A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique

Environmental Research Letters

LETTER

A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique

Mark V Bernhofen[1], Charlie Whyman[1,2], Mark A Trigg[1], P Andrew Sleigh[1], Andrew M Smith[3], Christopher C Sampson[3], Dai Yamazaki[4], Philip J Ward[5], Roberto Rudari[6], Florian Pappenberger[7], Francesco Dottori[8], Peter Salamon[8] and Hessel C Winsemius[9]

[1] School of Civil Engineering, University of Leeds, Leeds, LS2 9JT, United Kingdom
[2] Stantec, High Wycombe, HP11 1JU, United Kingdom
[3] Fathom Global, Engine Shed, Temple Meads, Bristol, BS1 6QH, United Kingdom
[4] Institute of Industrial Science, The University of Tokyo, Tokyo, 153-8505, Japan
[5] Institute for Environmental Studies (IVM), Vrije Universiteit Amsterdam, 1081HV Amsterdam, The Netherlands
[6] CIMA Research Foundation, I-17100, Savona, Italy
[7] European Centre for Medium-Range Weather Forecasts, Reading, RG2 9AX, United Kingdom
[8] European Commission, Joint Research Centre, I-21027 Ispra, Italy
[9] Deltares, 2629 HV Delft, The Netherlands

E-mail: cn13mvb@leeds.ac.uk

Keywords: global flood models, validation, flooding, flood hazard, flood risk, rivers

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
13 April 2018

REVISED
28 August 2018

ACCEPTED FOR PUBLICATION
10 September 2018

PUBLISHED
1 October 2018

Abstract

Global flood models (GFMs) are becoming increasingly important for disaster risk management internationally. However, these models have had little validation against observed flood events, making it difficult to compare model performance. In this paper, we introduce the first collective validation of multiple GFMs against the same events and we analyse how different model structures influence performance. We identify three hydraulically diverse regions in Africa with recent large scale flood events: Lokoja, Nigeria; Idah, Nigeria; and Chemba, Mozambique. We then evaluate the flood extent output provided by six GFMs against satellite observations of historical flood extents in these regions. The critical success index of individual models across the three regions ranges from 0.45 to 0.7 and the percentage of flood captured ranges from 52% to 97%. Site specific conditions influence performance as the models score better in the confined floodplain of Lokoja but score poorly in Idah's flat extensive floodplain. 2D hydrodynamic models are shown to perform favourably. The models forced by gauged flow data show a greater level of return period accuracy compared to those forced by climate reanalysis data. Using the results of our analysis, we create and validate a three-model ensemble to investigate the usefulness of ensemble modelling in a flood hazard context. We find the ensemble model performs similarly to the best individual and aggregated models. In the three study regions, we found no correlation between performance and the spatial resolution of the models. The best individual models show an acceptable level of performance for these large rivers.

Introduction

Flooding is the most frequent and the most damaging of natural disasters globally [1]. From 1995–2015, floods affected 2.3 billion people, killing 157 000 [2]. Fluvial (river) flooding is the most common type of flood event and with over half of the world's population living within 3 km of a freshwater body, it has truly global implications [3]. Flood impacts will continue to increase in severity, as the population exposed to fluvial flooding is expected to rise by 31% over the next 30 years. Certain vulnerable regions, such as Sub-Saharan Africa, are predicted to see an increase in exposed population by as much as 104% [4]. Given current CO_2 emission trends, global temperatures could rise by up to 4 °C by 2100 [5]. To put this into a fluvial flooding context, a temperature rise of 4 °C could result in 70% of the global population

experiencing a 500% increase in flood risk [6]. Increased population exposure, coupled with the increased frequency and severity of flooding, means that reducing the risks associated with flooding is of vital importance to the United Nations Office for Disaster Risk Reduction (UNISDR) as outlined in their global assessment reports [7]. Reducing disaster vulnerability is a key target in goal 11 of the United Nation's Sustainable Development Goals [8] and specific risk reduction targets, to be met by 2030, were introduced in the Sendai Framework for Disaster Risk Reduction [9].

Flood models are an integral tool for managing and reducing the risks associated with flooding. In the past decade, increased computing power and precision of remote sensing datasets has led to the development of global flood models (GFMs) [10]. These models are being developed by a number of different groups that include consultancies [11], research groups [12], intergovernmental organizations [13, 14], academia [15], and academic affiliated companies [16, 17]. GFMs are being actively used for disaster risk management: providing flood hazard maps in data-scarce countries where there is little local or national information about flood risk [18]. They are also being used extensively in research: for evaluating the benefits of flood protection investments globally [19] and to determine changes in future flood risk due to climate change [6, 20, 21].

Despite their extensive applicability, each flood model has only had limited, internal, validation against either observed events, existing regional models, or reported fatalities and financial losses [12–17, 19, 22]. The Global Flood Partnership (GFP) (https://gfp.jrc.ec.europa.eu/), a cooperation framework between developers and users of global flood tools, made the comparison of GFMs a research priority at their annual meeting in 2014 [23]. The resulting GFM Intercomparison Project (GFMIP) was the first study to compare the flood hazard output of six GFMs on the continent of Africa. Research from the GFMIP showed there was wide variation in the flood hazard output of the six GFMs [24]. The GFMIP identified the need for collective validation of the GFMs against observed flood extents.

This study is a continuation to the GFMIP, using its outputs and original GFM model output data to validate against observed flood events and expand on the testing of collective model output. It is the first study to validate multiple GFMs under the same framework and against the same observed events, allowing model performance to be easily compared. This study should help identify which GFMs perform best and how different model structures influence performance. The results should also provide further insight into the reasons for model disagreement originally identified in the GFMIP [25].

The collective validation presented in this paper expands the rigorous GFM comparison begun in the GFMIP. As the models are improved and are used more extensively for disaster risk reduction, the need to compare model performance becomes increasingly apparent. The results of a rigorous comparison provide both users and model developers with information pertinent to the potential applicability of GFMs.

In this study, we identify regions with recent, large scale flood events with good observational validation data. We then develop a validation framework under which we test the output of six GFMs and the aggregated output of the GFMIP. We aim to answer which models perform best and identify the most important model characteristics affecting GFM performance. We also investigate whether an ensemble of the best individual GFMs improves the predicted flood extent.

Data and methodology

Models

The six GFMs compared in the GFMIP and in this study are the Catchment-Based Macro-scale Floodplain (CaMa-Flood) model [15], the Centro Internazionale in Monitoraggio Ambientale and United Nations Environment Program (CIMA-UNEP) model [13], the European Centre for Medium-Range Weather Forecasts (ECMWF) [14] model, the Global Flood Risk with Image Scenarios (GLOFRIS) model [17, 22], the Joint Research Centre (JRC) model [12], and the SSBN model (now known as Fathom Global Ltd) [16]. GFM output was provided for this study by each of the six developers in the form of flood extent maps. The models use different techniques to predict flood extent and depth for a given return period flow. These range in complexity from 1D hydraulic modelling (CIMA-UNEP) and simple 2D flood redistribution methods (GLOFRIS) to more complex 2D (ECMWF and CaMa-UT) and 2D hydrodynamic models (JRC and SSBN). GFM forcing can be split into cascade model type (CaMa-Flood, GLOFRIS, ECMWF, JRC) and gauged flow model type (SSBN, CIMA-UNEP) [24]. Cascade models use climate reanalysis data over 40 years to determine the probability that a cell is flooded. Gauged flow models use a growth curve to determine extreme flow. This flow is then input into a hydraulic model that predicts the flood extent for a given return period flow. Model output resolutions at the equator vary between ~90 m (SSBN, CIMA-UNEP), ~540 m (CaMa-UT, ECMWF), and ~900 m (GLOFRIS, JRC). All the GFMs use the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) as their input DEM. Further information regarding model setup and the differences in model forcing and computational engine can be found in the supplementary material (available online at stacks.iop.org/ERL/13/104007/mmedia). The aggregated fluvial flood extent (figure 1(d)), an output of the GFMIP that shows the level of agreement in flood extent between all six models, was also validated in this study to assess the

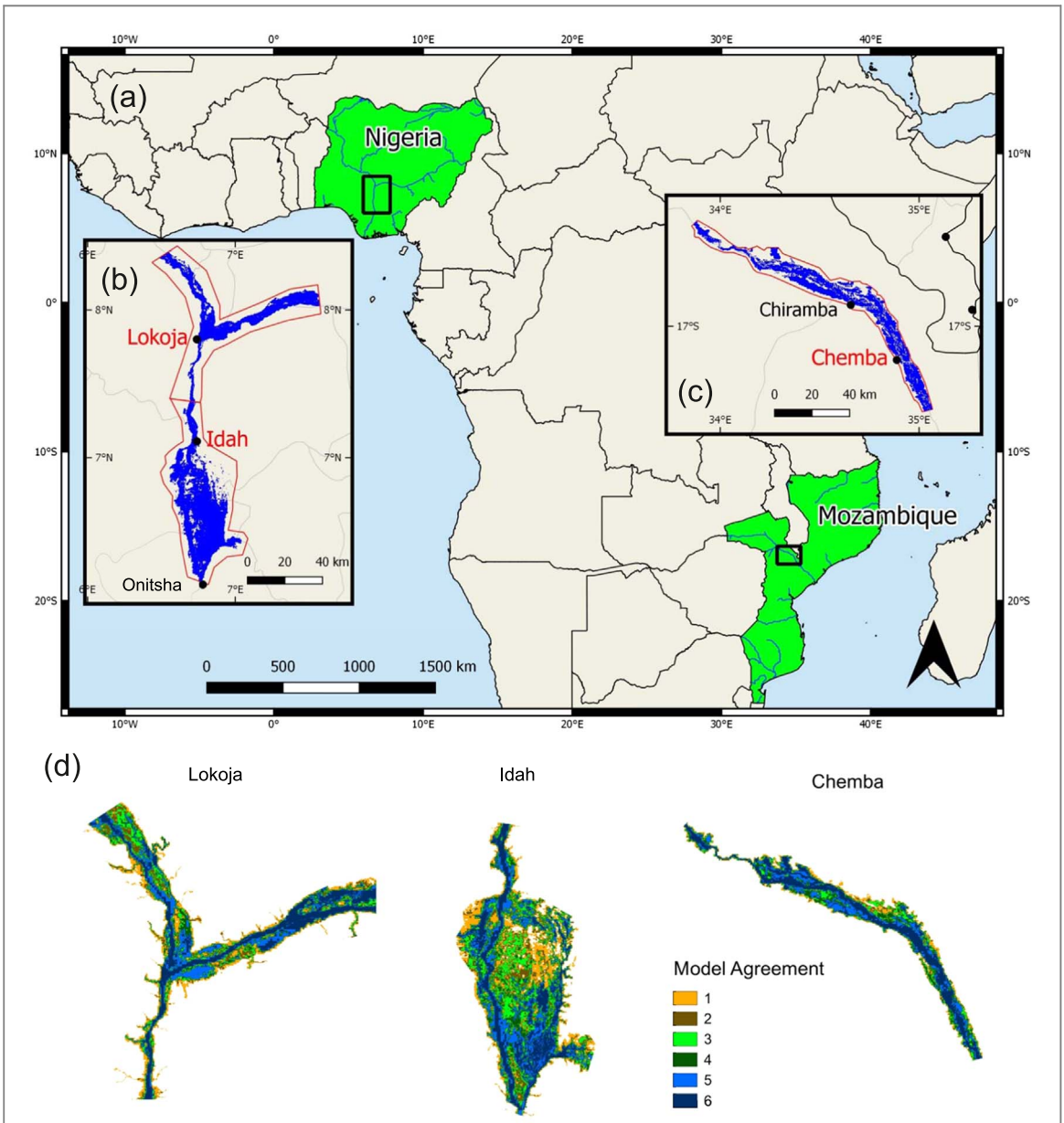

Figure 1. (a) Location of study> regions in Africa. (b) Lokoja and Idah study regions with MODIS imagery of 2012 flooding [27]. (c) Chemba study region with MODIS imagery of 2007 flooding [27]. (d) Global flood model aggregated fluvial flood extent output (25 year return period) for each region where the cell colour represents the number of models that predict it will flood in the corresponding cell [26].

potential for using multiple model combinations for flood extent prediction [24, 26].

Case study

Three hydraulically varied regions in Africa were chosen for validation: two in Nigeria and one in Mozambique (figure 1). Nigeria and Mozambique were identified in the GFMIP as countries with high exposure to flooding [24]. An important factor in the choice of study regions was the size of the river. All the reaches contained rivers sufficiently large that they should be accurately represented in the GFMs regardless of the model spatial resolution. Validating model performance on rivers narrower than the resolution of the coarsest GFM would produce unfair results. In addition to this, delta regions were avoided for analysis to prevent issues associated with the demarcation of fluvial and coastal flooding, the latter of which is not currently represented in the GFMs, although recently CaMa-Flood was coupled with the results of a Global Tide and Surge Model [28] to simulate the influence of tide and surge on river levels [29].

The first region in Nigeria, referred to in this study as Lokoja, is at the confluence of the Niger and Benue rivers. It is a region with narrow, confined floodplains. The second region in Nigeria, located south of Lokoja between the cities of Idah and Onitsha, is referred to as Idah in this study. The Idah region is relatively flat and contains an extensive floodplain that has a number of smaller channels and streams. Downstream of the Idah floodplain is a tectonic constricted outlet.

Located in central Mozambique, the final analysis region is referred to as Chemba and is situated in the lower Zambezi basin, upstream of the delta. The Zambezi River in the Chemba region can be classified as anabranching (more than one channel) with a very wide valley floor trough [30].

The flood events used as the benchmark validation datasets were the floods of 2007 in Mozambique and of 2012 in Nigeria. These events were chosen as they were recent large-scale disasters with good observational validation data. Torrential rain between December 2006 and February 2007, coupled with the landfall of Cyclone Favio in February 2007, caused flooding in Mozambique that affected more than 130 000 people [31]. The 2012 flooding in Nigeria was even more devastating; affecting almost four million people [32]. The floods in Nigeria were caused by heavy rainfall between July and October 2012.

Data

Flood imagery of both events was taken from the Dartmouth Flood Observatory (DFO) archive [27]. The DFO uses Moderate Resolution Image Spectroradiometer (MODIS) imagery to capture flood events globally, and stores them online in an open-access archive. Vegetation bias was determined to have a negligible effect on the MODIS flood imagery in the three study regions [33]. The Chemba region is dominated by shrubbery and grasslands, and any woodland is sparse [34] and although there are forests in both regions in Nigeria, these have not detrimentally affected the observed MODIS flood imagery. For the 2012 event in Nigeria, 45 days of imagery (15 September–29 October) were downloaded from the DFO archive and merged into one flood extent. Using over six weeks of data ensured that the entire event (maximum extent) was captured. The flood extent for the 2007 event in Mozambique was taken from a flood map image on the DFO website. The process of georeferencing the image for analysis is outlined in the supplementary material.

Both flood events had, very approximately, estimated return periods of around 50 years [35, 36]. The GFMIP compared the flood extent outputs of six return periods: 25, 100, 250, 500, and 1000 years. Not all of the individual GFMs had a 50 year return period output. Therefore, to ensure that the validation results best represent the skill of the models, two return periods were tested in the individual analysis: 25 and 100 years. For the aggregated analysis, only a 25 year return period was used. The return periods mentioned in this study, both reported and modelled, should be interpreted with an understanding of their associated uncertainties. Both events' 50 year flood return period was reported in news reports with no indication of how the value was calculated [35, 36]. Individual GFM return periods will not be consistent with one another due to the different approaches each takes to determine a given return period flood extent. Depending on the GFM model type, the climate model used, or the gauge data used, each GFM will have different estimated return period extents.

All the datasets used for validation in this study are open access, with the thought that the regions and events studied can be used for future GFM validation. The datasets are available from Research Data Leeds for academic research and education purposes (https://doi.org/10.5518/340).

Analysis

The analysis in this study was done in QGIS (v2.18). Individual GFM outputs were converted from extents with pixels indicating depth of flooding to binary (wet/dry) water masks representing only flood extent. No specific flood depth threshold was used, only the wet/dry threshold of each individual GFM output. The modelled and observed extents were then overlapped in each of the study regions. The MODIS flood imagery used in this study was obtained in ~250 m resolution. In order to preserve the detail of the highest resolution models, and because comparison needs to be carried out at the same spatial resolution, the MODIS imagery and all GFM outputs that were not previously of ~90 m resolution were resampled using the nearest neighbour method to ~90 m resolution. Because the datasets are binary, false accuracy errors associated with resampling to a higher resolution are not introduced. This is because interpolation between binary pixels during resampling does not result in new values (as is the case when resampling a continuous value dataset). Resampling may have introduced geospatial overlap errors, however, these errors occur regardless of the resolution resampled to and they are unlikely to have affected the validation results. The degree of overlap between the modelled flood extents and the observed DFO extents was calculated in terms of the number of pixels that showed model agreement, overprediction, and underprediction. Maps visualizing this overlap were produced (figure 3). The numerical data from these calculations was then used to calculate performance scores. The aggregated GFM output (figure 1(d)) was extracted in six different model agreement levels. The extents ranged from largest to smallest: from any model agreement ($\geqslant$1 models agree) to all model agreement (six models agree). Each of the six model agreement levels was converted to a binary water mask and underwent the same analysis as the individual GFMs.

The performance metrics used in the analysis of the flood models are commonly used in flood model assessments and for forecast verification in the atmospheric sciences [37]. The scores were also used by a number of GFM providers for their own in-house validation [12, 16, 37–39]. The three performance scores were chosen as their results represent the most important aspects of model performance: model fit, model bias, and the proportion of total flood captured. The

first, and most comprehensive, score is the $F^{<2>}$ score or the critical success index (CSI) [37]:

$$CSI = \frac{F_m \cap F_o}{F_m \cup F_o} \quad (1)$$

where $F_m \cap F_o$ is the intersection of the modelled and observed flood extent, or number of correct forecasts, and $F_m \bigcup F_o$ is the union of modelled and observed extent. The CSI ranges from 1 (best) to 0 (worst). The CSI has been shown to favourably bias larger floods [40]. However, because the floods compared in this study have a similar return period and because model performance is being compared within the same flood, CSI was deemed appropriate. The second score, the hit rate (HR) [37], measures the proportion of the observed flood that was captured by the model:

$$\mathrm{HR} = \frac{F_m \cap F_o}{F_o}, \quad (2)$$

where F_o is the total observed flood extent. The HR ranges from 1 (entire flood captured) to 0. The third score is the Bias score [37], which measures whether a forecast is biased towards underprediction or overprediction:

$$\mathrm{Bias} = \frac{(F_m \cap F_o) + F_m}{(F_m \cap F_o) + F_o} - 1, \quad (3)$$

where F_m is the total modelled flood extent. A Bias score of 0 indicates an unbiased model. Positive and negative bias scores indicate bias towards overprediction and underprediction respectively.

Although there are a number of other forecast verification scores that could have been used, the three performance scores chosen for this study were deemed appropriate because they do not consider the dry area in the validation regions. Performance scores such as the Pierce skill score, false alarm rate, and $F^{<1>}$ that account for dry area in their formulae are not desirable in situations where correct 'no' forecasts dominate the analysis, as would be the case for the large validation regions in this study [40].

The variation in flood hazard output between the GFMs identified in the GFMIP [24] raises the question of whether an ensemble model performs better than any individual flood model. Multiple model combinations have been used extensively in the atmospheric sciences in the form of model ensembles [41–46]. The ensemble model proposed in this study is a simple composite of the best performing individual models. In theory, this ensemble should reduce the uncertainty associated with using any individual model. Using a combination of the best performing individual models should reduce uncertainty as using multiple models with different modelling methods would negate any errors associated with a single modelling method. The best performing individual models to include in the ensemble are determined by the following ensemble score (ES):

$$\mathrm{ES} = \text{Average CSI} - |0.2 * \text{Average Bias}|. \quad (4)$$

In order to have one common ensemble model output, the average of the 25 year return period performance scores across the regions was used to determine the ES. A Bias adjustment factor of 0.2 was added to the ES to penalize for any significant bias towards overprediction or underprediction. The value of 0.2 was chosen as it was large enough to penalize for bias, but small enough that the CSI remained the most important score in the ES. The bias adjustment factor reduces the likelihood that any GFM that is heavily biased towards over or underprediction is included in the ensemble model. Excessive overprediction is especially detrimental to the ensemble model as the resulting flood footprint would be dominated by the model that tends towards overprediction. The number of individual models to include in the ensemble model was decided based on the performance scores of the different model agreement levels in the aggregated model validation.

Once the best individual models to include in the ensemble model had been determined, the ensemble model was created in QGIS by combining the flood extents of the individual models into one, binary, ensemble flood extent. The ensemble extent was then validated using the same methodology as for the individual and aggregated models.

Results and discussion

Individual models

The performance scores are represented graphically in figure 2 and the GFMs are arranged from left to right in order of resolution from coarsest to finest. The results indicate that there is a significant variation between the GFMs ability in modelling the flood events in each region. The average CSI of the GFMs range from 0.45 (GLOFRIS) to 0.70 (JRC) for a 25 year flow. To put these scores into context, CSI scores from other flood validation literature, in different validation regions, range from 0.3 to 0.9 [12, 16, 39], with >0.7 considered good and <0.5 poor.

Lokoja stands out as the region in which almost all of the models perform best. The higher CSI scores in Lokoja are likely a reflection of the region's narrow confined floodplain, and the relative simplicity of modelling the flood where extent is not sensitive to flood discharge magnitude. The increased complexity in flood modelling in flat extensive floodplains such as the one in Idah is reflected in the lower CSI scores for the region. The overlap of the observed and modelled extents (figure 3) illustrates the varied success of the GFMs at modelling floodplain inundation in Idah.

GLOFRIS, which uses a simple flood volume distribution method for modelling inundation, had the lowest average CSI score across the three regions and

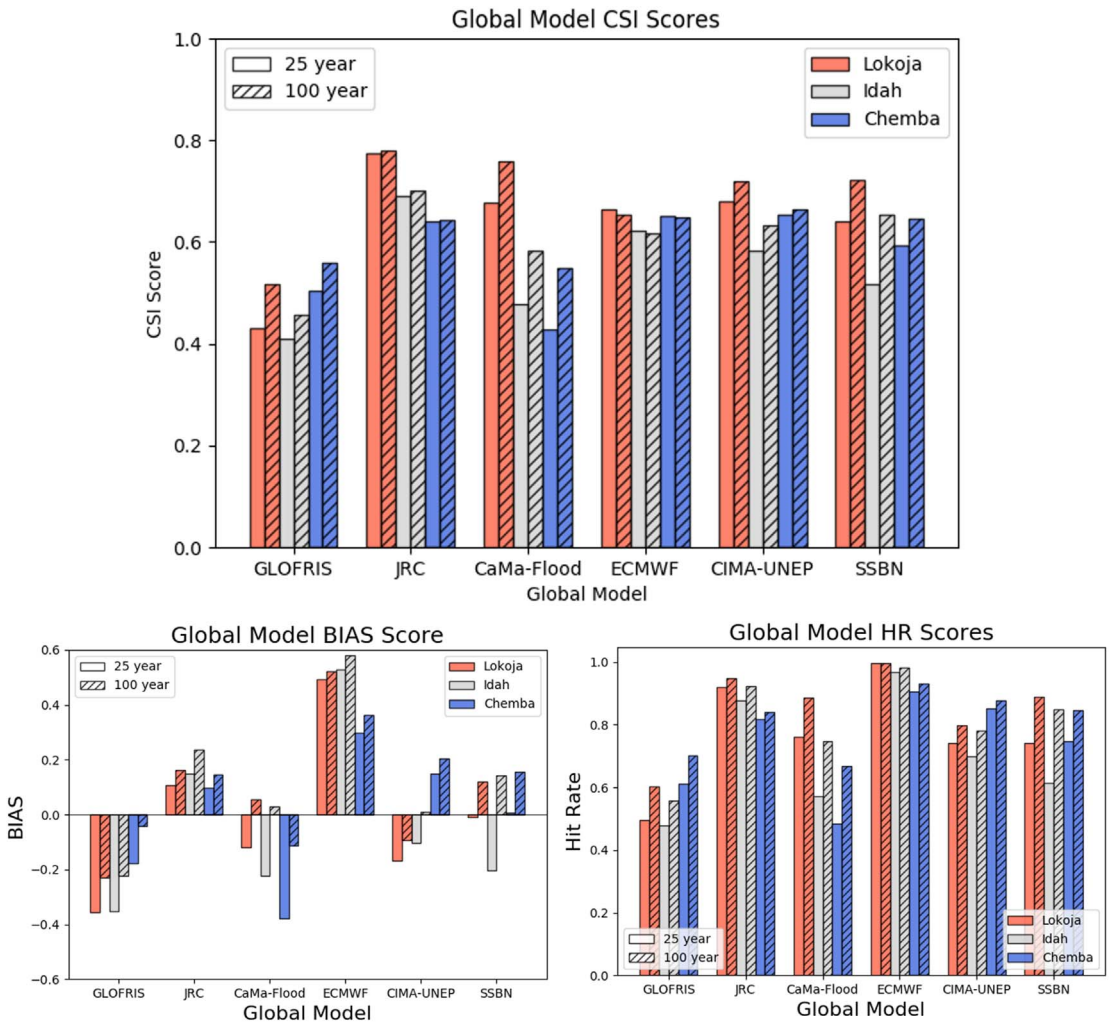

Figure 2. Individual global flood model (GFM) critical success index (CSI), Bias, and hit rate (HR) performance scores when compared against observed events in Lokoja, Idah, and Chemba. Results shown for 25 and 100 year return period GFM output.

showed very large regions of underprediction in Idah. The other 2D models, which have a more hydrodynamic flood modelling scheme, scored better across the three regions. This could be due to a more accurate representation of the physics of floodplain flow or a better characterization of the river floodplain. This is evident in Idah, where CaMa-Flood, SSBN, and JRC performed better, possibly due to the greater connectivity modelled within the floodplain by their native sub-grid models. Although implementing similar schemes, the subtleties of their 2D model structures differ. This could explain why the JRC model had higher performance scores across the three regions. The benefits of CIMA-UNEP's simpler 1D cross-section approach to modelling floodplains proved successful at modelling much of the central floodplain missed by GFMs as the 1D section implicitly connects low areas along the cross-section. However, this can also lead to overprediction if the 1D approach models inundation in low lying floodplain areas with no connectivity to the channel.

The GFM with one of the highest CSI scores in Chemba is ECMWF, whereas the GFM with the lowest CSI score in Chemba is CaMa-Flood. This highlights the importance of input flow in GFM performance: CaMa-Flood and ECMWF share the same core hydrodynamic model, but differ in their flow generation model. The performance of CaMa-Flood also significantly improves as the modelled return period is increased from 25 to 100 years. This suggests that the input flow was the limiting factor affecting the performance scores of the 25 year output. Apart from ECMWF, increasing the return period from 25 to 100 years generally increased the CSI scores of the GFMs. Increasing the GFM return period resulted in averaged CSI percentage increases of 14% (GLOFRIS), 0.1% (JRC), 5% (CIMA-UNEP), 19% (CaMa-Flood), and 15% (SSBN). These findings show that in these three study areas, GLOFRIS, CaMa-Flood and SSBN are sensitive to input flow. However, the level of return period sensitivity could be exaggerated by the fact that these three models all showed higher bias towards underprediction at the 25 year return period than the rest of the models. Increasing the return period of an underpredicting and an unbiased flood model would likely result in a comparatively greater proportion of additional flooding being captured by the underpredicting model at the higher return period, thus leading to a larger increase in CSI. JRC

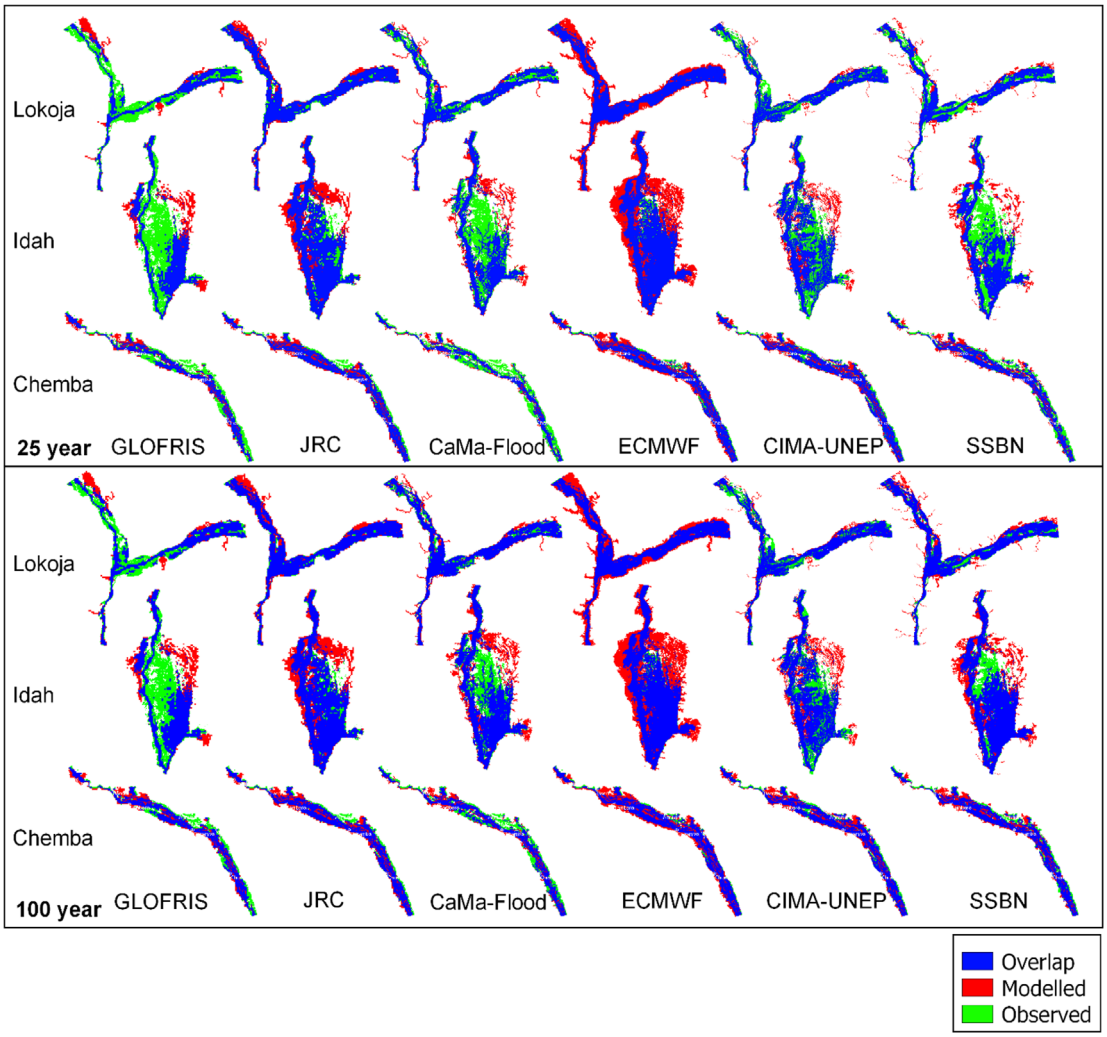

Figure 3. Overlap of individual global flood model extent for return period flows of 25 and 100 years and MODIS observed flood extent for Lokoja, Idah, and Chemba.

continues to perform the best of all the models at either return period when averaged across the three regions. The variation in input flow is reflected in the HR and BIAS scores of the GFMs. Averaged across all three regions, ECMWF captures almost all the flooding, with an HR of 0.96 for a 25 year flood. However, it is the GFM that showed the largest bias in either direction: 0.44 for the 25 year return period and 0.49 for the 100 year return period. These results suggest that ECMWF is significantly overestimating input flow at both return periods.

The differences in model forcing (climate reanalysis data versus gauged discharge data) is apparent in the bias scores of the GFMs. CIMA-UNEP and SSBN, both based on gauged discharge data, show an average bias towards underprediction at the 25 year return period and an average bias towards overprediction at the 100 year return period. This suggests that both gauge forced models are doing a good job at estimating the reported 50 year return period of the observed flooding. Three of the four models forced by climate reanalysis data show bias in only one direction at both return periods, suggesting that the climate forced models have greater difficulty predicting a representative return period. This could be due to the fact that the validation regions are in the tropics and reanalysis datasets have been found to poorly represent precipitation in the tropics [47]. However, caution should be taken before drawing general conclusions because input flow is not the only parameter influencing floodplain extent (for instance, poorly represented floodplain connectivity might cause a systematic estimation bias on flood extent).

The improved connectivity offered by higher spatial resolution GFMs is evident in the Idah floodplain (figure 3). CIMA-UNEP and SSBN, both with outputs of 90 m resolution at the equator, are able to model some of the smaller channels within the floodplain (either implicitly or explicitly). Despite the improved connectivity representation, there is no discernible correlation between the performance scores and GFM spatial resolution, indicating that the models still need further improvements in capturing river/floodplain connectivity. At present, there is currently no

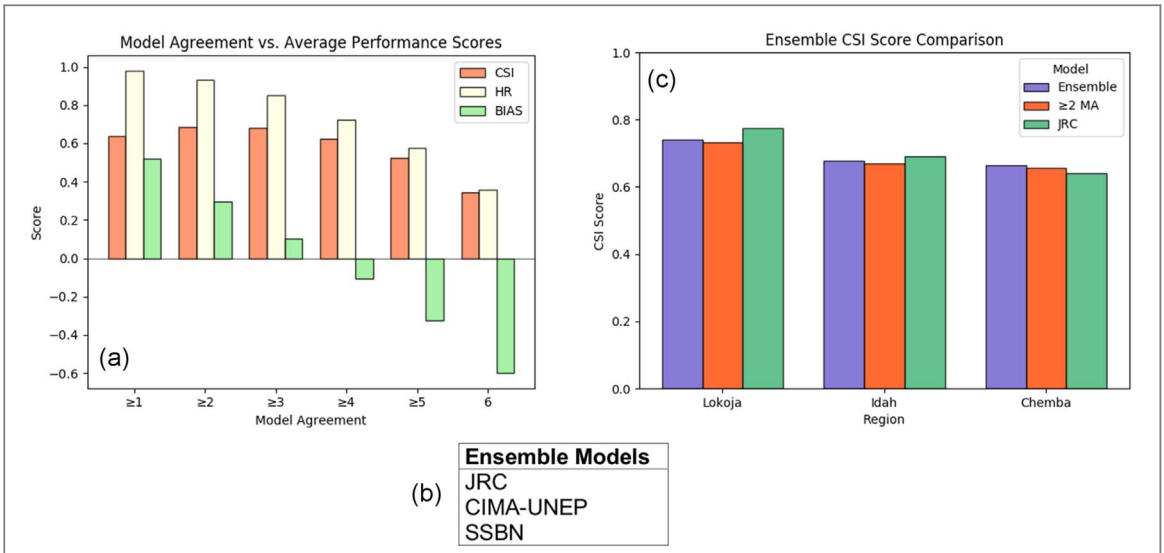

Figure 4. (a) Critical success index (CSI), hit rate (HR), and Bias scores averaged across all three study regions for the six different levels of aggregated model agreement for a 25 year return period flood. (b) The three best performing individual global flood models (GFMs) that are included in the ensemble model as determined by the ensemble score. (c) CSI scores for the composite model, the best performing individual GFM (JRC), and the best performing aggregated model agreement level (⩾2 Model Agreement) for a 25 year return period flood.

well-developed method to represent channel bifurcation in 1D fluvial models. A better representation of bifurcation would improve the performance of both 1D and 2D sub-grid models in areas of high bifurcation, such as floodplains [48].

The comparative usefulness of GFMs and regional flood models is a point of contention in flood modelling literature [18, 49]. Thomas [50] developed a regional flood model for southern Nigeria and validated it against the 2012 floods. The model incorporated local bathymetric and hydrographic data. When compared with MODIS data of the flood event, the regional model's CSI scores were 0.73 and 0.53 for Lokoja and Idah, respectively. Comparison with the best GFM performance scores show that JRC and CaMa-Flood outperform the regional model in Lokoja with CSIs of 0.78 and 0.75, respectively. The case for the GFMs is even stronger in Idah as five GFMs outperform the regional model: JRC, SSBN, CaMa-Flood, ECMWF and CIMA-UNEP with CSI scores of 0.70, 0.65, 0.58, 0.62 and 0.58 respectively. Comparison of performance scores between the studies should be approached with some caution as the analysis areas in Thomas' [50] study varied slightly compared with the ones used in this study. However, in the cases shown here, the performance of GFMs is comparable to, or in some cases better than, the performance of a locally calibrated regional model.

Aggregated model

The performance scores of the different levels of model agreement for the 25 year return period aggregated model (figure 4(a)) show that the CSI peaks at ⩾2 and ⩾3 model agreement. These results correspond with the results of the individual model validation: two or three models consistently outperform the rest. A HR of 0.36 at 6 model agreement shows that all six models are correctly capturing at least 36% of the observed flood events. The bias trends steadily from overprediction to underprediction as the model agreement level increases. The least bias in either direction occurs at ⩾3 and ⩾4 model agreement, this is likely due to the fact that the opposite bias of the individual models shown in figure 2 balanced one another out.

Ensemble model

The aggregated model validation found that the ⩾2 and ⩾3 model agreement groups had the highest CSI scores. As a result, the number of models chosen to include in the ensemble model was three. The individual models included in the ensemble model, chosen using the ES, were JRC, CIMA-UNEP, and SSBN (figure 4(b)). The validation performance scores of the ensemble model are compared (figure 4(c)) with the best performing models from the individual and aggregate group: JRC and ⩾2 model agreement. The results show that there is little difference between the CSI scores of the ensemble model, JRC, and ⩾2 model agreement. Furthermore, the JRC GFM scores higher than the ensemble model in Lokoja and Idah. The aim of an ensemble model is to reduce the uncertainty associated with using a single model. For an ensemble model to perform better than individual models, the individual models that make up the ensemble model need to compensate for the uncertainty in the other models either through different input data or different modelling methods. Judging from the results of the analysis, it seems that the combination of individual models did not improve the results as a whole. If anything, they added to the uncertainty in the form of increased overprediction, which resulted in the reduced CSI scores. Although the ensemble model did

not outperform the best individual model, it did score comparably well. There are situations where this ensemble approach could be of use. For example, in regions where it is not possible to validate flood models to determine the best individual model, the use of a multiple model ensemble could reduce the uncertainty associated with using only one model, whilst not significantly reducing the flood extent prediction accuracy.

Observational data

It is imprudent to discuss our validation findings without making some reference to the observational data used and the inherent uncertainty that is associated with flood observation mapping. This study used extents from the DFO archive, which is currently the most extensive global flood database. However, work is being done to develop a global database of historic flood events in Google Earth Engine (GEE) [51, 52]. The DFO flood extents used in this study and the equivalent extents from the new GEE global database [52] were analysed to examine the agreement between the two data sources. The results of the analysis show that there is 12% disagreement in Lokoja (CSI 0.88), 11% disagreement in Idah (CSI 0.89), and 63% disagreement in Chemba (CSI 0.37) between the observed flood extents from the two data sources. The bias scores are also always in the direction of the DFO extents (Lokoja 0.02, Idah 0.01, and Chemba 1.32) indicating that the DFO extents are larger. Figures showing the observational agreement and disagreement are included in the supplementary material. This observational disagreement between data sources highlights an underlying problem with flood mapping. Satellite imagery, both optical and radar, faces issues with observational bias. Optical imagery is affected by cloud cover and radar imagery is affected by vegetation. Data sources differ in the methods they use to reduce the effects of such observational bias. As a result, flood maps for the same event can differ if they are obtained from different sources. Neither source captures all of the flooding; each misses different parts. The task faced by the end user when confronted with the uncertainty associated with two disagreeing datasets is to decide which most closely represents the actual event. Even then, the chosen extent is used under the assumption that it is entirely correct. If these observational uncertainties could be incorporated into flood maps, it would allow for a measure of confidence to be calculated relating to the accuracy of the observations and as a result, the accuracy of the validation findings.

Conclusions

This paper has outlined the first validation intercomparison between GFMs. Validation of the individual models against observed events in Nigeria and Mozambique showed that there is a significant variation in GFM performance, with average CSI scores ranging from 0.45 to 0.7. Site specific conditions played an important role in model performance. The GFMs scored well in Lokoja, where flood extents were restricted by a confined floodplain. Conversely, the models showed less skill in Idah, a flat extensive floodplain with complex morphology. The underlying hydraulic models showed varied success in modelling floodplain inundation. CIMA-UNEP's 1D approach was able to implicitly model greater connectivity within the Idah floodplain. Generally however, the connectivity provided by 2D models was evident in both the performance scores and the inundation maps. 2D hydrodynamic models showed significantly more skill at predicting inundation than 2D volume redistribution methods. Input flow was identified as a crucial factor in modelling a representative flood inundation extent and increasing the return period of the GFMs resulted in significant improvements for half of the GFMs. The GFMs forced by gauged data showed better return period accuracy than those forced by climate reanalysis data. This was attributed to the poor reanalysis representation of precipitation in the tropics. Spatial resolution, although showing some improvement in floodplain connectivity, did not obviously improve model performance.

Comparison of the GFMs with a regional flood model developed for Nigeria showed that some of the GFMs outperformed the regional model. Through validation, the three best models were identified and combined into a composite model. The validation of the composite model, showed that it performed similarly, but not better than the best individual GFM.

Outlook

This study has demonstrated the usefulness of a collective GFM validation procedure. The comparisons and conclusions that can be drawn from the common validation data cannot be made using the individual internal GFM validation data that has been available thus far. The focus area of this study has been limited to three regions in Africa and has looked only at flood extents. The GFMs tested in this study have a multitude of uses beyond only flood extent mapping. These include, but are not limited to: flood forecasting, estimating future impacts, and real time disaster response. Going further, a more extensive validation procedure that incorporates a comparison of flow velocity [53], inundated depth, and flood duration [54] would allow more conclusions to be drawn about both the performance and different uses of the models. The validation also needs to be extended across different climates and continents. To do this, a catalogue of appropriate validation regions needs to be developed and the observational data used for validation needs to be shared openly. Future studies should

also incorporate more GFMs such as insurance catastrophe models to encourage the knowledge transfer between research and industry. Incorporating advanced methods of model output validation and applying them across more regions would allow for a truly global validation comparison study of GFMs.

Acknowledgments

The work in this paper was in part supported by UK National Environmental Research Council grant NE/R008949/1 and iCASE funding from Fathom Global. PJW received funding from the Netherlands Organisation for Scientific Research (NWO) VIDI grant VIDI 016.161.324. We thank the members of the Global Flood Partnership who provided feedback on the research at the GFP workshop in Delft, 2018. We also thank Beth Tellman and Jonathan Sullivan of Cloud to Street for the flood extents from their global database.

ORCID iDs

Mark V Bernhofen https://orcid.org/0000-0002-4919-0111
Mark A Trigg https://orcid.org/0000-0002-8412-9332
Philip J Ward https://orcid.org/0000-0001-7702-7859

References

[1] Berz G, Kron W, Loster T, Rauch E, Schimetschek J, Schmieder J, Siebert A, Smolka A and Wirtz A 2001 World map of natural hazards—a global view of the distribution and intensity of significant exposures *Nat. Hazards* **23** 443–65
[2] Wallemacq P, Guha-Sapir D, McClean D and CREDUNISDR 2015 The Human Cost of Weather Related Disasters: 1995 - 2015 https://doi.org/10.13140/rg.2.2.17677.33769
[3] Kummu M, de Moel H, Ward P J and Varis O 2011 How close do we live to water? A Global analysis of population distance to freshwater bodies *PLoS One* **6** 13
[4] Jongman B, Ward P J and Aerts J 2012 Global exposure to river and coastal flooding: long term trends and changes *Glob. Environ. Change-Human Policy Dimens.* **22** 823–35
[5] Sherwood S C, Bony S and Dufresne J L 2014 Spread in model climate sensitivity traced to atmospheric convective mixing *Nature* **505** 37
[6] Alfieri L, Bisselink B, Dottori F, Naumann G, de Roo A, Salamon P, Wyser K and Feyen L 2017 Global projections of river flood risk in a warmer world *Earth Future* **5** 171–82
[7] Desai B, Maskrey A, Peduzzi P, De Bono A and Herold C 2015 Making development sustainable: the future of disaster risk management *Global Assessment Report on Disaster Risk Reduction* United Nations Office for Disaster Risk Reduction https://archive-ouverte.unige.ch/unige:78299
[8] United Nations General Assembly 2015 Transforming our World: The 2030 Agenda for Sustainable Development http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E
[9] United Nations International Strategy for Disaster Reduction 2015 Sendai Framework for Disaster Risk Reduction 2015–2030 http://www.wcdrr.org/uploads/Sendai_Framework_for_Disaster_Risk_Reduction_2015-2030.pdf
[10] Wood E F *et al* 2011 Hyperresolution global land surface modeling: meeting a grand challenge for monitoring Earth's terrestrial water *Water Resour. Res.* **47** 10
[11] Michel G (ed) 2018 *Risk Modeling for Hazards and Disasters* (Amsterdam: Elsevier) pp xi–xii
[12] 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
[13] Rudari R, Silvestro F, Campo L, Rebora N, Boni G and Herold C 2015 *Improvement of the global flood model for the GAR 2015* United Nations Office for Disaster Risk Reduction http://www.preventionweb.net/english/hyogo/gar/2015/en/bgdocs/risk-section/CIMA%20Foundation,%20Improvement%20of%20the%20Global%20Flood%20Model%20for%20the%20GAR15.pdf
[14] Pappenberger F, Dutra E, Wetterhall F and Cloke H L 2012 Deriving global flood hazard maps of fluvial floods through a physical model cascade *Hydrol. Earth Syst. Sci.* **16** 4143–56
[15] 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** 21
[16] Sampson C C, Smith A M, Bates P B, Neal J C, Alfieri L and Freer J E 2015 A high-resolution global flood hazard model *Water Resour. Res.* **51** 7358–81
[17] Ward P J, Jongman B, Weiland F S, Bouwman A, van Beek R, Bierkens M F P, Ligtvoet W and Winsemius H C 2013 Assessing flood risk at the global scale: model setup, results, and sensitivity *Environ. Res. Lett.* **8** 10
[18] Ward P J *et al* 2015 Usefulness and limitations of global flood risk models *Nat. Clim. Change* **5** 712–5
[19] Ward P J *et al* 2017 A global framework for future costs and benefits of river-flood protection in urban areas *Nat. Clim. Change* **7** 642
[20] 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
[21] Winsemius H C *et al* 2016 Global drivers of future river flood risk *Nat. Clim. Change* **6** 381–5
[22] Winsemius H C, Van Beek L P H, Jongman B, Ward P J and Bouwman A 2013 A framework for global river flood risk assessments *Hydrol. Earth Syst. Sci.* **17** 1871–92
[23] Groeve T D *et al* 2015 Joining forces in a global flood partnership *Bull. Am. Meteorol. Soc.* **96** ES97–100
[24] Trigg M A *et al* 2016 The credibility challenge for global fluvial flood risk analysis *Environ. Res. Lett.* **11** 10
[25] Hoch J M, Neal J C, Baart F, van Beek R, Winsemius H C, Bates P D and Bierkens M F P 2017 GLOFRIM v1.0-A globally applicable computational framework for integrated hydrological-hydrodynamic modelling *Geosci. Model Dev.* **10** 3913–29
[26] Trigg M A *et al* 2016 *Aggregated fluvial flood hazard output for six Global Flood Models for the African Continent* University of Leeds (https://doi.org/10.5518/96)
[27] Brakenridge G R *Global Active Archive of Large Flood Events* Dartmouth Flood Observatory, University of Colorado http://floodobservatory.colorado.edu/Archives/index.html
[28] Muis S, Verlaan M, Winsemius H C, Aerts J and Ward P J 2016 A global reanalysis of storm surges and extreme sea levels *Nat. Commun.* **7** 11
[29] Ikeuchi H, Hirabayashi Y, Yamazaki D, Muis S, Ward P J, Winsemius H C, Verlaan M and Kanae S 2017 Compound simulation of fluvial floods and storm surges in a global coupled river-coast flood model: model development and its application to 2007 Cyclone Sidr in Bangladesh *J. Adv. Model. Earth Syst.* **9** 1847–62
[30] Davies B R, Beilfuss R D and Thoms M C 2000 Cahora Bassa retrospective, 1974–1997: effects of flow regulation on the Lower Zambezi River *SIL Proc., 1922-2010* (https://doi.org/10.1080/03680770.1998.11901620)
[31] Rana R 2007 *A Review of the Mozambique Floods Response Shelter Working Group* International Federation of Red Cross and Red Crescent Societies p 43 https://www.alnap.org/system/files/content/resource/files/main/mozambique-shelter-review.pdf

[32] Federal Government of Nigeria 2013 *Nigeria: Post-Disaster Needs Assessment 2012 Floods* Global Facility for Disaster Reduction and Recovery p 154 https://www.gfdrr.org/sites/gfdrr/files/NIGERIA_PDNA_PRINT_05_29_2013_WEB.pdf
[33] Nigro J, Slayback D, Policelli F and Brakenridge G R 2014 NASA/DFO MODIS near real-time (NRT) global flood mapping product evaluation of flood and permanent water detection https://floodmap.modaps.eosdis.nasa.gov/documents/NASAGlobalNRTEvaluationSummary_v4.pdf
[34] CES 2014 Resettlement Plan *Ecofarm Irrigation and Organic Sugarcane Project Mozambique* Grahamstown: Coastal & Environmental Services http://www.cesnet.co.za/pubdocs/Eco%20Farm%20ESIA%20Addendum%20Sugar%20Mill%20TKBR%2017.12.15_136/Volume%205%20-%20Ecofarm%20Resettlement%20Plan.pdf
[35] BBC 2007 Mozambique seeks urgent flood aid http://news.bbc.co.uk/2/hi/africa/6361957.stm
[36] Reuters 2012 Nigeria floods kill 363 people, displace 2.1 mln - agency https://in.reuters.com/article/nigeria-floods/nigeria-floods-kill-363-people-displace-2-1-mln-agency-idINDEE8A40EH20121105
[37] Wilks D 2006 *Statistical Methods in the Atmospheric Sciences* (United States of America: Elsevier)
[38] Alfieri L, Salamon P, Bianchi A, Neal J, Bates P and Feyen L 2014 Advances in pan-European flood hazard mapping *Hydrol. Process.* **28** 4067–77
[39] Wing O E J, Bates P D, Sampson C C, Smith A M, Johnson K A and Erickson T A 2017 Validation of a 30 m resolution flood hazard model of the conterminous United States *Water Resour. Res.* **53** 7968–86
[40] Stephens E, Schumann G and Bates P 2014 Problems with binary pattern measures for flood model evaluation *Hydrol. Process.* **28** 4928–37
[41] Ehrendorfer M 1997 Predicting the uncertainty of numerical weather forecasts: a review *Meteorol. Z.* **6** 147–83
[42] Leith C E 1974 Theoretical skill of Monte-Carlo forecasts *Mon. Weather Rev.* **102** 409–18
[43] Demeritt D, Cloke H, Pappenberger F, Thielen J, Bartholmes J and Ramos M 2007 Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting *Environ. Hazards* **7** 115–27
[44] Schellekens J *et al* 2017 A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset *Earth Syst. Sci. Data* **9** 389–413
[45] Siqueira V A, Collischonn W, Fan F M and Chou S C 2016 Ensemble flood forecasting based on operational forecasts of the regional Eta EPS in the Taquari-Antas basin *RBRH-Rev. Bras. Recur. Hidr.* **21** 16
[46] Gneiting T and Raftery A E 2005 Atmospheric science—weather forecasting with ensemble methods *Science* **310** 248–9
[47] Beck H E, Vergopolan N, Pan M, Levizzani V, van Dijk A, Weedon G P, Brocca L, Pappenberger F, Huffman G J and Wood E F 2017 Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling *Hydrol. Earth Syst. Sci.* **21** 6201–17
[48] Mateo C M R, Yamazaki D, Kim H, Champathong A, Vaze J and Oki T 2017 Impacts of spatial resolution and representation of flow connectivity on large-scale simulation of floods *Hydrol. Earth Syst. Sci.* **21** 5143–63
[49] World Bank 2014 *Understanding Risk in an Evolving World* (Washington, DC: World Bank Group) p 220
[50] Ekeu-Wei I and Blackburn A 2018 *Application of Open-access and 3rd Party Geospatial Technology for Integrated Flood Risk Management in Data Sparse Regions of Developing Countries PhD Thesis* Lancaster University http://eprints.lancs.ac.uk/89716/
[51] Tellman B, Sullivan J, Doyle C, Kettner A, Brakenridge G R, Erickson T and Slayback D 2017 *A Global Geospatial Database of 5000 + Historic Flood Event Extents (New Orleans)* (American Geophysical Union)
[52] Tellman B, Sullivan J A, Kuhn C, Kettner A J, Doyle C S, Brakenridge G R, Eriksen T and Slayback D 2018 A global inventory of flood events and exposure trends in preparation
[53] Kreibich H, Piroth K, Seifert I, Maiwald H, Kunert U, Schwarz J, Merz B and Thieken A 2009 Is flow velocity a significant parameter in flood damage modelling? *Nat. Hazards Earth Syst. Sci.* **9** 1679–92
[54] Dang N, Babel M and Luong H 2011 Evaluation of food risk parameters in the Day River Flood Diversion Area, Red River Delta, Vietnam *Nat. Hazards* **56** 169