, selected basins, (Köppen–Geiger) climate classifications and human impact conditions (socio-econ-scenarios) as applied in the GHMs. By comparing the selected basins under VARSOC conditions (figure 1), the simulations of flood-relevant indices from the GHMs-CaMa-Flood model chain (multi-model mean Pearson correlation coefficients) varied greatly across the globe. Clearly, global universal parameterization cannot fit all basins due to their varied hydrological, ecological and hydraulic characteristics [43]. In general, the flood conditions...
simulations perform reasonably well between 30°N and 30°S, but rather poor beyond 60°N. This might be related to the relatively low spatial resolution of terrain used beyond 60°N for generating routing direction within the CaMa-Flood model, constraining the accuracy of simulations in boreal regions [12]. Moreover, the low Pearson correlation coefficients in basins with high latitude and elevation (for day of year with maximum discharge in figures S4 and S5) indicate difficulties in reproducing snowmelt and soil freezing could be other sources of errors [5].

The simulations of peak discharge timing showed relatively poor performance than peak discharge amplitude. The mean PCCs derived from all selected stations are 0.41, 0.45, 0.53, 0.24 and 0.25 for maximum discharge, maximum 7 d averaged discharge, 90-percentile maximum discharge, day of year (1–366) with maximum discharge and day of year (1–366) with maximum 7 d averaged discharge, respectively. Overall, the simulation of peak discharge was unsatisfactory from GHMs-CaMa-Flood model chain.

3.2. Evaluation of GHMs-based flood simulation under different climate zones

We then examined GHMs-based flood simulations under different climate zones (Köppen–Geiger climate classification) and socio-economic scenarios (VAR-SOC, NONSOC and PENSSOC). The results showed great discrepancies between models under different socio-climate zones (figure 2). In basins located in BWh (arid hot desert) and DWa (cold region with dry winter and hot summer) climate zoned, all GHMs performed satisfactorily except for the DBH model. In basins located in DWd (cold region with dry and very cold winter) climate zone, the LPJML model demonstrated poor performance. Most models exhibited poor performance in basins located in DWb (cold region with dry winter and warm summer) climate zone except for the H08 and PCR-FLOBWB models. The VIC model showed the best performance in basins located in Dsb (cold region with dry and warm summer) and Dfd (cold region with very cold winter but dry season) climate zones. Overall, the performances of the five GHMs showed great discrepancy across climate zones and flood-relevant indices [5], but exhibited small difference among various socio-economic scenarios (figure S1).

Taylor diagram is a useful tool for multi-model evaluation [44]. Based on Taylor diagram, we compared the performances of five GHMs in terms of five multi-station-averaged flood-relevant indices (section 2.2). For annual maximum discharge, maximum 7 d averaged and annual 90-percentile maximum discharge, the PCR-GLOBWB and VIC models showed relatively good performances. For the days of maximum and 7 d averaged discharge, the DBH, PCR-GLOBWB and VIC models experienced better performance. The GHMs performances varied considerably across river basins and flood-relevant indices, due to their different process representation and spatially generalized parameters [45]. The large inter-model differences in seasonal variability might be due to different process

Figure 2. Pearson correlation coefficients of five flood-relevant indices under five GHMs (DBH, H08, LPJML, PCR-GLOBWB and VIC) and different climate zones (based on Köppen–Geiger climate classification).
In general, the uncertainties of GHMs-based flood simulations might come from (1) the biases in climate forcing data as inputs of the GHMs, (2) the errors in describing hydrological processes (representation of hydrological process and discrepancy of models with real hydrological process) at the scale of grid cell (3) the limitation of river routing approach such as model parameters related to physical routing process [10, 11, 46–48]. The large variations in GHMs’ performance in different basins, climate zones and various flood-relevant indices, call for a careful selection of GHMs and improvement in techniques adopted for future flood simulations.

3.3. Improvement of GHMs-based flood simulations using machine learning

Given the temporal lag between basin-scale peak discharge and historical input data (precipitation, temperature) [49], we chose an LSTM to improve the performance of GHMs-based streamflow simulation. We used basin-averaged daily mean air temperature, precipitation, wind speed and simulated daily discharge (corresponding to the selected gauging stations) from VIC-CaMa-Flood model chain as the inputs of LSTM.

Subsequently, we compared the Nash-Sutcliffe efficiency (NSE) and RE of the VIC-driven CaMa-Flood streamflow simulations and LSTM outputs against the daily discharge observations. The NSE and RE of daily streamflow (figure 4) suggested significant improvements for all basins after using the physics-guided LSTM, which dramatically improved the NSE and RE for all stations (median value) from −0.54 to 0.49 and from −14.39 to −7.52, respectively.

Apparently, the VIC-CaMa-Flood model chain (figure 5) overestimated maximum discharge, especially for the lower amplitudes zones. In general, the timing points of the peak discharge were dispersive from observation section of DOYMAX and DOYMAX7 in figure 5, which indicates the poor performance of VIC-CaMa-Flood model chain in the timing of peak discharge. Liu et al [49] and Zhao et al [12] also highlighted the poor performance of GHMs in the timing of peak discharge. The combination of LSTM and VIC-CaMa-Flood model chain (PGNN) drastically improved the accuracies of the five flood-relevant indices. Specifically, the physics-guided LSTM improved $R^2$ from 0.27 to 0.95 from model simulations to LSTM cooperation for maximum annual discharge, 0.3–0.98 for annual maximum 7 d averaged discharge, 0.29–0.99 for annual 90-percentile maximum discharge, 0.17–0.24 for day of year with maximum discharge and 0.19–0.27 for day of year with maximum 7 d averaged discharge.

Yet, beside these dramatic improvements, LSTM could not significantly improve the timing of peak discharge.

Figure 3. Taylor diagrams for comparing five typical GHMs (DBH, H08, LPJML, PCR-GLOBWB and VIC) and observations. Figures 3(a)–(e) indicate annual maximum discharge (MAX), annual maximum 7 d averaged discharge (MAX7), annual 90-percentile maximum discharge (P90), day of year (1–366) with maximum discharge (DOYMAX); day of year (1–366) with maximum 7 d averaged discharge (DOYMAX7).
discharge, even though with a time lag set to 30 d (both forward and backward). This might be because of the sensitivity of peak discharge to extreme precipitation, snowmelt (or soil freezing) and soil moisture excess [50], which was mistaken in the simulations cannot entirely be corrected by the LSTMs. Moreover, the biases in the climate forcing data, the representation of real hydrological process and the limitation of routing could further limit the capacity to accurately predict correct timing [10, 11, 46, 51]. Thus, the improvement of timing might depend more on the traditional GHMs-CaMa-Flood model chain than tools like LSTM used in this study, e.g. the higher temporal/spatial resolution for both hydrological and routing process, higher-precision forcing data.

Based on meteorological data (precipitation, temperature, wind speed, day of year used in this study), we compared the performances of physics-guided LSTM and pure LSTM. The results (figure S2) showed obvious better values in \( R^2 \) for physics-guided LSTM (used as PGNN here) than pure LSTM for all five selected indices, such as annual maximum discharge, \( R^2 \) is 0.95 for PGNN, higher than 0.86 for a pure neural network based on meteorological data. To test the performance of meteorological data and simulated discharge in LSTM, we conducted a new LSTM model using pure simulated discharge as input (figure S3 in supplementary material). Compared with figure S2 (with pure meteorological data as input), the higher \( R^2 \) with pure discharge implied the dominant role of simulated discharge in the performance of flood simulations within the physics-guided LSTM. These results indicate that the PGNN used in this study are improved and strongly rely on both the VIC-CaMa-Flood and meteorological data, while the combination of LSTM and VIC-CaMa-Flood gave the best performance in flood simulations (with the highest \( R^2 \)).

Our analyzes showed that the LSTM can correct the streamflow simulation biases through harvesting existing data. It could be a powerful tool for more robust and confidential flood risk assessment after combine with present discharge observations. Moreover, the combination of LSTM and GHMs-CaMa-Flood model chain generated in this study might serve as a useful tool for projecting future flood disasters under climate change.

4. Summary and conclusions

In this study, we performed a global evaluation of flood simulations from five GHMs against flood-relevant indices calculated from GSIM dataset. We made the first attempt to improve GHMs-based flood simulations using a hybrid physics-guided machine learning approach (LSTM). The following conclusions were drawn:

(1) The hybrid physically based neural network used in this study demonstrated excellent performance
in global flood simulation, which implies that the combination of classical flood simulations and machine learning techniques might be a powerful tool for more robust and confident flood risk assessment.

(2) The large variations in GHMs’ performances in different basins, climate zones and for various flood-relevant indices, call for a careful selection of GHMs and improvement techniques adopted for future flood simulations.

(3) Both GHMs and the combination with LSTM could not optimally simulate the timing of peak discharge. This calls for improvements of flood timing in simulation such as routing and snowmelt processes.

Acknowledgments

This study was financially supported by the National Research and Development Program.
of China (2019YFA0606903, 2016YFC0401401 and 2016YFA0602402), the National Natural Sciences Foundation of China (41601035), the Key Research Program of the Chinese Academy of Sciences (ZDRW-ZS-2017-3-1 and ZDRW-ZS-2019-3) and the Top-Notch Young Talents Program of China (Fubao Sun). We appreciate the editor and two anonymous reviewers for their constructive and insightful comments.

Data availability statement

This work has been conducted under the framework of phase 2 of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2a: www.isimip.org). The daily runoff data from the GHMs can be downloaded from https://esg.pik-potsdam.de/search/isimip/under ISIMIP2a. The observed discharge data is available from the Global Runoff Data Center (GRDC, 56068 Koblenz, Germany, www.bafg.de/GRCEN/Home/homepage_node.html) and U.S. Geological Survey (USGS, https://waterdata.usgs.gov/nwis), streamflow indices from GSIP (https://doi.org/10.1594/PANGAEA.887470), precipitation data from MSWEP (https://doi.org/10.1175/BAMS-D-17-0138.1), air temperature and wind speed data from ERA-interim group (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim). The codes and results used in this study are available from the first author upon request (yangt.16b@igsrr.ac.cn).

ORCID iDs

Tao Yang https://orcid.org/0000-0002-7662-4798
Pierre Gentine https://orcid.org/0000-0002-0845-8345
Wenbin Liu https://orcid.org/0000-0002-9569-6762

References

[1] Munich Re 2016 Munich Re Insurance Company Geo Risks Research NatCatSERVICE Database (www.munichre.com/en/insurance/business/non-life/natcatservice/index.html) (Accessed: 9 May 2018)
[2] Arnell N W and Gosling S N 2016 The impacts of climate change on river flood risk at the global scale Clim. Change 134 387–401
[3] Yin J, Gentine P, Zhou S, Sullivan S C, Wang R, Zhang Y and Guo S 2018 Large increase in global storm runoff extremes driven by climate and anthropogenic changes Nat. Commun. 9 4389
[4] Long D, Yang Y, Wada Y, Hong Y, Liang W, Chen Y, Yong B, Hou A, Wei J and Chen L 2015 Deriving scaling factors using a global hydrological model to restore GRACE total water storage changes for China’s Yangtze river basin Remote Sens. Environ. 168 177–93
[5] Masaki Y, Hanasaki N, Biemann H, Schmied H M, Tang Q, Wada Y, Gosling S N, Takahashi K and Hijiioka Y 2017 Intercomparison of global river discharge simulations focusing on dam operation—multiple models in two case-study river basins, Missouri-Mississippi and Green–Colorado Environ. Res. Lett. 12 055002
[6] Schumacher M, Forootan E, van Dijk A J M, Müller Schmied H, Crossbie R S, Kusche J and Doll P 2018 Improving drought simulations within the Murray–Darling Basin by combined calibration/assimilation of GRACE data into the WaterGAP Global Hydrology Model Remote Sens. Environ. 204 212–28
[7] Gudmundsson L, Leonard M, Do H X, Westra S and Seneviratne S I 2019 Observed trends in global indicators of mean and extreme streamflow Geophys. Res. Lett. 46 756–66
[8] Liu W and Sun F 2019 Increased adversely-affected population from water shortage below normal conditions in China with anthropogenic warming Sci. Bull. 64 567–9
[9] Liu W, Lim W H, Sun F, Mitchell D, Wang H, Chen D, Bethke I, Shiogama H and Fischer E 2018 Global freshwater availability below normal conditions and population impact under 1.5 and 2 °C stabilization scenarios Geophys. Res. Lett. 45 9803–13
[10] Veldkamp T I E et al 2018 Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study Environ. Res. Lett. 13 035008
[11] Willner S N, Levermann A, Zhao F and Frieler K 2018 Adaptation prepared to preserve future high-end river flood risk at present levels Sci. Adv. 4 1–9
[12] Zhao F et al 2017 The critical role of the routing scheme in simulating peak river discharge in global hydrological models Environ. Res. Lett. 12 075003
[13] Alcamo J and Gallopin G 2009 Building a 2nd generation of water scarcity projections Water Policy 11 159–79
[14] IPCC 2013 Climate Change 2013: The Physical Science Basis (Geneva: IPCC)
[15] UNEP 2007 Global Environment Outlook 4: Environment for Development (Valletta: Progress Press) (http://hdl.handle.net/20.500.11822/7646)
[16] Lim W H, Yamazaki D, Koira S, Hirabayashi Y, Kanae S, Dadson S J, Hall J W and Sun F 2018 Long-term changes in global socioeconomic benefits of flood defences and residual risk based on CMIP5 climate models Earth’s Future 6 938–54
[17] Kühnlein M, Appelhans T, Thies B and Nauss T 2014 Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—a random forests-based approach applied to MSG SEVIRI Remote Sens. Environ. 141 129–43
[18] Requena-Mesa C, Reichstein M, Mahecha M, Kraft B and Denzler J 2018 Predicting landscapes as seen from space from environmental conditions Int. Geosci. Remote Sens. Symp. (IGARSS) 2018 1768–71
[19] Zhao W and Du S 2016 Learning multiscale and deep representations for classifying remotely sensed imagery ISPRS J. Photogramm. Remote Sens. 113 153–63
[20] Pathak J, Wikner A, Fussell R, Chandra S, Hunt R B, Girvan M and Ott E 2018 Hybrid forecasting of chaotic processes: using machine learning in conjunction with a knowledge-based model Chaos 28 041101
[21] Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N and Prabhat 2019 Deep learning and process understanding for data-driven Earth system science Nature 566 195–204
[22] Zhang D, Holland E S, Lindholm G and Ratnaweera H 2018 Hydraulic modeling and deep learning based flow forecasting for optimizing inter catchment wastewater transfer J. Hydrol. 567 792–802
[23] Zhu X, Zhang Q, Xu C Y, Sun P and Hu P 2019 Reconstruction of high spatial resolution surface air temperature data across China: a new geo-intelligent multisource data-based machine learning technique Sci. Total Environ. 665 300–13
[24] Gentine P, Pritchard M, Rasp S, Reinaud G and Yacalis G 2018 Could machine learning break the convection parameterization deadlock? Geophys. Res. Lett. 45 5742–51
[25] Yang T, Asanjan A A, Welles E, Gao X, Sorooshian S and Liu X 2017 Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information Water Resour. Res. 53 2796–812
[26] Karpaterne A, Watkins W, Read J and Kumar V 2017 Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling (arXiv:1710.11431)

[27] Yamazaki D, De Almeida G A M and Bates P D 2013 Improving computational efficiency in global river models by implementing the local inertial flow equation and a vector-based river network map Water Resour. Res. 49 7221–35

[28] Yamazaki D, Kanae S, Kim H and Oki T 2011 A physically based description of floodplain inundation dynamics in a global river routing model Water Resour. Res. 47 1–21

[29] Tang Q, Oki T, Kanai S and Hu H 2007 The influence of precipitation variability and partial irrigation within grid cells on a hydrological simulation J. Hydrometeorol. 8 499–512

[30] Hanasaki N, Kanae S, Oki T, Masuda K, Motoya K, Shirakawa N, Shen Y and Tanaka K 2008b An integrated model for the assessment of global water resources: I. Model description and input meteorological forcing Hydrolog. Earth Syst. Sci. 12 1007–25

[31] Hanasaki N, Kanae S, Oki T, Masuda K, Motoya K, Shirakawa N, Shen Y and Tanaka K 2008b An integrated model for the assessment of global water resources: II. Applications and assessments Hydrolog. Earth Syst. Sci. 12 1027–37

[32] Rost S, Gerten D, Bondeau A, Lucht W, Rohwer J and Schaphoff S 2008 Agricultural green and blue water consumption and its influence on the global water system Water Resour. Res. 44 1–17

[33] Wada Y, van Beek L P H, van Kempen C M, Beckman J W T M, Vasak S and Bierkens M F P 2010 Global depletion of groundwater resources Geophys. Res. Lett. 37 L20002

[34] van Beek L P H, Wada Y and Bierkens M F P 2011 Global monthly water stress: I. Water balance and water availability Water Resour. Res. 47 W07517

[35] Liang X, Lettenmaier D P, Wood E F and Burges S J 1994 A simple hydrologically based model of land surface water and energy fluxes for general circulation models J. Geophys. Res. 99 14415

[36] Hu X, Hall J W, Shi P and Lim W H 2016 The spatial exposure of the Chinese infrastructure system to flooding and drought hazards Nat. Hazards 80 1083–1118

[37] Gumudsson L, Do H X, Leonard M and Westra S 2018 The global streamflow indices and metadata archive (GSIM): II. Quality control, time-series indices and homogeneity assessment Earth Syst. Sci. Data 10 787–804

[38] Do H X, Gumudsson L, Leonard M and Westra S 2018 The global streamflow indices and metadata archive (GSIM): I. The production of a daily streamflow archive and metadata Earth Syst. Sci. Data 10 765–85

[39] Wang H, Sun F, Xia J and Liu W 2017 Impact of LUCC on streamflow based on the SWAT model over the Wei River basin on the Loess Plateau in China Hydrol. Earth Syst. Sci. 21 1–17

[40] Hrechitre S and Schmidhuber J 1997 Long short-term memory Neural Comput. 9 1735–80

[41] Beck H E, Wood E F, Pan M, Fisher C K, Miralles D G, van Dijk A J M, McVicar T R and Adler R F 2019 MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment Bull. Am. Meteorol. Soc. 100 475–500

[42] Dee D P et al 2011 The ERA-interim reanalysis: configuration and performance of the data assimilation system Q. J. R. Meteorol. Soc. 137 553–97

[43] Jung H C, Hamski J, Durand M, Alsdorf D, Hossain F, Lee H, Hossain A K M A, Hasan K, Khan A S and Hoque A K M Z 2010 Characterization of complex fluvial systems using remote sensing of spatial and temporal water level variations in the Amazon, Congo, and Brahmaputra Rivers Earth Surf. Process. Landf. 35 294–304

[44] Liu W and Sun F 2016 Assessing estimates of evaporative demand in climate models using observed pan evaporation over China J. Geophys. Res. Atm. 121 8329–49

[45] Zehnerpour J et al 2019 Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models Environ. Model. Softw. 114 112–28

[46] Hattermann F F et al 2018 Sources of uncertainty in hydrological climate impact assessment: a cross-scale study Environ. Res. Lett. 13 015006

[47] Wang T, Sun F, Lim W H, Wang H, Liu W and Liu C 2018 The predictability of annual evapotranspiration and runoff in humid and nonhumid catchments over china: comparison and quantification J. Hydrometeorol. 19 533–45

[48] Zillani M G, Ghiotline R, Ait-El-Fquih B, McCabe M F and Hotteit I 2019 Enhanced flood forecasting through ensemble data assimilation and joint state-parameter estimation J. Hydrol. 577 123924

[49] Liu X, Tang Q, Cui H, Mu M, Gerten D, Gosling S N, Masaki Y, Sato0 Y and Wada Y 2017 Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations Environ. Res. Lett. 12 025009

[50] Berghuijs W R, Harrigan S, Molnar P, Slater L J and Kirchner J W 2019 The relative importance of different flood-generating mechanisms across Europe Water Resour. Res. 55 1–12

[51] Bloeschl G, Hall J, Papaik J, Perdigão R A P, Merz B, Arheimer B and Chirico G B 2017 Changing climate shifts timing of European floods Science 357 588–90