ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Evaluation of river flood extent simulated with multiple global hydrological models and climate forcings

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
17 February 2021

REVISED
16 July 2021

ACCEPTED FOR PUBLICATION
28 July 2021

PUBLISHED
13 August 2021

LETTER

Evaluation of river flood extent simulated with multiple global hydrological models and climate forcings

Benedikt Mester[1,2,*], Sven Norman Willner[1], Katja Frieler[1] and Jacob Schewe[1]

[1] Potsdam Institute for Climate Impact Research, Potsdam, Germany
[2] Institute of Environmental Science and Geography, Potsdam University, Potsdam, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: benedikt.mester@pik-potsdam.de

Keywords: global flood model, validation, model intercomparison, flood risk, global hydrological model

Supplementary material for this article is available online

Abstract

Global flood models (GFMs) are increasingly being used to estimate global-scale societal and economic risks of river flooding. Recent validation studies have highlighted substantial differences in performance between GFMs and between validation sites. However, it has not been systematically quantified to what extent the choice of the underlying climate forcing and global hydrological model (GHM) influence flood model performance. Here, we investigate this sensitivity by comparing simulated flood extent to satellite imagery of past flood events, for an ensemble of three climate reanalyses and 11 GHMs. We study eight historical flood events spread over four continents and various climate zones. For most regions, the simulated inundation extent is relatively insensitive to the choice of GHM. For some events, however, individual GHMs lead to much lower agreement with observations than the others, mostly resulting from an overestimation of inundated areas. Two of the climate forcings show very similar results, while with the third, differences between GHMs become more pronounced. We further show that when flood protection standards are accounted for, many models underestimate flood extent, pointing to deficiencies in their flood frequency distribution. Our study guides future applications of these models, and highlights regions and models where targeted improvements might yield the largest performance gains.

1. Introduction

Of all natural disasters worldwide, fluvial (river) flooding is among the most frequent and devastating hazards (Jha *et al* 2012). In the recent years of 2010–2018, it caused 115 million human displacements (IDMC 2019), 49 595 fatalities, and US$ 360 billion in economic losses (Munich Re 2020). For example, 11 million displacements (IDMC 2019), 1985 deaths, and US$ 9.5 billion in economic losses (EM-DAT 2020) were recorded in the aftermath of the Pakistan floods in 2010. Flooding killed 6054 people in India in 2013 (EM-DAT 2020) and caused an estimated US$ 33 billion losses in China in 2016, with only 2% of losses insured (Floodlist 2016). Beyond these records, one can expect further losses such as of cultural heritage and ecosystem services, which are, however, difficult to assess (Hurlbert 2018).

Continental-scale changes in flood discharge have been observed recently, in line with theoretical expectations about the effects of global warming on the hydrological cycle (IPCC 2014, Blöschl *et al* 2019). This poses the question to what extent the societal impacts of floods have already been shaped by anthropogenic climate change. However, displacements, damages and losses associated with floods are a function not only of the physical flood hazard, but also of socioeconomic factors. The latter, in particular, determine exposure—the number of people or the value of assets potentially affected by flooding—, and vulnerability—the susceptibility of exposed elements to the hazard (IPCC 2012, Jongman *et al* 2015). Together, these controlling risk factors form a dynamic, spatially and temporally variable balance used for risk assessment. Since not all three factors are generally known, it is challenging to quantify their relative contributions to the ultimate impacts of historical floods.

Global flood models (GFMs) can be used to estimate historical flood extents based on observed weather, and could thereby provide the physical flood hazard component for such assessments when direct observations of flood extent are lacking. This approach is increasingly being used, for instance, to estimate past changes in vulnerability (Tanoue *et al* 2016) or attribute trends in reported flood-induced damages (Sauer *et al* 2021). However, the degree to which such studies can explain the observed variations in damages and affected population varies substantially, and can be fairly low for many parts of the world. It is unclear to what extent this is due to shortcomings in the simulated flood hazard, exposure, or assumptions about vulnerability.

This highlights that a thorough understanding of the reliability of global flood hazard estimates is important. However, validation and benchmarking studies are rare (Hoch and Trigg 2019), which is mainly due to the scarce availability of reference flood maps outside of some high-income countries and regions such as the European Union, North America, or Australia (Dottori *et al* 2016). A comparison of several GFMs to satellite imagery in three African river sections showed considerable differences between models in terms of how accurately they reproduced observed flood extent; most models both missed flooding in some areas and falsely simulated flooding in others (Bernhofen *et al* 2018). The agreement between these models in simulated flood extent was shown to be only 30%–40%, with considerable differences in hazard magnitude and spatial patterns (Trigg *et al* 2016).

It is hardly known to what extent such differences between GFM simulations and their predictive capacities are related to the GFMs themselves, for instance, due to differences in model structure or the underlying digital elevation models; and to what extent they are related to the boundary conditions used to force the GFMs. Depending on the modelling framework, these boundary conditions consist either of gauged river flow datasets, or—for the majority of GFMs—of gridded runoff estimates from global hydrological model (GHM) or land-surface models (Trigg *et al* 2016). Those runoff simulations in turn need meteorological variables as input, which come from global climate reanalyses or climate models. Most global flood hazard simulations thus are the result of a cascade of different models and data products, with multiple options available at each step in the cascade. The influence of choices in the upstream steps of this cascade on the resulting flood extent estimate has hardly been systematically investigated (Zhou *et al* 2021).

In this study we address this research gap. We run a state-of-the-art GFM with runoff forcing from 11 different GHMs, each in turn forced by three different climate reanalyses. We evaluate the resulting simulation ensemble against satellite-derived flood extent observations for eight recent large flood events on four continents, covering different climatic and hydraulic environments, and assess the influence of the choice of both climate forcing and GHM on the performance of the GFM simulations. We do this under different assumptions about flood protection, to also assess the realism of simulated return intervals.

2. Data and methodology

2.1. Models

We use the GFM CaMa-Flood (Yamazaki *et al* 2011), driven by an ensemble of 11 GHMs and three gridded climate forcing datasets, leading to 33 combinations in total. The climate forcing datasets used to drive the GHMs are the Princeton Global Forcing data set version 2 (PGFv2) (Sheffield *et al* 2006), the Global Soil Wetness Project phase 3 forcing data set (GSWP3) (Hyungjun 2014) and the WATCH forcing data methodology applied to ERA-Interim reanalysis data (WFDEI) (Weedon *et al* 2011, 2014). All three datasets are based on reanalysis products (ERA-Interim for WFDEI; 20CR for GSWP3; NCEP/NCAR for PGFv2) that assimilate information from local weather stations, and subsequently apply corrections to the precipitation data and other variables using station-based observational data; two datasets (WFDEI and GSWP3) also correct for precipitation undercatch by rain gauges. Given these methodologies, and the gridded nature of the forcing products, direct comparison with local station data is not straightforward, but existing validation exercises show reasonable agreement with station data as well as with gridded observational datasets (Weedon *et al* 2014, Essou *et al* 2017).

The set of GHMs comprises CLM4.0 (Leng *et al* 2015), DBH (Tang *et al* 2007), H08 (Hanasaki *et al* 2008), JULES-W1 (Best *et al* 2011), LPJmL (Sitch *et al* 2003), MATSIRO (Pokhrel *et al* 2014), MPI-HM (Stacke and Hagemann 2012), ORCHIDEE (Traore *et al* 2014), PCR-GLOBWB (Wada *et al* 2014), VIC (Liang *et al* 1994) and WaterGAP2 (Müller Schmied *et al* 2016). An overview of the GHMs' main characteristics, e.g. evaporation and runoff schemes, is available in the supplementary material—table S1. All GHM simulations follow a common protocol (ISIMIP2a, www.isimip.org) to ensure a standardized input scheme for CaMa-Flood. Simulations are performed under naturalized conditions, i.e. storage in man-made reservoirs or agricultural water withdrawal are not included.

The runoff of the respective GHM then constitutes the input for CaMa-Flood v3.6.2 which yields discharge as well as flood depth on a 0.25° resolution grid. The underlying river network in CaMa-Flood has been derived by the model author based on the flow direction maps HydroSHEDS (Lehner *et al* 2008) and GDBD (Masutomi *et al* 2009) as well as the digital elevation model SRTM3 (Farr *et al* 2007) using their FLOW method (Yamazaki *et al* 2009). Using these same data, we downscale flood depth to 18 arc sec. i.e. the daily flood volume in a low-resolution (0.25°) grid cell is distributed onto the underlying high-resolution (18 arc sec) grid cells according to their elevation. We then assign to each high-resolution grid cell the annual maximum daily value, resulting in an annual flood depth timeseries. The event duration according to the satellite imagery matches with the rising limb or the peak of the flood simulations for most regions of interest; and coincides with no second flood event in the year of investigation, which legitimizes this approach (figures S22–S33 in the supplementary material (available online at stacks.iop.org/ERL/16/094010/mmedia)). This dataset is used to produce figure 2. Finally, we calculate the flooded fraction on an intermediate-resolution (2.5 arc min) grid, i.e. the fraction of flooded high-resolution grid cells within each intermediate-resolution grid cell. This flood fraction dataset is used to calculate the performance scores (see below).

Whereas this constitutes our default simulation setup ('default'), we also assess setups assuming protection against floods with an average recurrence interval (ARI) of 2 years ('protect 2y' setup), and assuming flood protection standards according to the FLOPROS database (Scussolini *et al* 2016) ('protect FLOPROS' setup). FLOPROS incorporates modelled protection, infrastructure, and policy measures in a best estimate ('merged' layer) on a sub-national level. Fitting, for each climate forcing and GHM, a generalized extreme value (GEV) distribution to the annual maximum discharge for each cell and in the simulation period available for all models (1971–2010), we obtain the return period in dependence of discharge. We then compare the return period for each studied event to the protection level for the respective cell; i.e. either 2 years, or the protection level given by FLOPROS. In the 'protect' settings, we thereby only account for flood events in cells in which this protection level is exceeded and assume no flooding for events with lower return period.

It should be noted that there are well-established practices used in floodplain planning processes internationally (World Meteorological Organization 2009) that do not rely on GFMs but instead use more complex, locally calibrated hydrodynamic models (Raadgever and Hegger 2018). However, these techniques require elaborate calibration for each individual catchment (Canning and Walton 2014), and rely on local observational data that is not commonly available in all parts of the world. For instance, in a new, comprehensive global streamflow database (Do *et al* 2018, Gudmundsson *et al* 2018), local gauge records are available for only one of the events studied in this paper (the 2010 flood in Dalby, Australia; see below), and are entirely unavailable for some of the study regions. Thus, while the global models evaluated here are likely inferior to more complex, locally-informed flood prediction models where those exist, the global models nonetheless are important tools widely used in continental- or global-scale applications (Bates *et al* 2021).

2.2. Observational data

For the comparison of our simulated flood extent we use satellite imagery from the archive of the Dartmouth Flood Observatory (DFO), which is based on NASA MODIS satellite sensors (https://floodobservatory.colorado.edu/) (Brakenridge 2006), and from the UNOSAT Flood Portal (UFP) providing flood extent maps derived from a variety of satellite sensors (http://floods.unosat.org/geoportal/catalog/main/home.page). The number of eligible events is limited, because consistent geospatial imagery starts in 2010 for DFO and 2006 for UFP, respectively, and most the climate reanalysis products used here extend only until 2010 or (for PGFv2) 2012. Only large-scale disasters with a large river size are taken into account to ensure that the inundated areas can be adequately captured given the spatial resolution of the GFM. It is essential that observational validation data is available, consistent and comprehensive for the entire area of interest. Additionally, a spread across different climate zones and continents is desirable for a comprehensive global comparison study (Dottori *et al* 2016). We exclude flash flood events, storm surge flooding, as well as floods caused by mismanagement or failure of man-made structures, since these types of floods cannot be modelled by the GFM.

We identify ten regions of particular interest in the context of eight major flood events, as shown in figure 1. The following regions, named after the central city or town affected, are used for validation: Huainan in China (year of flood event: 2007), Sayaxché in Guatemala (2008; the westernmost part is located in Mexico), Trinidad in Bolivia (2008), Alipur and Ghotki in Pakistan (2010), Phimai in Thailand (2010), Dalby in Australia (2010), Chemba in Mozambique (2007), as well as Lokoja and Idah in Nigeria (2012). Chemba (MOZ; we will indicate each location's ISO3 country code throughout the paper for ease of reference), Lokoja (NGA), and Idah (NGA) are studied in a recent GFM intercomparison study, thus facilitating comparison of our results with that study (Bernhofen *et al* 2018). The selected areas are located in monsoon climates, tropical savannas and rainforests, subtropical climates, and deserts, on four continents (Peel *et al* 2007). The selection covers a variety of hydraulic characteristics, ranging from confined watercourses of the Niger River and the anabranching Condamine River in Queensland to highly braided and slow sections of the Indus River. A more detailed description of the chosen events and the river hydraulics can be found in the supplementary material—table S2.

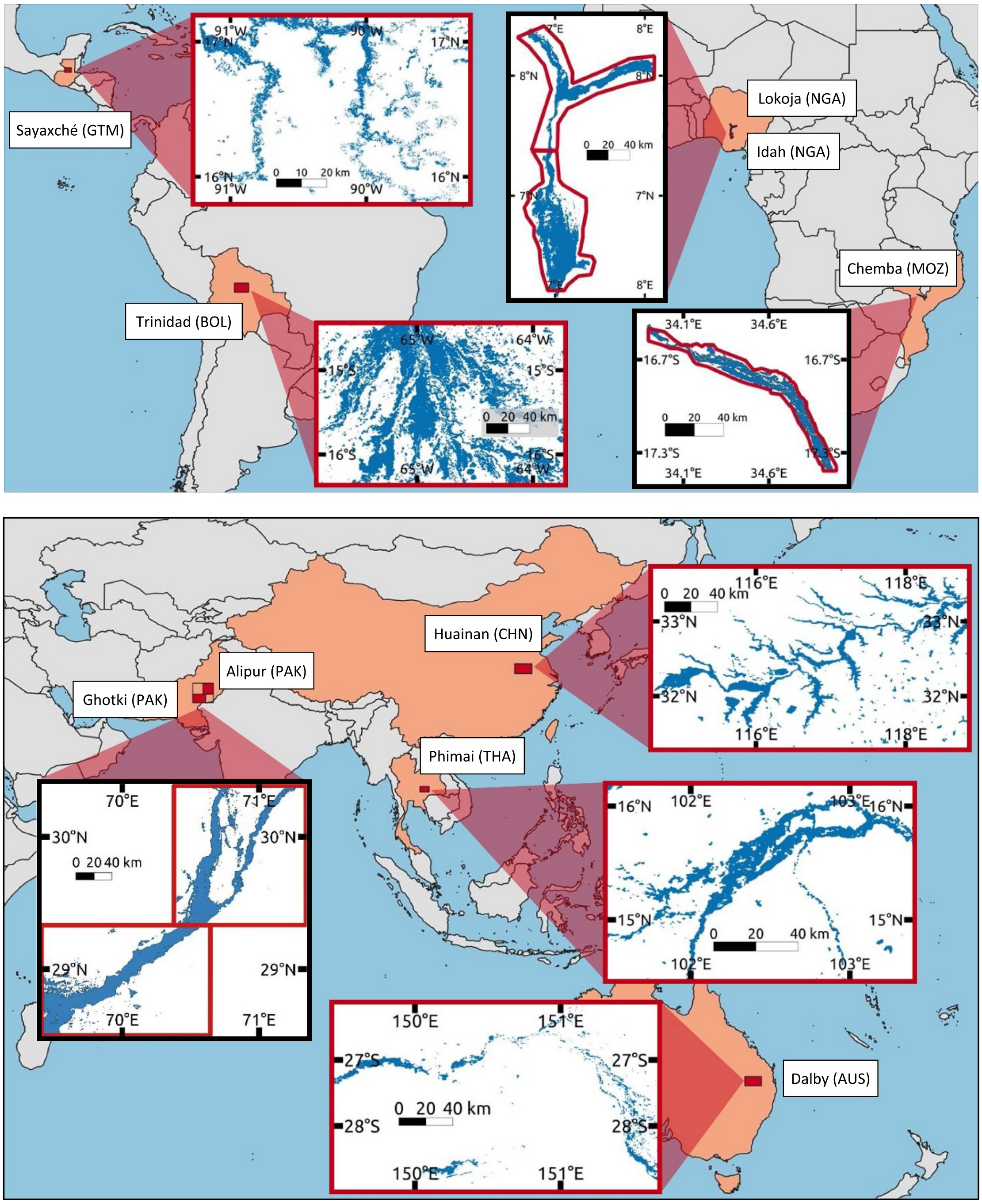


Figure 1. Satellite imagery of flood extent (blue) in the study areas (red). Note: the outlines of Chemba (MOZ), Lokoja (NGA) and Idah (NGA) follow the flood shape and are not rectangular. Country shapes reproduced from https://gadm.org/data.html.

The satellite imagery of flood extents for the regions Alipur (PAK), Ghotki (PAK), Phimai (THA), and Dalby (AUS) are taken from the DFO. Satellite imagery for Huainan (CHN), Sayaxché (GTM) and Trinidad (BOL) are downloaded from the UFP. For some events, the data consists of several days of imagery and is hence merged into one maximum flood extent per event. Rectangular analysis regions are defined for the events above (red rectangles in figure 1). Flood footprints and outline data for the regions Chemba (MOZ), Lokoja (NGA) and Idah (NGA) are derived from Research Data Leeds (http://archive.researchdata.leeds.acuk/411/), which is intended for GFM validation (Bernhofen *et al* 2018) and also based on the DFO archive. In contrast to the other regions, data for these regions is only available inside an irregularly shaped polygon roughly outlining the main inundation area, which limits the analysis of both observations and simulations to these polygons (we will discuss the implications of this below). The satellite imagery for all regions is at a 209 m resolution. Only the simulations of eight

GHMs forced with PGFv2 extend to the year 2012, which limits the analysis of Lokoja (NGA) and Idah (NGA).

2.3. Analysis

The analysis procedure starts with a visual comparison of the simulated and the observed flood extent. For the former, we use the gridded CaMa-Flood flood depth output downscaled to 18 arc sec resolution (approx. 550 m at the equator). This yields a binary grid with two grid cell states (flooded or not flooded) for each climate forcing and GHM. For each climate forcing and grid cell we then count the number of GHMs showing flooding in the respective cell, to create a model agreement map. We use the flood outlines for each region (red outlines in figure 1) to mask both the model agreement map and the satellite imagery. The masked images are then superimposed onto each other (figure 2). Next, in order to quantify model performance, we use two different performance metrics (Bates and de Roo 2000, Aronica *et al* 2002, Werner *et al* 2005, Bernhofen *et al* 2018). First, the critical success index (CSI) is defined as:

$$\mathrm{CSI} = \frac{F_\mathrm{m} \cap F_\mathrm{o}}{F_\mathrm{m} \cup F_\mathrm{o}}$$

where F_m is the modelled flooded area by CaMa-Flood and F_o is the observed flooded area by the satellite imagery. $F_\mathrm{m} \cap F_\mathrm{o}$ is the intersection area between modelled and observed flood extent, i.e. the area correctly simulated as flooded by the model; and $F_\mathrm{m} \cup F_\mathrm{o}$ is the union area between modelled and observed flooded area. The CSI is perceived as the one of the most comprehensive scores (Bernhofen *et al* 2018). It ranges from 0 to 1, where 1 represents a perfect model 'fit' (Sampson *et al* 2015), and penalizes overprediction. The Bias score is defined as:

$$\mathrm{Bias} = \frac{(F_\mathrm{m} \cap F_\mathrm{o}) + F_\mathrm{m}}{(F_\mathrm{m} \cap F_\mathrm{o}) + F_\mathrm{o}} - 1.$$

An unbiased model has a Bias score of 0, positive and negative values indicate a tendency towards over- or under prediction of flood extents, respectively. The Bias score rewards a large intersection area between modelled and observed flood extent.

At high spatial resolution, mismatches in river geometry between the satellite imagery and the digital elevation models used in the GFM could deteriorate the performance scores in confined floodplains; e.g. if a river channel in the DEM is offset relative to its real location (Yamazaki *et al* 2011). Since we want to evaluate simulated flood extent per event, rather than DEM accuracy, we therefore calculate performance scores at the coarser resolution of 2.5 arc min. For that, we downsample both the binary satellite imagery and the CaMa-Flood binary flood data to 2.5 arc min using simple linear sampling, yielding the share of flooded area per cell. Incorporating absolute cell area we thus compute absolute flooded area per cell for both CaMa-Flood output and satellite imagery. Summing over all cells within an analysis region yields F_m and F_o respectively. The intersection area $F_\mathrm{m} \cap F_\mathrm{o}$ is calculated analogously but multiplying, in each cell, the smaller flooded fraction of either CaMa-Flood or satellite data with the cell area; for the union area $F_\mathrm{m} \cup F_\mathrm{o}$ the larger value of either model or data is used. This approach is based on the assumption that the location of flooding at the sub-grid scale is, for a given grid-scale flood extent, constrained by topography. For comparison, we also show CSI and Bias scores computed directly on 18 arc sec resolution in supplementary material figures S36 and S37.

In the section 3, along with the performance metrics for individual simulations and regions, we also show the median over all regions as well as median, minimum, maximum and spread over all hydrological models. Lokoja (NGA) and Idah (NGA) were excluded from the computation of the regional median, in order to allow a fair comparison of the median values between the three climate forcing datasets.

3. Results

We first analyse results for the default simulations not accounting for flood protection; and subsequently, in section 3.3, discuss the simulations with flood protection.

3.1. Model agreement map

Figure 2 displays the model agreement overview for all three climate forcings and ten regions, for the default simulations. Results differ substantially between the regions. The agreement between models is high, and the simulated flood outline in relatively good agreement with observations, for Sayaxché (GTM), Trinidad (BOL), Lokoja (NGA), Idah (NGA), and Phimai (THA), although the models miss some extended parts of the flood in Trinidad (BOL) and Phimai (THA), and somewhat overestimate the flood extent in Idah (NGA). Agreement between models is also high (indicated by reddish colours) in Huainan (CHN); however, the models overestimate the extent of flooding there, including a large area in the northwestern part of the region where most models falsely simulate flooding. In Chemba (MOZ), as well as in Alipur (PAK) and Ghotki (PAK), most models agree on flooding along the main river branches, but partly underestimate the extent of this flooding; while many models falsely simulate flooding in an extensive area to the east of the Indus River in Ghotki (PAK), and along the Sutlej river estuary in Alipur (PAK). Finally, in Dalby (AUS), only some models capture the more extensive parts of the flood; at the same time, several models simulate flooding alongside long stretches of the river channel where no flooding is observed.

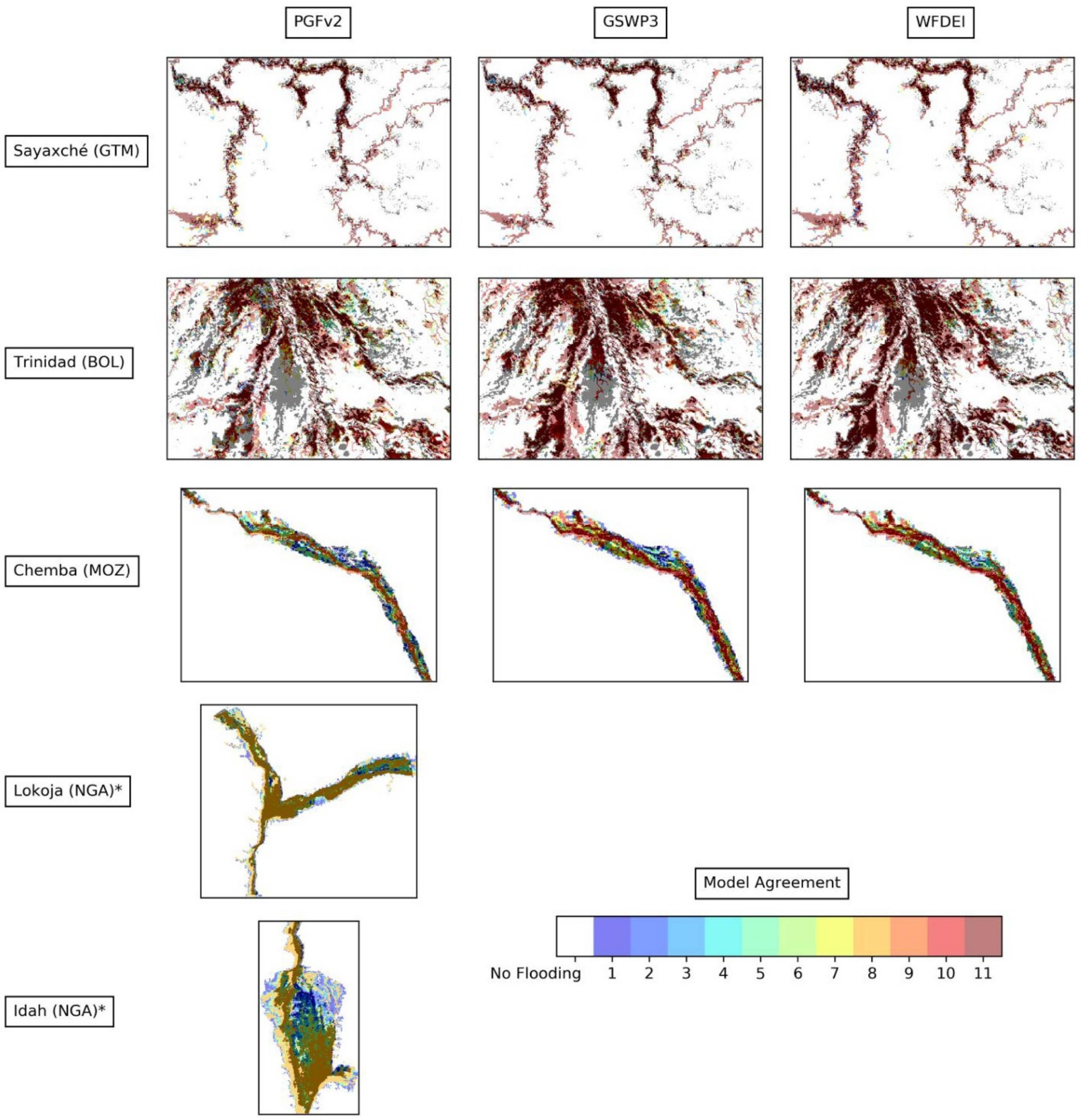


Figure 2. Model agreement map indicating the flood extent overlap between the 11 GHMs and the satellite data for each study region (row) and climate forcing (column). The cell colour represents the number of GHMs that computed the corresponding cell to be flooded. The underlying flood extent of the satellite imagery (light grey) is assigned a dark colour tone if it matches with at least one GHM. For the 2012 flood in Lokoja (NGA) and Idah (NGA) (marked with an asterisk *), only eight GHMs, driven with PGFv2, were available.

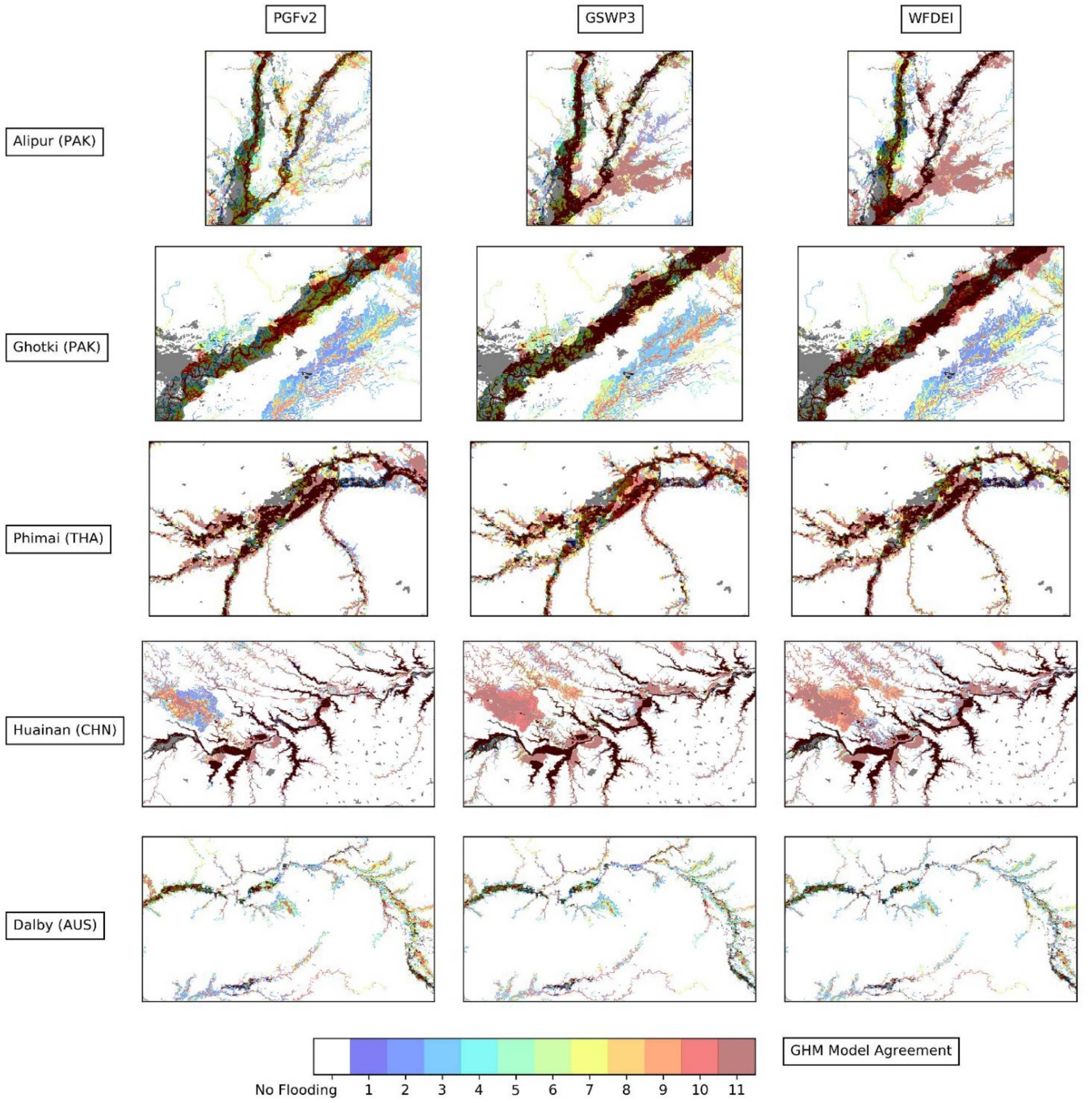


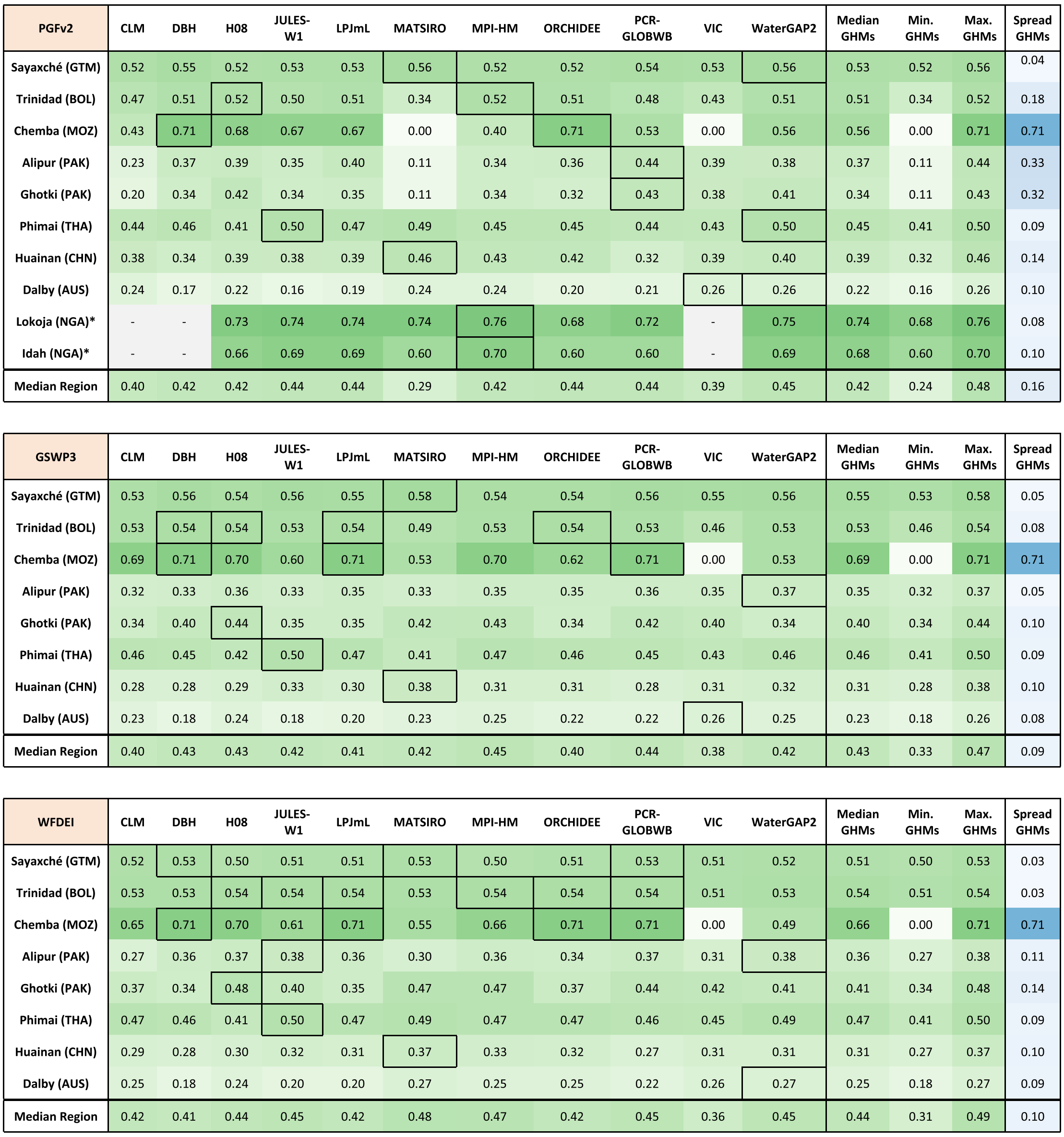
Figure 2. (Continued.)

The simulation of flooding in large areas where no flooding was observed—found for the floods in China, Pakistan, and Australia—could theoretically be induced by contamination of our flood extent estimate by a different flood event: i.e. an event occurring in the same year which had an even higher flood magnitude in that part of the region, and would thus be picked up when taking the annual maximum discharge in every grid cell. However, analysis of daily flood data confirms that this is not the case; while there can be a considerable delay of the flood peak in CaMa-Flood compared to observations (Zhao *et al* 2017), the estimated flood extent is largely related to one coherent flood event, except for outliers in marginal grid cells (supplementary material, figures S1–S33).

Regarding the climate forcing, the GSWP3 and WFDEI reanalysis datasets lead to very similar results. The PGFv2 dataset leads to markedly smaller simulated flood extents in Alipur (PAK), Ghotki (PAK), and Huainan (CHN), with most or all GHMs. This partly remedies the overestimation of flooding outside the main river floodplains, but also leads to a more substantial underestimation of flood extent along the main rivers in Alipur (PAK) and Ghotki (PAK), compared to the other two forcing datasets.

3.2. Model performance scores

In line with the observations from the model agreement maps, the CSI scores vary substantially between the different regions (figure 3). A comparatively low CSI is found for Alipur (PAK), Ghotki (PAK), Huainan (CHN), with scores between 0.3 and 0.4, and especially Dalby (AUS) with scores below 0.3. High CSI scores of around 0.5 and higher are found for Sayaxché (GTM), Trinidad (BOL), Chemba (MOZ), Lokoja (NGA), and Idah (NGA). Intermediate scores of around 0.45 are found for Phimai (THA). Since the CSI score also depends on the flood magnitude (Stephens *et al* 2014), this numerical comparison



PGFv2	CLM	DBH	H08	JULES-W1	LPJmL	MATSIRO	MPI-HM	ORCHIDEE	PCR-GLOBWB	VIC	WaterGAP2	Median GHMs	Min. GHMs	Max. GHMs	Spread GHMs
Sayaxché (GTM)	0.52	0.55	0.52	0.53	0.53	0.56	0.52	0.52	0.54	0.53	0.56	0.53	0.52	0.56	0.04
Trinidad (BOL)	0.47	0.51	0.52	0.50	0.51	0.34	0.52	0.51	0.48	0.43	0.51	0.51	0.34	0.52	0.18
Chemba (MOZ)	0.43	0.71	0.68	0.67	0.67	0.00	0.40	0.71	0.53	0.00	0.56	0.56	0.00	0.71	0.71
Alipur (PAK)	0.23	0.37	0.39	0.35	0.40	0.11	0.34	0.36	0.44	0.39	0.38	0.37	0.11	0.44	0.33
Ghotki (PAK)	0.20	0.34	0.42	0.34	0.35	0.11	0.34	0.32	0.43	0.38	0.41	0.34	0.11	0.43	0.32
Phimai (THA)	0.44	0.46	0.41	0.50	0.47	0.49	0.45	0.45	0.44	0.43	0.50	0.45	0.41	0.50	0.09
Huainan (CHN)	0.38	0.34	0.39	0.38	0.39	0.46	0.43	0.42	0.32	0.39	0.40	0.39	0.32	0.46	0.14
Dalby (AUS)	0.24	0.17	0.22	0.16	0.19	0.24	0.24	0.20	0.21	0.26	0.26	0.22	0.16	0.26	0.10
Lokoja (NGA)*	-	-	0.73	0.74	0.74	0.74	0.76	0.68	0.72	-	0.75	0.74	0.68	0.76	0.08
Idah (NGA)*	-	-	0.66	0.69	0.69	0.60	0.70	0.60	0.60	-	0.69	0.68	0.60	0.70	0.10
Median Region	0.40	0.42	0.42	0.44	0.44	0.29	0.42	0.44	0.44	0.39	0.45	0.42	0.24	0.48	0.16

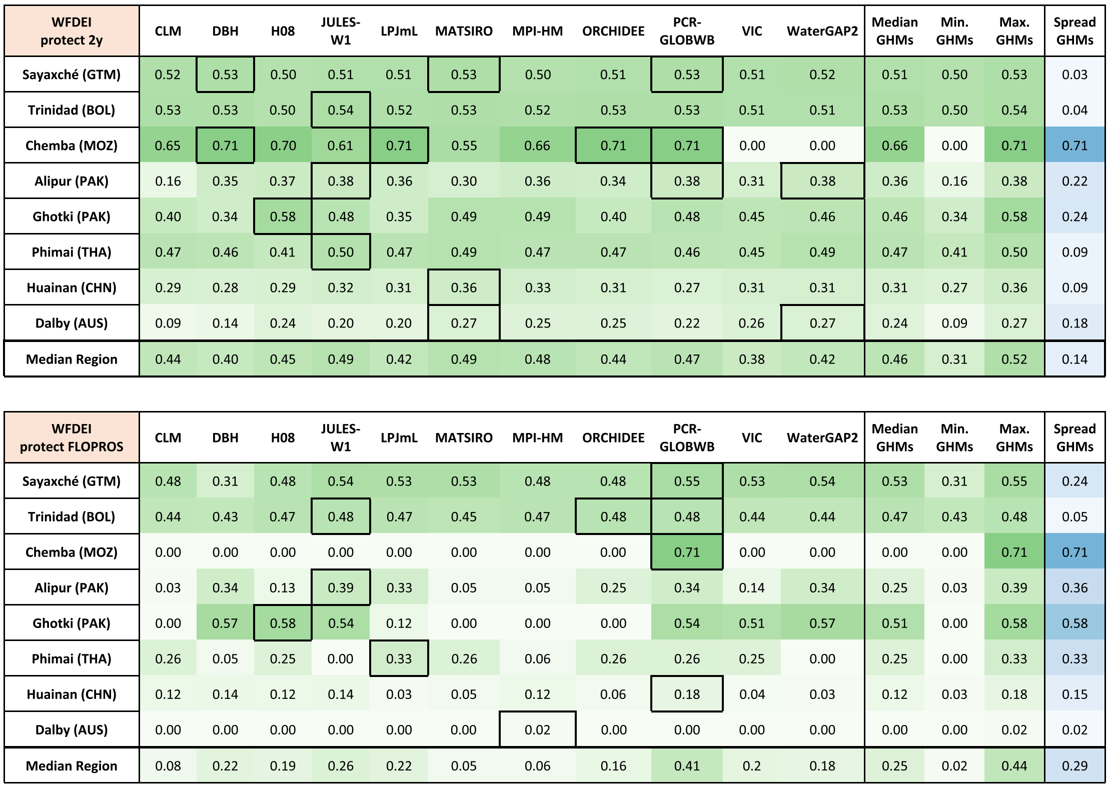
GSWP3	CLM	DBH	H08	JULES-W1	LPJmL	MATSIRO	MPI-HM	ORCHIDEE	PCR-GLOBWB	VIC	WaterGAP2	Median GHMs	Min. GHMs	Max. GHMs	Spread GHMs
Sayaxché (GTM)	0.53	0.56	0.54	0.56	0.55	0.58	0.54	0.54	0.56	0.55	0.56	0.55	0.53	0.58	0.05
Trinidad (BOL)	0.53	0.54	0.54	0.53	0.54	0.49	0.53	0.54	0.53	0.46	0.53	0.53	0.46	0.54	0.08
Chemba (MOZ)	0.69	0.71	0.70	0.60	0.71	0.53	0.70	0.62	0.71	0.00	0.53	0.69	0.00	0.71	0.71
Alipur (PAK)	0.32	0.33	0.36	0.33	0.35	0.33	0.35	0.35	0.36	0.35	0.37	0.35	0.32	0.37	0.05
Ghotki (PAK)	0.34	0.40	0.44	0.35	0.35	0.42	0.43	0.34	0.42	0.40	0.34	0.40	0.34	0.44	0.10
Phimai (THA)	0.46	0.45	0.42	0.50	0.47	0.41	0.47	0.46	0.45	0.43	0.46	0.46	0.41	0.50	0.09
Huainan (CHN)	0.28	0.28	0.29	0.33	0.30	0.38	0.31	0.31	0.28	0.31	0.32	0.31	0.28	0.38	0.10
Dalby (AUS)	0.23	0.18	0.24	0.18	0.20	0.23	0.25	0.22	0.22	0.26	0.25	0.23	0.18	0.26	0.08
Median Region	0.40	0.43	0.43	0.42	0.41	0.42	0.45	0.40	0.44	0.38	0.42	0.43	0.33	0.47	0.09

WFDEI	CLM	DBH	H08	JULES-W1	LPJmL	MATSIRO	MPI-HM	ORCHIDEE	PCR-GLOBWB	VIC	WaterGAP2	Median GHMs	Min. GHMs	Max. GHMs	Spread GHMs
Sayaxché (GTM)	0.52	0.53	0.50	0.51	0.51	0.53	0.50	0.51	0.53	0.51	0.52	0.51	0.50	0.53	0.03
Trinidad (BOL)	0.53	0.53	0.54	0.54	0.54	0.53	0.54	0.54	0.54	0.51	0.53	0.54	0.51	0.54	0.03
Chemba (MOZ)	0.65	0.71	0.70	0.61	0.71	0.55	0.66	0.71	0.71	0.00	0.49	0.66	0.00	0.71	0.71
Alipur (PAK)	0.27	0.36	0.37	0.38	0.36	0.30	0.36	0.34	0.37	0.31	0.38	0.36	0.27	0.38	0.11
Ghotki (PAK)	0.37	0.34	0.48	0.40	0.35	0.47	0.47	0.37	0.44	0.42	0.41	0.41	0.34	0.48	0.14
Phimai (THA)	0.47	0.46	0.41	0.50	0.47	0.49	0.47	0.47	0.46	0.45	0.49	0.47	0.41	0.50	0.09
Huainan (CHN)	0.29	0.28	0.30	0.32	0.31	0.37	0.33	0.32	0.27	0.31	0.31	0.31	0.27	0.37	0.10
Dalby (AUS)	0.25	0.18	0.24	0.20	0.20	0.27	0.25	0.25	0.22	0.26	0.27	0.25	0.18	0.27	0.09
Median Region	0.42	0.41	0.44	0.45	0.42	0.48	0.47	0.42	0.45	0.36	0.45	0.44	0.31	0.49	0.10

Figure 3. CSI scores for all combinations of GHMs and PGFv2 (top), GSWP3 (middle), and WFDEI (bottom) forcing. The 'Median Region' across the even number of regions is calculated as the mean of the two middle values. Lokoja (NGA) and Idah (NGA) (marked with *) were excluded from the computation of the 'Median Region'. '-' means no input data was available and a black box indicates the best-performing GHM(s) for a given region.

between regions and events should be interpreted with caution, and in conjunction with the model agreement maps. In particular, for a given region with a concave topography (e.g. extensive floodplain), the larger the flood magnitude, the more the flood extent will be constrained by topography, and the less variation in flood extent will be induced by a given variation in flood discharge; potentially leading to more favourable CSI scores than for smaller floods. These caveats however do not affect comparison between models and datasets within one region and event, which we turn to next.

We find differences in performance between climate forcings and between GHMs to be mutually dependent, and we discuss them together in the following. The spread across GHMs (rightmost column in figure 3) is by far largest for Chemba (MOZ). This is primarily because there the CSI with the VIC model is zero for all three forcings, and the CSI with the MATSIRO model is zero for the PGFv2 forcing; indicating that there is no intersection between simulated and observed flood extent. The MATSIRO model with PGFv2 forcing also has very low CSI scores in Alipur (PAK) and Ghotki (PAK). However, even if the MATSIRO and VIC models were excluded, Chemba (MOZ), Alipur (PAK) and Ghotki (PAK) would still remain the regions with the largest differences in CSI scores across GHMs, under the PGFv2 forcing; for instance, compare CLM and PCR-GLOBWB for Alipur (PAK) and Ghotki (PAK), or MPI-HM and ORCHIDEE for Chemba (MOZ). Using the other two climate forcings, these inter-GHM differences are



WFDEI protect 2y	CLM	DBH	H08	JULES-W1	LPJmL	MATSIRO	MPI-HM	ORCHIDEE	PCR-GLOBWB	VIC	WaterGAP2	Median GHMs	Min. GHMs	Max. GHMs	Spread GHMs
Sayaxché (GTM)	0.52	0.53	0.50	0.51	0.51	0.53	0.50	0.51	0.53	0.51	0.52	0.51	0.50	0.53	0.03
Trinidad (BOL)	0.53	0.53	0.50	0.54	0.52	0.53	0.52	0.53	0.53	0.51	0.51	0.53	0.50	0.54	0.04
Chemba (MOZ)	0.65	0.71	0.70	0.61	0.71	0.55	0.66	0.71	0.71	0.00	0.00	0.66	0.00	0.71	0.71
Alipur (PAK)	0.16	0.35	0.37	0.38	0.36	0.30	0.36	0.34	0.38	0.31	0.38	0.36	0.16	0.38	0.22
Ghotki (PAK)	0.40	0.34	0.58	0.48	0.35	0.49	0.49	0.40	0.48	0.45	0.46	0.46	0.34	0.58	0.24
Phimai (THA)	0.47	0.46	0.41	0.50	0.47	0.49	0.47	0.47	0.46	0.45	0.49	0.47	0.41	0.50	0.09
Huainan (CHN)	0.29	0.28	0.29	0.32	0.31	0.36	0.33	0.31	0.27	0.31	0.31	0.31	0.27	0.36	0.09
Dalby (AUS)	0.09	0.14	0.24	0.20	0.20	0.27	0.25	0.25	0.22	0.26	0.27	0.24	0.09	0.27	0.18
Median Region	0.44	0.40	0.45	0.49	0.42	0.49	0.48	0.44	0.47	0.38	0.42	0.46	0.31	0.52	0.14

WFDEI protect FLOPROS	CLM	DBH	H08	JULES-W1	LPJmL	MATSIRO	MPI-HM	ORCHIDEE	PCR-GLOBWB	VIC	WaterGAP2	Median GHMs	Min. GHMs	Max. GHMs	Spread GHMs
Sayaxché (GTM)	0.48	0.31	0.48	0.54	0.53	0.53	0.48	0.48	0.55	0.53	0.54	0.53	0.31	0.55	0.24
Trinidad (BOL)	0.44	0.43	0.47	0.48	0.47	0.45	0.47	0.48	0.48	0.44	0.44	0.47	0.43	0.48	0.05
Chemba (MOZ)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.00	0.00	0.71	0.71
Alipur (PAK)	0.03	0.34	0.13	0.39	0.33	0.05	0.05	0.25	0.34	0.14	0.34	0.25	0.03	0.39	0.36
Ghotki (PAK)	0.00	0.57	0.58	0.54	0.12	0.00	0.00	0.00	0.54	0.51	0.57	0.51	0.00	0.58	0.58
Phimai (THA)	0.26	0.05	0.25	0.00	0.33	0.26	0.06	0.26	0.26	0.25	0.00	0.25	0.00	0.33	0.33
Huainan (CHN)	0.12	0.14	0.12	0.14	0.03	0.05	0.12	0.06	0.18	0.04	0.03	0.12	0.03	0.18	0.15
Dalby (AUS)	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02
Median Region	0.08	0.22	0.19	0.26	0.22	0.05	0.06	0.16	0.41	0.2	0.18	0.25	0.02	0.44	0.29

Figure 4. CSI scores for all combinations of GHMs and WFDEI for 'protect 2y' (top) and 'protect FLOPROS' (bottom). The 'Median Region' across the even number of regions is calculated as the mean of the two middle values. Lokoja (NGA) and Idah (NGA) were excluded from the computation of the 'Median Region'. A black box indicates the best-performing GHM(s) for a given region.

much smaller, mostly with a difference of about 0.1 or less between the scores of the best and worst performing model.

The median CSI scores over all regions (bottom row in each subplot in figure 3) indicate that none of the GHMs performs consistently better or worse than the others. Even the VIC model, which fails to capture the flood in Chemba (MOZ), achieves reasonable scores in the other regions, and has a region-median CSI only slightly below the other GHMs (between 0.36 and 0.39 depending on the climate forcing). The MATSIRO model, which shows very low CSI scores in three of the regions under PGFv2 forcing, is among the best performing GHMs in some of the other regions; in particular, it achieves the highest CSI score of all GHMs in Huainan (CHN) for all three forcings, and has the highest region-median CSI score under the WFDEI forcing.

Similarly, the statistics over all GHMs (rightmost four columns in figure 3), and the combined statistics across regions and GHMs (bottom right in each subplot in figure 3, as well as figure 5), show that neither of the three climate forcing datasets is generally superior to the others: the GHM-region median CSI is 0.42, 0.43, and 0.44, respectively, for PGFv2, GSWP3, and WFDEI. PGFv2 exhibits slightly lower GHM-median CSI scores, and lower minimum scores, than the other two forcings in many of the regions; on the other hand, in Huainan (CHN), all GHMs achieve substantially higher scores with PGFv2 than with GSWP3 or WFDEI.

There is thus no climate forcing dataset that performs best in all regions; nor is there one GHM which consistently performs best in all regions and with all forcings (see black boxes in figure 3).

We now turn to the Bias scores. While a low CSI score generally already implies a high bias, the Bias score additionally indicates whether the observed total flood extent is over- or under predicted by the model. We find that with WFDEI and GSWP3 forcing, Bias scores are generally either small or positive (with the exception of VIC in Chemba (MOZ)) (figure 4). This confirms the observation made in the model agreement maps that flood extent is generally either matched relatively well, or overestimated, by the model simulations. With the PGFv2 climate forcing, the overall result is similar, but substantial negative Bias scores—indicating an under prediction of flood extent—occur in several cases: for Alipur (PAK), Ghotki (PAK), and Chemba (MOZ), with CLM, MATSIRO, and MPI-HM; corresponding to the anomalous CSI values for these combinations described above. On the other hand, the PGFv2 forcing leads to much smaller (positive) biases than the other two forcings in Huainan (CHN).

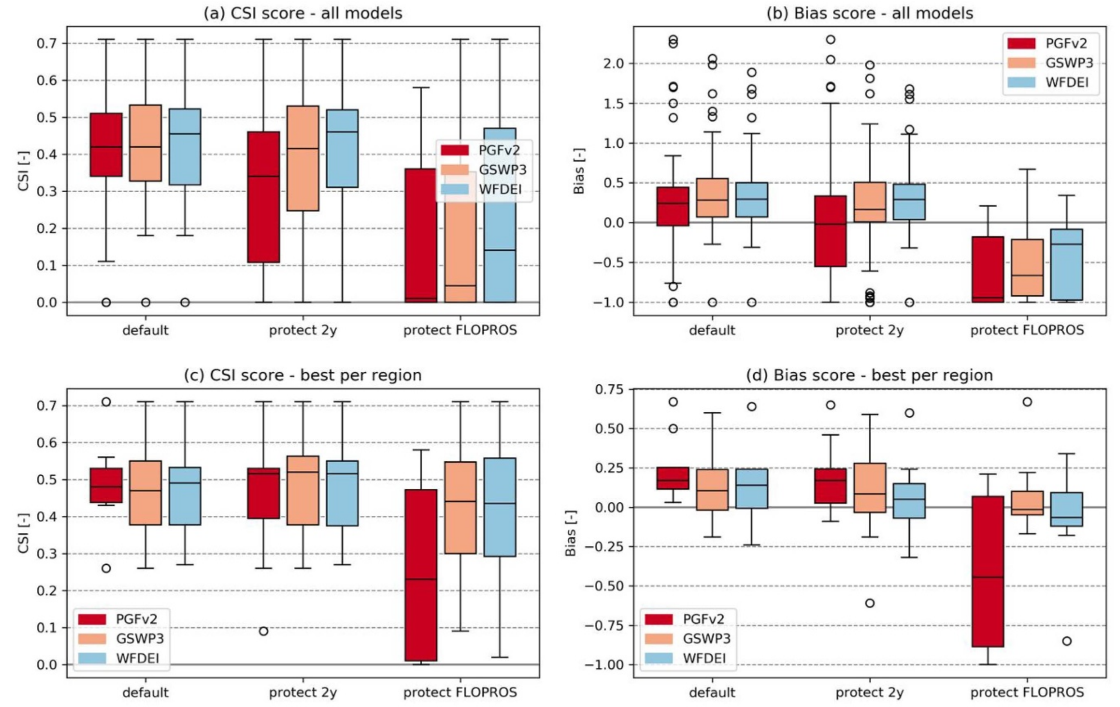


Figure 5. Comparison of CSI and Bias scores between the default setting ('default'), protection against floods with an ARI of 2 years ('protect 2y') and protection standards according to FLOPROS ('protect FLOPROS') for the climate forcings PGFv2, GSWP3 and WFDEI. Figures (a) and (b) include all GHMs and regions, (c) and (d) only account for the best GHM per region. The regions Lokoja (NGA) and Idah (NGA) were excluded from the computation.

Calculating the performance scores on the higher-resolution (18 arc sec) binary flood outputs confirms these overall results, though the scores are somewhat lower (supplementary material, figures S36 and S37).

3.3. Flood protection

We now analyse the simulations accounting for flood protection infrastructure, by counting only those flooded pixels whose ARI (estimated through fitting a GEV function) is longer than either 2 years ('protect 2y') or the protection standard indicated in the global database FLOPROS (Scussolini *et al* 2016) ('protect FLOPROS').

Under the WFDEI forcing, CSI scores in the 'protect 2y' simulations change only little compared to the default simulations. Most notably, there are now two GHMs simulating no flooding in Chemba (MOZ): WaterGAP2, in addition to VIC; and the CLM model shows lower CSI scores in Alipur (PAK) and Dalby (AUS) (figure 4). More widespread deterioration of CSI scores appears under the other two forcings. With GSWP3, there are multiple GHMs each in Chemba (MOZ), Alipur and Ghotki (PAK) that show no or almost no flooding (supplementary figure S39). With PGFv2, the CSI drops to zero in Chemba (MOZ) for almost all models, with the exception of DBH which still shows a high CSI score there; and CSI scores in Huainan (CHN) are seriously degraded for all GHMs. In summary, while for most regions the maximum CSI scores achieved are preserved when assuming protection against ARI of 2 years, the spread among models and among forcings increases notably compared to the default simulations (figures 5(a) and (c)).

This pattern becomes even more pronounced when assuming flood protection according to FLOPROS. CSI scores degrade in the majority of regions and forcing-GHM combinations (figures 4, 5 and supplementary figure S41). While all of the 'protect FLOPROS' simulations still exhibit significant CSI scores in Trinidad (BOL), they all show zero or almost zero CSI in Dalby (AUS). In most other regions, we find both: multiple GHM-forcing combinations that still achieve relatively high CSI scores, often preserving the maximum CSI for the region found in the 'default' setup; as well as many combinations showing zero or very low CSI scores. Chemba (MOZ) is an interesting case as here the CSI score drops to zero for most of the 'protect FLOPROS' simulations, while, with WFDEI forcing, one single GHM (PCR-GLOBWB) still achieves the same CSI as in the default setup; and the same is true for three GHMs (CLM, ORCHIDEE, and PCR-GLOBWB) with GSWP3 forcing.

While the Bias scores were mostly positive in the 'default' simulations, the 'protect FLOPROS' setup mostly results in negative biases, which are in many parts of substantial magnitude (figures 5(b) and supplementary figure S42). This corresponds to the decreased and, often, zero CSI scores found in this setup. Bias scores in the 'protect 2y' setup are, as expected, closer to those in the 'default' setup; with more pronounced overestimation of flood extent under WFDEI forcing, and more pronounced underestimation in some regions under PGFv2 forcing (supplementary figure S40).

Considering the entire simulation ensemble, the lowest median bias in flood extent is achieved in the 'protect 2y' setup, in particular the simulations forced with PGFv2 and GSWP3 yielding a lower median bias than in the 'default' setup (figure 5(b)). On other hand, the spread among simulations is largest for 'protect 2y'. In the 'protect FLOPROS' setup, the majority of simulations show a large negative bias, as discussed above. However, when considering for each region only the best-performing model, the least biased results are achieved with WFDEI forcing in the 'protect FLOPROS' setup; the corresponding GSWP3 simulations being only slightly more biased (figure 5(d)).

4. Discussion and conclusions

A number of key results emerge from our analysis:

(1) The performance of a global river flood modelling chain in reproducing observed flood extent for major recent floods differs considerably between events. While CSI scores of 0.7 and higher are obtained for Chemba (MOZ), Idah (NGA), Lokoja (NGA) (similar as in Bernhofen *et al* (2018)), scores are much lower for many other events, dropping below 0.3 for Dalby (AUS).

(2) The choice of GHM and climate forcing have mutually dependent effects on flood model performance.

(3) PGFv2 performs somewhat poorer than the other two climate forcing datasets for many regions and GHMs, but better for some. The performance spread between GHMs is largest with PGFv2.

(4) No climate forcing or GHM performs best for all regions. Considering the median over all regions, the PCR-GLOBWB model stands out as achieving the best, or among the best, results for all forcings and in particular for the 'protect' setups.

(5) Accounting for flood protection according to the FLOPROS database dramatically degrades the average agreement between simulations and observations, by reducing or eliminating simulated flood extent in many cases. However, individual forcing-GHM combinations remain in almost every region that achieve high CSI and low Bias scores.

Regarding key result no. 1, we reiterate that CSI scores should not be compared between regions at face value, because of the varying flood extents. However, it is also evident from the maps in figure 2 that both the overall match between models and observations, and the level of agreement among models, differ depending on the region and event that is analysed. This may be explained by varying topographies among the regions, leading to different errors in simulated flood extent due to cross-floodplain slopes. Another important caveat is that the shape of the study area is not consistent across regions. While the study areas for most regions are rectangular and include large parts where no flooding was observed, the study areas for Chemba (MOZ), Lokoja (NGA), and Idah (NGA) are irregularly shaped polygons narrowly outlining the observed flood extent along the main river channels. This means that potential flooding along tributaries is excluded from the analysis (supplementary figures S45–S48). Considering an example of a model overestimating observed flood extent, the overestimation may appear less severe if parts of the flood that are further away from the main channel are cut off. With the rectangular study areas, such excess simulated flood extent would be more visible and would degrade the CSI score to a larger degree. This might go some way in explaining why Chemba (MOZ), and especially Lokoja (NGA) and Idah (NGA), exhibit systematically higher CSI scores than other regions. These scores are similar to those found for the same three regions by Bernhofen *et al* (2018), who used the same study area outlines. This may indicate that the general level of model performance found in our study is comparable to that in Bernhofen *et al* (2018), and that the lower CSI scores in the additional regions in our study may also be related to the layout of the study area, rather than only to a poorer model performance in those regions.

Regarding key result no. 2, the importance of the choice of GHM is confirmed by a recent study that compared different sources of uncertainty in CaMa-Flood estimates: the GHM runoff inputs, variables for flood frequency analysis and fitting distributions (Zhou *et al* 2021). Of these three, the GHMs were found to be by far the largest source of variation in the estimate of flood depth and inundation. That study used a single reanalysis forcing dataset (WFDEI) as input to the GHMs and did not evaluate different GHMs' performance in relation to observed flooding; rendering our study a useful complement. It should be pointed out that the hydrological modelling approach underlying many of these GHMs is traditionally aimed at evaluating water and energy balances at longer timescales (also termed catchment yield-type models, in contrast to rainfall-runoff-type models), without specific focus on flood generation; which may partly explain their deficiencies in estimating flood magnitudes and timing. Disparities between GHM-simulated and actual runoff may thus be a fundamental reason for the relatively poor performance of the flood modelling methodology applied here, compared to conventional basin-scale flood analysis;

also affecting the results summarized in key results no. 3 and 4.

Regarding key result no. 3, Müller Schmied *et al* (2016) evaluated hydrological simulations with a single GHM (WaterGAP2) driven by the same climate forcing datasets as in our study, and found substantial differences in long-term average runoff and other hydrological indicators. In particular, relatively low precipitation and runoff estimates were found with PGFv2, compared to GSWP3 and WFDEI; likely related to the lack of a precipitation undercatch correction in this dataset, but potentially also to a different observational product that precipitation was corrected against (CRU TS3.21 for PGFv2; GPCC version 6 for GSWP3 and WFDEI). While runoff extremes were not directly assessed in that study, systematically lower precipitation forcing could nonetheless explain the larger negative biases in flood extent that we find in our default simulations with PGFv2 for some regions and GHMs; and might likewise explain the smaller positive biases and higher CSI scores in Huainan (CHN), where flood extent is strongly overestimated with GSWP3 and WFDEI forcing.

Nevertheless, the differences even between GSWP3- and WFDEI-driven flood simulations highlight some remaining uncertainty in the forcing data. To test whether using observational datasets directly as input to the GHMs might be beneficial, we performed a set of simulations where, for each event and grid cell, the return period simulated by a given GHM-forcing combination was mapped to the corresponding flood volume given by a benchmark simulation of MATSIRO which was driven with station-based (GPCC) rather than reanalysis precipitation data (Kim *et al* 2009). This adjustment procedure was originally devised to correct biases in climate model-derived runoff (Hirabayashi *et al* 2013), and has been applied in other global flood modelling studies (Dottori *et al* 2018, Willner *et al* 2018, Sauer *et al* 2021). However, in our study, the adjustment using the MATSIRO benchmark simulation does not systematically improve the agreement between simulated and observed flood extent (supplementary material figures S43 and S44); suggesting that neither a particular GHM nor the observations-based precipitation forcing are generally superior to the GHM ensemble and the reanalysis-based forcings studied here, respectively.

Regarding key result no. 5, one striking finding is that the incorporation of flood protection standards according to FLOPROS leads to zero simulated flood extent in some regions with many or all GHMs and climate forcings. This can either mean that the protection standard indicated in the FLOPROS database is higher than in reality for these regions; or that the return period simulated by the models is too short—in other words, that the models simulate too frequent flooding; or both. It must be remembered that the FLOPROS protection levels are purely model-based estimates for most of the events; except for those in Mozambique, China and Australia, where information about the actual design standards or about corresponding policy regulations entered the estimates provided in the database. Thus, FLOPROS values may not exactly reflect the flood protection standards actually implemented in the study regions.

Reported estimates of the average recurrence interval (ARI) are only available for some of the flood events studied here, and available estimates often differ between sources and/or come with considerable uncertainties. Nevertheless, the actual ARI appears to be higher than the FLOPROS protection standard for most events (supplementary table S3). This suggests that the deterioration of model results when applying FLOPROS may not be predominantly related to errors in FLOPROS. Instead, it suggests that the ARI in the affected model simulations may be too short; i.e. the model too frequently simulates a flood of the given magnitude.

The 'protect FLOPROS' simulations thus serve to highlight those GHM-climate forcing combinations that correctly simulate an ARI larger than the protection standard, according to FLOPROS. For instance, applying FLOPROS, the CSI for Chemba (MOZ) drops to zero for almost all simulations except for PCR-GLOBWB with GSWP3 and WFDEI forcing, and for CLM and ORCHIDEE with GSWP3 forcing (supplementary figure S41). Similarly, imposing FLOPROS flood protection levels in Sayaxché (GTM), the majority of models still achieve a reasonable CSI with GSWP3 and WFDEI forcing but simulate no flooding with PGFv2 forcing. In Alipur (PAK) and Ghotki (PAK), only JULES-W1 and PCR-GLOBWB realistically simulate a large ARI with all three climate forcings. Interestingly, not a single simulation shows substantial flooding in Dalby (AUS) in the 'protect FLOPROS' setup; suggesting that here, the protection standard of a 100 year ARI assumed in FLOPROS may indeed be too high; which is also in line with reports of a 90 year ARI for the 2010/2011 flood event (supplementary table S3).

Generally though, many of the runoff simulations used here may still be in need of improvement with respect to the high-end of the runoff distribution. Indeed, a recent evaluation of monthly runoff simulated by six GHMs, including H08, LPJmL, MATSIRO, PCR-GLOBWB, and WaterGAP2, found that most models—except MATSIRO and WaterGAP2—tended to overestimate high-flow runoff (more precisely, Q5, the magnitude of runoff that is exceeded 5% of the time) (Zaherpour *et al* 2018). GHMs particularly struggle to capture the levels and variability of runoff and, consequently, river discharge in more arid environments such as parts of the Murray–Darling basin (Haddeland *et al* 2011, Hattermann *et al* 2017, Zaherpour *et al* 2018). While the CaMa-Flood river routing model has been shown to improve the discharge hydrograph compared to GHMs' native routing schemes in many basins (Zhao *et al* 2017), it may not always be able to compensate a systematic overestimation of high-flow runoff by the GHMs, which may then result in an overestimation of flood extent. Moreover, CaMa-Flood does not account for human water management such as water withdrawals or dams, which may in some cases have significant effects on flood volume and timing (Mateo *et al* 2014, Zhao *et al* 2017). Other limitations of the GFM include the accuracy of the baseline topography, and the use of a global empirical equation to calculate channel depth as function of annual river discharge, without separate calibration for each river (Yamazaki *et al* 2014).

We can thus derive from our study some recommendations for future applications of GFMs to simulate flood extent. One is that the choice of GHM (or more generally, the runoff model) and climate forcing should be carefully considered, because it can strongly impact performance. The good news is that serious losses in flood extent performance occurred only with a limited number of individual climate forcing-GHM combinations. Two of our three climate forcings, and the majority of GHMs used, showed very similar levels of performance. The more difficult news is that there is no general recommendation on which forcings or runoff models not to use; because even those that perform particularly poorly in some regions may actually be the best choice in a different region, or in a different GHM-climate forcing combination. A multi-model, multi-forcing ensemble approach may be advisable when there is no prior knowledge about a certain combination's performance for the specific type of event and region under investigation. That being said, validating the underlying runoff model(s) separately from the flood model is a crucial component of robust flood risk analysis; and the performance of each part in the modelling chain should be taken into account to determine whether the modelling chain is fit for a given purpose.

Global flood modelling capacities could profit from further development of GHMs, for instance, addressing the difficulty to accurately capture runoff extremes in arid and semi-arid regions. Weighted ensembles of models may provide a useful method when systematic differences in model performance can be identified (Zaherpour *et al* 2018). A closer coupling of runoff and flood modelling, accounting for human alterations of river flow, could improve flood estimates in highly managed river basins (Mateo *et al* 2014, Boulange *et al* 2021). Not least, improving the availability and fidelity of observational data, e.g. by extending direct observations of precipitation, runoff, or flood levels, and by making existing data more accessible—including on human-made alterations of the natural river flow—would help with both the calibration and validation of the different parts in the flood modelling chain (Müller Schmied *et al* 2016).

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The research was supported within the European Union Horizon 2020 projects RECEIPT, CASCADES, and HABITABLE, as well as BMBF project ISIpedia. The ISIMIP modelling teams are gratefully acknowledged. We would like to thank two anonymous reviewers for providing insightful and valuable comments that greatly improved the manuscript.

ORCID iDs

Benedikt Mester https://orcid.org/0000-0001-7731-476X
Sven Norman Willner https://orcid.org/0000-0001-6798-6247
Katja Frieler https://orcid.org/0000-0003-4869-3013
Jacob Schewe https://orcid.org/0000-0001-9455-4159

References

Aronica G, Bates P D and Horritt M S 2002 Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE *Hydrol. Process.* **16** 2001–16

Bates P D *et al* 2021 Combined modeling of US fluvial, pluvial, and coastal flood hazard under current and future climates *Water Resour. Res.* **57** e2020WR028673

Bates P D and de Roo A P J 2000 A simple raster-based model for flood inundation simulation *J. Hydrol.* **236** 54–77

Bernhofen M V *et al* 2018 A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique *Environ. Res. Lett.* **13** 104007

Best M J *et al* 2011 The joint UK land environment simulator (JULES), model description—part 1: energy and water fluxes *Geosci. Model Dev.* **4** 677–99

Blöschl G *et al* 2019 Changing climate both increases and decreases European river floods *Nature* **573** 108–11

Boulange J, Hanasaki N, Yamazaki D and Pokhrel Y 2021 Role of dams in reducing global flood exposure under climate change *Nat. Commun.* **12** 1–7

Brakenridge G R 2006 Global active archive of large flood events (Dartmouth Flood Observatory, University of Colorado) (available at: http://floodobservatory.colorado.edu/Archives/index.html) (accessed 06 July 2020)

Canning S and Walton R Western Downs Regional Council 2014 *2014 Flood Study Reports - Dalby Flood Study Volume I Detailed Technical Report* Western Downs Regional Council (available at: https://www.wdrc.qld.gov.au/wp-content/uploads/2015/08/Dalby-Flood-Study-Volume-I-Detailed-Technical-Report-April-2014.pdf)

Do H X, Gudmundsson L, Leonard M and Westra S 2018 The global streamflow indices and metadata archive (GSIM)—part 1: the production of a daily streamflow archive and metadata *Earth Syst. Sci. Data* **10** 765–85

Dottori F *et al* 2018 Increased human and economic losses from river flooding with anthropogenic warming *Nat. Clim. Change* **8** 781–6

Dottori F, Salamon P, Bianchi A, Alfieri L, Hirpa F A and Feyen L 2016 Development and evaluation of a framework for global flood hazard mapping *Adv. Water Resour.* **94** 87–102

EM-DAT 2020 The OFDA/CRED international disaster database, University Catholic Louvain-Brussels, Belgium (available at: www.emdat.be) (accessed 02 September 2020)

Essou G R C, Brissette F and Lucas-Picher P 2017 The use of reanalyses and gridded observations as weather input data for a hydrological model: comparison of performances of simulated river flows based on the density of weather stations *J. Hydrometeorol.* **18** 497–513

Farr T G *et al* 2007 The shuttle radar topography mission *Rev. Geophys.* **45** RG2004

Floodlist 2016 *Floodlist* (available at: http://floodlist.com/asia/china-july-2016-floods-cost-33-billion-dollars) (accessed 02 September 2020)

Gudmundsson L, Do H X, Leonard M and Westra S 2018 The global streamflow indices and metadata archive (GSIM)—part 2: quality control, time-series indices and homogeneity assessment *Earth Syst. Sci. Data* **10** 787–804

Haddeland I *et al* 2011 Multimodel estimate of the global terrestrial water balance: setup and first results *J. Hydrometeorol.* **12** 869–84

Hanasaki N, Kanae S, Oki T, Masuda K, Motoya K, Shirakawa N, Shen Y and Tanaka K 2008 An integrated model for the assessment of global water resources—part 1: model description and input meteorological forcing *Hydrol. Earth Syst. Sci.* **12** 1007–25

Hattermann F F *et al* 2017 Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins *Clim. Change* **141** 561–76

Hirabayashi Y, Mahendran R, Koirala S, Konoshima L, Yamazaki D, Watanabe S, Kim H and Kanae S 2013 Global flood risk under climate change *Nat. Clim. Change* **3** 816–21

Hoch J M and Trigg M A 2019 Advancing global flood hazard simulations by improving comparability, benchmarking, and integration of global flood models *Environ. Res. Lett.* **14** 034001

Hurlbert M A 2018 *Adaptive Governance of Disaster—Drought and Flood in Rural Areas* . In Water Governance - Concepts, Methods, and Practice (Berlin: Springer) (https://doi.org/10.1007/978-3-319-57801-9)

Hyungjun K 2014 Global Soil Wetness Project phase 3 forcing data set (GSWP3) (available at: http://hydro.iis.u-tokyo.ac.jp/GSWP3/exp1.html#initial-conditions) (accessed 21 December 2020)

IDMC 2019 *IDMC Global Report on Internal Displacement 2019 Displacement Dataset* (available at: www.internal-displacement.org/database/displacement-data) (accessed 02 September 2020)

IPCC 2014 Long-term climate change: projections, commitments and irreversibility pages 1029–1076 *Climate Change 2013—The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge: Cambridge University Press) pp 1029–136

IPCC 2012 *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change* vol 9781107025, ed C B Field *et al* (Cambridge: Cambridge University Press)

Jha A K, Bloch R and Lamond J 2012 *Cities and Flooding: A Guide to Integrated Urban Flood Risk Management for the 21st Century* (World Bank. © World Bank. License: CC BY 3.0 IGO) (available at: https://openknowledge.worldbank.org/handle/10986/2241) (accessed 02 September 2020)

Jongman B, Winsemius H C, Aerts J C J H, de Perez E C, van Aalst M K, Kron W and Ward P J 2015 Declining vulnerability to river floods and the global benefits of adaptation *Proc. Natl Acad. Sci. USA* **112** E2271–80

Kim H, Yeh P J-F, Oki T and Kanae S 2009 Role of rivers in the seasonal variations of terrestrial water storage over global basins *Geophys. Res. Lett.* **36** L17402

Lehner B, Verdin K and Jarvis A 2008 New global hydrography derived from spaceborne elevation data *EOS Trans. Am. Geophys. Union* **89** 93–4

Leng G, Huang M, Tang Q and Leung L R 2015 A modeling study of irrigation effects on global surface water and groundwater resources under a changing climate *J. Adv. Model. Earth Syst.* **7** 1285–304

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. Atmos.* **99** 14415–28

Masutomi Y, Inui Y, Takahashi K and Matsuoka Y 2009 Development of highly accurate global polygonal drainage basin data *Hydrol. Process.* **23** 572–84

Mateo C M, Hanasaki N, Komori D, Tanaka K, Kiguchi M, Champathong A, Sukhapunnaphan T, Yamazaki D and Oki T 2014 Assessing the impacts of reservoir operation to floodplain inundation by combining hydrological, reservoir management, and hydrodynamic models *Water Resour. Res.* **50** 7245–66

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. Discuss.* **20** 2877–98

Munich R 2020 NatCatSERVICE relevant flood/flash flood events worldwide 2010–2018 (available at: www.munichre.com/en/solutions/for-industry-clients/natcatservice.html) (accessed 03 September 2020)

Peel M C, Finlayson B L and McMahon T A 2007 Updated world map of the Köppen-Geiger climate classification *Hydrol. Earth Syst. Sci.* **11** 1633–44

Pokhrel Y N, Koirala S, Yeh P J-F, Hanasaki N, Longuevergne L, Kanae S and Oki T 2014 Incorporation of groundwater pumping in a global land surface model with the representation of human impacts *Water Resour. Res.* **51** 78–96

Raadgever T and Hegger D 2018 *Flood Risk Management Strategies and Governance* (Berlin: Springer) (https://doi.org/10.1007/978-3-319-67699-9)

Sampson C C, Smith A M, Bates P D, Neal J C, Alfieri L and Freer J E 2015 A high-resolution global flood hazard model *Water Resour. Res.* **51** 7358–81

Sauer I J, Reese R, Otto C, Geiger T, Willner S N, Guillod B P, Bresch D N and Frieler K 2021 Climate signals in river flood damages emerge under sound regional disaggregation *Nat. Commun.* **12** 2128

Scussolini P, Aerts J C J H, Jongman B, Bouwer L M, Winsemius H C, de Moel H and Ward P J 2016 FLOPROS: an evolving global database of flood protection standards *Nat. Hazards Earth Syst. Sci.* **16** 1049–61

Sheffield J, Goteti G and Wood E F 2006 Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling *J. Clim.* **19** 3088–111

Sitch S *et al* 2003 Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model *Glob. Change Biol.* **9** 161–85

Stacke T and Hagemann S 2012 Development and evaluation of a global dynamical wetlands extent scheme *Hydrol. Earth Syst. Sci.* **16** 2915–33

Stephens E, Schumann G and Bates P 2014 Problems with binary pattern measures for flood model evaluation *Hydrol. Process.* **28** 4928–37

Tang Q, Oki T, Kanae 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

Tanoue M, Hirabayashi Y and Ikeuchi H 2016 Global-scale river flood vulnerability in the last 50 years *Sci. Rep.* **6** 1–8

Traore A K, Ciais P, Vuichard N, Poulter B, Viovy N, Guimberteau M, Jung M, Myneni R and Fisher J B 2014 Evaluation of the ORCHIDEE ecosystem model over Africa against 25 years of satellite-based water and carbon measurements *J. Geophys. Res. Biogeosci.* **119** 1554–75

Trigg M A *et al* 2016 The credibility challenge for global fluvial flood risk analysis *Environ. Res. Lett.* **11** 094014

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

Weedon G P, Balsamo G, Bellouin N, Gomes S, Best M J and Viterbo P 2014 The WFDEI meteorological forcing data set: WATCH forcing data methodology applied to ERA-interim reanalysis data *Water Resour. Res.* **50** 7505–14

Weedon G P, Gomes S, Viterbo P, Shuttleworth W J, Blyth E, Österle H, Adam J C, Bellouin N, Boucher O and Best M 2011 Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century *J. Hydrometeorol.* **12** 823–48

Werner M G F, Hunter N M and Bates P D 2005 Identifiability of distributed floodplain roughness values in flood extent estimation *J. Hydrol.* **314** 139–57

Willner S N, Levermann A, Zhao F and Frieler K 2018 Adaptation required to preserve future high-end river flood risk at present levels *Sci. Adv.* **4** eaao1914

World Meteorological Organization 2009 *Integrated Flood Management Concept Paper* (Geneva: WMO)

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

Yamazaki D, Oki T and Kanae S 2009 Deriving a global river network map and its sub-grid topographic characteristics from a fine-resolution flow direction map *Hydrol. Earth Syst. Sci.* **13** 2241–51

Yamazaki D, Sato T, Kanae S, Hirabayashi Y and Bates P D 2014 Regional flood dynamics in a bifurcating mega delta simulated in a global river model *Geophys. Res. Lett.* **41** 3127–35

Zaherpour J *et al* 2018 Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts *Environ. Res. Lett.* **13** 065015

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

Zhou X, Ma W, Echizenya W and Yamazaki D 2021 The uncertainty of flood frequency analyses in hydrodynamic model simulations *Nat. Hazards Earth Syst. Sci.* **21** 1071–85