Uncertainty in the extreme flood magnitude estimates of large-scale flood hazard models

Laura Devitt 1, Jeffrey Neal 1,3, Thorsten Wagener 2,3,4 and Gemma Coxon 1,3

1 School of Geographical Sciences, University of Bristol, Bristol, United Kingdom
2 Department of Civil Engineering, University of Bristol, Bristol, United Kingdom
3 Cabot Institute, University of Bristol, Bristol, United Kingdom
4 Institute for Environmental Science and Geography, University of Potsdam, Potsdam, Germany

E-mail: laura.devitt@bristol.ac.uk

Keywords: large-scale flood hazard models, global hydrological model, flood risk

Abstract
The growing worldwide impact of flood events has motivated the development and application of global flood hazard models (GFHMs). These models have become useful tools for flood risk assessment and management, especially in regions where little local hazard information is available. One of the key uncertainties associated with GFHMs is the estimation of extreme flood magnitudes to generate flood hazard maps. In this study, the 1-in-100 year flood (Q100) magnitude was estimated using flow outputs from four global hydrological models (GHMs) and two global flood frequency analysis datasets for 1350 gauges across the conterminous US. The annual maximum flows of the observed and modelled timeseries of streamflow were bootstrapped to evaluate the sensitivity of the underlying data to extrapolation. Results show that there are clear spatial patterns of bias associated with each method. GHMs show a general tendency to overpredict Western US gauges and underpredict Eastern US gauges. The GloFAS and HYPE models underpredict Q100 by more than 25% in 68% and 52% of gauges, respectively. The PCR-GLOBWB and CaMa-Flood models overestimate Q100 by more than 25% at 60% and 65% of gauges in West and Central US, respectively. The global frequency analysis datasets have spatial variabilities that differ from the GHMs. We found that river basin area and topographic elevation explain some of the spatial variability in predictive performance found in this study. However, there is no single model or method that performs best everywhere, and therefore we recommend a weighted ensemble of predictions of extreme flood magnitudes should be used for large-scale flood hazard assessment.

1. Introduction
The vast worldwide impact of flood events, affecting 2.3 billion people and causing more than 662 US dollars in economic damage between 1995 and 2015 alone (CRED and UNISDR 2015), has motivated the development and application of global flood hazard models (GFHMs). GFHMs provide an understanding of the spatial extent and frequency of flood hazard at large spatial scales and can identify flood-prone areas in ungauged basins. They are increasingly being used for international disaster risk management and have been applied by the World Bank and the Global Facility for Disaster Reduction and Recovery (Ward et al 2015). These models have been successfully applied for large-scale flood mapping largely due to the development of new remotely sensed data sets, computing power and automated modelling frameworks (e.g. Yamazaki et al 2011, Pappenberger et al 2012, Winsenius et al 2013, Sampson et al 2015, Dottori et al 2016). However, the uncertainty in GFHM predictions is largely unknown.

One of the key problems for the hydrologic and hydraulic research community is producing reliable estimates of extreme flows in ungauged basins (Böschs et al 2013). Fundamentally, all GFHM frameworks combine a methodology that generates design flood magnitudes with a way to distribute
these flows over the floodplain (ranging from a simple volume filling method to two-dimensional hydrodynamic models). There are two key approaches for generating flood magnitudes (Trigg et al. 2016). These have been developed to account for the global scale of analysis and the need for making predictions in regions of the world that are data limited:

(a) use of continuous timeseries of simulated streamflow from a global hydrological model (GHM). Return period flows along the river network are derived by an at-site flood frequency analysis of the simulated discharge (e.g. Yamazaki et al. 2011, Winsemius et al. 2013, Dottori et al. 2016).

(b) Statistical methods using observed gauged discharge and regionalised flood frequency analysis (RFFA) techniques (e.g. Meigh et al. 1997, Merz and Blöschl 2009a), to estimate return period flows along the global river network (e.g. Smith et al. 2015, UNISDR 2015, Zhao et al. 2020 in review).

Each method inherently has its own limitations and uncertainties. When using a modelling approach there are large uncertainties associated with the representation of meteorological extremes in globally available satellite-derived data products (e.g. Kidd et al. 2012, Chen et al. 2014, Kendon et al. 2017, Beck et al. 2017a), combined with the difficulty of defining model structure and calibration of parameters in ungauged basins across global domains (Wagener and Wheater 2006, Beven and Cloke 2012, Blöschl et al. 2013, Archfield et al. 2015). For the RFFA approach, methods are often limited by the length of available observational flow timeseries and therefore statistical models are used for extrapolation in order to estimate the probability of yet unobserved extremes (Michele and Rosso 2001, Sousa et al. 2011). Additionally, RFFA methods applied at the global scale are limited by their data requirements, and mostly use basic basin characteristics to cluster catchments or build regression models (Smith et al. 2015). Common to both methods is the uncertainty of extreme discharge measurements (e.g. Di Baldassarre and Montanari 2009, Coxon et al. 2015), which are rarely accounted for when using extreme discharge measurements to calibrate parameters of GHMs (e.g. Döll et al. 2003, Beck et al. 2016), or to train the statistical models used by the RFFA methods (Lang et al. 2010, Neppel et al. 2010, Steinbakk et al. 2016). Previous studies have suggested differences in extreme flows are a key source of differences in flood hazard between models (Trigg et al. 2016, Bernhofen et al. 2018, Aerts et al. 2020), with calls for a framework to compare GFHMs at component level to detect causes for model differences (Hoch and Trigg 2019). However, this previous work has focused on the deterministic final outputs from a number of GFHMs which has not allowed for direct attribution of the differences between these models to any specific factors. Previous studies comparing GHMs have shown that there is variability in performance of model simulations of peak flows in space (Towner et al. 2019), the effect of routing scheme implemented can be critical in the simulation of peak discharge magnitudes (Zhao et al. 2017), that in many environments the uncertainty from input data quality is of a similar magnitude of model structure uncertainty (Marthens et al. 2020), and there is significant uncertainty in global scale hydrological modelling from precipitation data errors (Sperna Weiland et al. 2015). Studies comparing RFFA methods have also shown that performance of predicting flood magnitudes tends to have a correlation with climatological factors, i.e. performance tends to be highest in humid and lowest in arid catchments, and catchment size (Salinas et al. 2013), and that regional regressions that divide a domain into subregions and apply regression models separately always perform better than ek et al (2014).

Quantifying extreme flood estimates (such as a 1 in 100 year flood event) to drive GFHMs is an unresolved problem due to poorly understood and quantified uncertainties in flood magnitude estimation methods (Engeland et al. 2005, Read and Vogel 2015, Halbert et al. 2016). At present, the typical approach to developing a GFHM is to use the simulations of a single GHM or outputs from a single RFFA (e.g. Pappenberger et al. 2012, Winsemius et al. 2013, Sampson et al. 2015, Dottori et al. 2016). These are used for extreme flow and river conveyance estimation. However, there is little justification for the choice of methodology used by the GFHM groups. With GFHMs increasingly being used to inform national and international flood risk management strategies in regions where no local information exists (Ward et al. 2015), it is important that we understand the hydrological uncertainties that are associated with the estimation of flood magnitudes, and this information is often not included in policy making decision processes. Current research in model-intercomparison has focused on assessing the performance of simulated discharge timeseries, while benchmarking of GFHMs have focused on comparing flood hazard maps (Trigg et al. 2016, Bernhofen et al. 2018, Aerts et al. 2020). However, thus far there are no large-scale studies that have assessed and compared the methods used for estimating extreme flood magnitudes in the context of use in large-scale flood hazard mapping. Our aim is to evaluate the hydrological uncertainties associated with these methods. Therefore, this paper will explore:

(a) how well do RFFA and GHMs reproduce observed extreme flood magnitudes?
What controls the spatial variability in prediction performance?

Do ensemble predictions perform better than any individual prediction?

2. Methods

The conterminous United States was chosen as the study location. This is due to the availability of observed data and the wide variety of physiographical and climatological conditions, allowing for potential transfer of knowledge to similar regions outside of this case study.

2.1. Data

2.1.1. Observed

Daily river flow time series were extracted from the USGS website https://waterdata.usgs.gov/nwis (R-package dataRetrieval; de Cicco et al 2018) for gauging stations with an upstream catchment area of greater than 1000 km², and that were operational during the period 1980–2014. These criteria were chosen due to the spatial and temporal resolution of the GHMs used in this study (see table 1). Any station with >5% missing data were removed, leaving 1784 gauging stations for analysis.

2.1.2. GHMs

Four GHMs composed of different meteorological datasets, hydrological models/LSMs, and river routing models are used to each simulate global river discharge (details are summarised in table 1). GloFAS (Harrigan et al 2020) couples runoff from the LSM HTESS (Balsamo et al 2009) with the Lisflood (Burek et al 2013) hydrological and channel routing model (van der Knijff et al 2010), forced with ERA5 reanalysis data (Hersbach et al 2020). Lisflood is calibrated against daily river discharge from the GRDC database (Hirpa et al 2018). PCR-GLOBWB (Sutajudaja et al 2018) is a grid-based GHM which is forced with downscaled CRU TS 2.1 datasets (Harris et al 2014) to daily values assistance of ERA-Interim reanalysis (Dee et al 2011). The model was run with a default parameterisation and is uncalibrated. CaMa-Flood (Yamazaki et al 2011) is a hydrodynamic global river routing model. CaMa-Flood v3.95 was run, uncalibrated, forced with daily run-off product calculated by HTESS (Balsamo et al 2009) for the earth2Observe project (Schellekens et al 2017). All models within this project were forced using the WATCH forcing data methodology applied to ERA-Interim reanalysis (WFEDt; Weedon et al 2014). These models were chosen as they have previously been integrated in GFHM frameworks (see for example Pappenberger et al 2012, Winsemius et al 2013, Dottori et al 2016).

We also included WorldWideHYPE (Arheimer et al 2020) in our study, which is the first attempt of applying a catchment model at the global scale and was chosen here as a comparison with the ‘traditional’ global gridded approach. WorldWideHYPE is discretised into hydrological response units and is developed from the catchment-level upwards, calibrating parameters in a stepwise approach for groups of parameters regulating specific processes and catchment characteristics. Although this model has not been integrated into any of the GFHM frameworks, the HYPE catchment model (Lindström et al 2010) is used operationally in Sweden for flood forecasting, and therefore this global version of the model could provide the global flood hazard modelling community with a potentially useful dataset.

The model data used here is from freely available datasets, and as a result, simulations are composed of a combination of routing models and meteorological datasets and do not all use the same precipitation input or hydrological set-up. In GFHM frameworks, flood magnitude estimates are often made by taking a model ‘off the shelf’, and therefore our study is aiming to understand the uncertainties that are associated in this approach. Our aim is not to provide a detailed model component intercomparison here, as this is a separate research question and is an immensely challenging task (e.g. Ceola et al 2015). Instead we intent to provide an understanding of the hydrological uncertainties of the GFHM frameworks.

Daily streamflow simulated by each model were extracted at each gauge location for a direct comparison with the observed timeseries. Due to the coarseness of the gridded hydrological models, the USGS gauge upstream drainage area was compared with the model hydrography and needed to be within ±20%. Where this condition was not met, the gauge location was removed from the analysis. This resulted in a total of 1350 gauges.

2.1.3. Global flood frequency analyses

Two regionalised flood frequency analysis (RFFA) datasets, that have been produced for global scale application, were also compared. The estimated flood return periods were extracted from the datasets at the locations of the gauging stations where daily streamflow simulations from the GHMs were available. This allows for direct comparison between the estimated flood magnitudes at these locations.

The first dataset used is from Smith et al (2015) who published the first attempt of a regionalised flood frequency analysis at the global scale. Annual maxima data from 703 selected gauging stations from the GRDC were used in their analysis. These data were first subdivided into the five main categories of the Köppen-Geiger climate classification. A hybrid-clustering approach was used to divide the gauging station locations into homogenous regions using three catchment descriptors: area, annual average rainfall, and slope. The index-flood method was applied to provide estimates of the mean annual flood (MAF) and extreme value distributions for each of the
Table 1. Description of the GHMs used in this study.

| Model version and availability | GloFAS (Harrigan et al 2020) | PCR-GLOBWB (Sutanudjaja et al 2018) | CaMa-Flood (Yamazaki et al 2011) | World Wide HYPE (Arheimer et al 2020) |
|--------------------------------|-----------------------------|-------------------------------------|-------------------------------|---------------------------------------|
|                               | GloFAS-ERA5 operational global river discharge reanalysis v2.1. [1] | PCR-GLOBWB 2 | CaMa-Flood v3.95 | World Wide HYPE v1.3 |
|                               | [https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-glofas-historical?tab=overview](https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-glofas-historical?tab=overview) | [https://geo.data.uu.nl/research-pcrglobwb/pcr-globwb_gmd_paper_sutanudjaja_et_al_2018/](https://geo.data.uu.nl/research-pcrglobwb/pcr-globwb_gmd_paper_sutanudjaja_et_al_2018/) | [http://hydro.iis.u-tokyo.ac.jp/~yamadai/cama-flood/](http://hydro.iis.u-tokyo.ac.jp/~yamadai/cama-flood/) | [https://hypeweb.smhi.se/](https://hypeweb.smhi.se/) |
| Spatial resolution            | 0.1° × 0.1°                  | Approx. 0.08° × 0.08°               | 0.1° × 0.1°                  | HRUs with average size 1000 km². Approx. 1800 000 HRUS and 131 296 catchments globally |
|                               | Approx. 11 km × 11 km at equator | Approx. 9 km × 9 km at equator     | Approx. 10 km × 10 km at equator | Applies catchment hydrological model HYPE globally (Lindström et al 2010). |
| Runoff generation             | Land surface model (LSM; HTESSEL; Balsamo et al 2009) coupled with channel routing model. | Grid-based global hydrological and water resources model (Sutanudjaja et al 2017). | LSM (HTESSEL; Balsamo et al 2009) coupled with channel routing model. | |
| Channel routing scheme        | LIsflood (Burek et al 2013) Kinematic wave | Kinematic wave | CaMa-Flood (v3.95) Local inertia equation | Flow time delay based on river length and flood wave velocity. Damping in a linear reservoir with recession coefficient for attenuation. |
| Meteorological forcing        | ERA5 reanalysis (Hersbach et al 2020) | Downscaled CRUTS 2.1 (Harris et al 2014) with ERA-Interim (Dee et al 2011) to daily values. | WATCH Forcing Data methodology applied to ERA-Interim reanalysis (Weedon et al 2014). | Hydrological Global Forcing Data (HydroGFD; Berg et al 2018). |
| Precipitation data product    | Reanalysis 0.1° × 0.1°        | Station records and reanalysis 0.5° × 0.5° | Gauge corrected reanalysis 0.5° × 0.5° | Gauge corrected reanalysis 0.5° × 0.5° |

(Continued.)
|                      | GloFAS                         | PCR-GLOBWB                     | CaMa-Flood                  | World Wide HYPE                                                                 |
|----------------------|-------------------------------|--------------------------------|-----------------------------|---------------------------------------------------------------------------------|
| **Topography**       | SRTM30 (USGS)                 | HydroSHEDS and HYDRO1K (Lehner et al 2008) | MERIT DEM (Yamazaki et al 2017) | GWD-LR (Yamazaki et al 2014) GIMP-DEM (Howat et al 2014) HYDRO1K (USGS 2020) SRTM (USGS) |
| **Evap scheme**      | Penman-Montieth               | Penman-Montieth                 | Penman-Montieth             | Jensen–Haise in temperate areas, modified Hargreaves in arid and equatorial areas, and Priestly–Taylor in polar and snow-/ice-dominated areas. |
| **Snow scheme**      | Energy balance                | Degree day                      | Energy balance              | Degree day Reservoirs and dams database incorporated from GRanD database (Lehner et al 2011). |
| **Reservoirs and Dams** | Incorporated 436 lakes and 667 reservoirs into river network (Zajac et al 2017). | Simple reservoir scheme applied to 6000 reservoirs and dams from GRanD database (Lehner et al 2011). | Not incorporated | Reservoirs and dams database from GRanD database (Lehner et al 2011). |
| **Calibration**      | LISFLOOD calibrated against daily river discharge from 1287 gauging stations worldwide (Hirpa et al 2018). | Uncalibrated                    | Uncalibrated                 | Stepwise approach for groups of parameters regulating specific processes and catchment characteristics in representative gauged catchments. |
pooled regions. From this, flood magnitudes for different return periods can be estimated.

The second dataset used is from Zhao et al (2020 in review), which aimed to make improvements to the Smith et al (2015) global RFFA. This approach involved two stages. Firstly, approximately 12 000 gauging stations from the GSIM dataset (Do et al 2018) were clustered into subareas by a K-means model. Twelve catchment descriptors covering meteorological, physiographic, hydrological, and anthropogenic factors were used. The second stage was developing a regression model in each of the subareas for the estimation of flood magnitudes using these same descriptors. It should be noted here that many of the gauges that were used in our evaluation are part of the training dataset used by Zhao et al (2020). The aim of our study is to provide the global flood hazard modelling community with some insight into how they could advance their frameworks in the future, and therefore we have included this dataset in our comparison as it is a potentially useful dataset of flood frequency estimates on the global scale.

2.2. Analysis

The annual maximum flow (AMAX) series was extracted from the daily streamflow timeseries at each of the gauge locations for observed and the four GHMs. To evaluate the sensitivity of the underlying streamflow data to extrapolation, we bootstrapped the AMAX from observed and GHMs to produce 100 resampled series for each gauge. A log Pearson type III distribution was fitted to these 100 resampled series, producing a frequency curve for all 100 samples at each site. The magnitude of the 1-in-100 year (Q100) flood was extracted from the curve to compare across the GHMs and RFFA methods as this is a common design flood magnitude used by the flood risk research community (e.g. Merz 2010). A log Pearson type III distribution was chosen as this is the standard technique for flood frequency analysis that is used by the USGS in the United States (England et al 2018), and has shown good performance in different hydrological settings (England et al 2003, Griffis and Stedinger 2007). The Q100 magnitudes were extracted from the RFFA datasets, but as we were unable to bootstrap the data used by these methods, a single estimate of Q100 is available at each gauge. To evaluate the predictive performance of each method, the per cent bias (%bias) was calculated for each estimate in relation to the observed data for all GHM bootstrap samples and at each gauge for the RFFA datasets.

Although this study focuses on the estimation of extreme flood magnitudes, the ability of the GHMs to capture the MAF was also analysed. MAF is important as GHMs use the MAF to condition the size of channels in their river networks (especially where no observations exist) (see e.g. Andreadis et al 2013, Harman et al 2008, Sampson et al 2015). Furthermore, comparisons between MAF and Q100 allow us to examine whether the results are impacted by issues with data extrapolation when fitting the log Pearson type III distribution to calculate Q100. The %bias was calculated for each gauge for the MAF of the observed and GHM timeseries. The %bias scores for MAF and Q100 where sorted from largest underprediction (i.e. most negative) to largest overprediction (i.e. largest positive) and each given a rank score. A Spearman’s rank correlation coefficient was then calculated for each of the GHMs to analyse the correlation between the biases at the gauges.

In order to identify the controls on the spatial variability of the model biases in Q100 estimates, multiple climatic attributes and basin characteristics were compared with the %bias scores. The GAGES-II (Falcone 2011) dataset provides geospatial data and classification for 9322 stream gauges maintained by the USGS, and was used here to define different descriptor variables for the gauges in this study. This dataset was also used to group the gauges based on their climatic and topographic settings. These regions were based on the hydrologic landscape regions (HLRs) concept developed by Winter (2001) and aimed to organise basins in terms of their hydrologic similarity. This was then implemented in the US by Wolock et al (2004) who used statistical and GIS analyses applied to land-surface form, geologic texture, and climate variables that describe the physical and climatic setting of 43 931 catchments in the US. These variables are chosen in order to conceptualise the actual hydrologic system in a uniform way. The dominant HLR for basins is given in the GAGES-II dataset. These were extracted for the gauges used in this study and were modified to provide nine regions of climate and topography classes. Climate classes are humid, subhumid, arid and semi-arid. Topography classes are plains, plateaus and mountains. A map of the region assigned to each of the 1350 gauges in this study is provided in the supplementary information (figure S1 (available online at stacks.iop.org/ERL/16/064013/mmedia)).

3. Results

3.1. How well do RFFA and GHMs reproduce observed extreme flood magnitudes?

To compare how well GHMs and RFFA methods reproduce extreme flood magnitudes the %bias was calculated at each gauge between the Q100 estimate calculated from each model and the observed data. These %bias scores were then grouped into categories of within error, over- and under-predictions. The gauges with %bias scores considered to be ‘within error’ here are those that have a value of between −25% and 25%. This threshold is chosen to represent and acknowledge the often large uncertainties of extreme discharge measurements (e.g. McMillan et al 2012, Coxon et al 2015). Figure 1 shows the spatial distribution of bias groups for each GHM and
Figure 1. The AMAX series extracted from the observed records and the model simulations were bootstrapped to produce 100 samples. The per cent bias (bias) was calculated between the bootstrap samples of Q100 estimates for the observed and GHMs. The scores were grouped into categories of ‘within error’ (−25% to 25%), under- and over-predicting, and these groups were used to create the legend for the mapped gauges. Point size represents the per cent of the bootstrap samples that are classified with the same bias group. Note: the Q100 estimates from the RFFA datasets could not be included in the bootstrap sample analysis, and therefore the deterministic results are presented. Bar charts in bottom left hand corner of each map represent the number of gauges for each dataset in each of the per cent bias categories.

For gauges with predictive performance in the ‘within error’ category, there are 17% for GloFAS, 23% for PCR-GLOBWB, 26% for CaMa-Flood, 26% for HYPE, 11% for the Fathom RFFA, and 40% for Zhao RFFA. Figure 1 also shows that there is a large difference between the Q100 estimates calculated using GHMs and RFFA methods and there are generally different spatial patterns of Q100 predictive performance.

The GHMs tend to underpredict Q100 in the east and overpredict in the west and central areas of the US. There are 223 gauges where all of the GHM bootstrap samples underpredict Q100, and 53% of these are in the East and Great Lakes regions. There are 72 gauges where all GHM bootstrap samples overpredict, and 70% are in the central and western areas of the US. GloFAS tends to underpredict Q100 consistently with 68% of bootstrap samples having a bias of less than −25%, and 77% of all underpredictions less than −50%. HYPE also tends to underpredict Q100, with 52% of bootstrap samples in the underpredicting bias groups, and 58% of these gauges have a bias value less than −50%. PCR-GLOBWB and CaMa-Flood tend to overpredict Q100 in the west and central areas of the US with 60% and 65% of all overpredictions in these regions respectively. CaMa-Flood tends to perform well in the east coast and Appalachians region, with 64% of gauges in the ‘within error’ category. The number of GHM bootstrap samples that were classified with the same bias group at each of the gauges was also assessed. The mode of the bias groups was calculated, and then the per cent of the bootstrap samples where the bias group agreed with this mode was analysed. The per cent of bootstrap samples where more than 50% were classified in the same bias group for each gauge is 83% for GloFAS, 83% for PCR-GLOBWB, 86% for CaMa-Flood, and 86% for HYPE.

For the RFFA methods, there are large differences in the spatial patterns of bias in estimates of Q100. The Fathom RFFA consistently overpredicts Q100 magnitudes, with 83% of all the gauges overestimating, and 73% of these gauges overpredicting Q100 by more than 100%. The Zhao RFFA performs well in most of the Eastern US, but tends to overpredict in the central regions, with 56% of the gauges in this area overpredicting, and 52% of those gauges overestimating Q100 by more than 100%. Although the spatial patterns and the distribution of bias groups (represented by the bar graphs in figure 1) look different for the RFFA methods, there are some similarities. 235 gauges experience an agreement in bias category, and of these 47% are in the Western US. This is seen in figure 1 along the West coast and the gauges in Arizona.

Although the key focus of this research is to understand the uncertainties in estimates of extreme flood magnitudes, there are often issues with the
extrapolation of observed and modelled data to extreme events. In figure 2, the %bias scores for deterministic MAF and Q100 for each GHM were ranked from lowest to highest (i.e. the largest under-prediction to the largest over-prediction) and compared. A Spearman’s rank correlation coefficient for each model was also calculated to quantify the correlation of the %bias ranks, which were 0.79, 0.79, 0.82 and 0.78 for GloFAS, PCR-GLOBWB, CaMa-Flood and HYPE, respectively, indicating a strong correlation between MAF and Q100 biases. Spearman’s rank correlation was also calculated for the ranked bootstrap samples for all GHMs, and coefficients were 0.81 for GloFAS, 0.79 for PCR-GLOBWB, 0.82 for CaMa-Flood, and 0.80 for HYPE. Figure 2 also shows that there is a relationship between GHM bias in Q100 predictions and basin mean elevation, with GloFAS, PCR-GLOBWB and CaMa-Flood overestimating in basins with highest elevation, and HYPE underestimating in these basins.

3.2. What controls the spatial variability in prediction performance?

To identify general controls on the spatial variability in prediction performance of the different flood estimation methods, the grouped Q100%bias scores were plotted against different climatic attributes and catchment characteristics. A wide range of different attributes were analysed (not shown, but can be found in supplement figure S1). We found basin area and topography to show the strongest relationship with %bias scores.

Figure 3(A) shows that for all GHMs and RFFA methods Q100 tends to be underpredicted the most in the smallest basins. For the largest basins, although the median basin size is small, there are large overestimations of Q100. The largest underpredictions are found in basins with area <10 000 km². The mean basin area of the gauges within the largest under-prediction group (i.e. %bias < −50%) is 2776 km² for GloFAS, 2719 km² for PCR-GLOBWB, 2698 km² for CaMa-Flood, 2732 km² for HYPE, 3468 km² for the Fathom RFFA, and 2256 km² for the Zhao RFFA. The GHMs tend to overpredict in the largest basins. For all the basins with area greater than 30 000 km², 31% for GloFAS, 78% for PCR-GLOBWB, 52% for CaMa-Flood, and 29% for HYPE, are overestimating Q100 by more than 100%.

Mean elevation, standard deviation of elevation and basin mean per cent slope were used as indicators of topographic influence on %bias. Figures 3(B) and (C) shows that topography has a different influence when the GHMs are compared with the RFFA data-sets. There is a consistent underprediction in basins with low mean and standard deviation of elevation and an over-prediction in basins with high mean and standard deviation of elevation by the GHMs, except for HYPE. This indicates that the models perform poorly in low-lying land with low terrain complexity, such as deltas and floodplains. This is likely due to the simplified routing schemes that are used by the models which limit the accuracy of river discharge especially in flat areas in high-flow periods (Zhao et al 2017). The models also perform poorly in high
altitude regions with complex terrain, such as mountainous regions. This may be caused by the significant uncertainty found over complex terrain caused by inaccuracy of the meteorological forcing datasets (Spers Weiland et al 2015). The mean elevation of the basins in the overpredicting %bias groups are
994 m for GloFAS, 951 m for PCR-GLOBWB, 1559 m for CaMa-Flood, and 409 m for HYPE. For all the basins that have an elevation of more than 1000 m (359 gauges from the whole sample) the percentage of these where each model overpredicts Q100 is 30% for GloFAS, 53% for PCR-GLOBWB, 70% for CaMa-Flood, and 21% for HYPE. The GHMs also tend to underpredict Q100 the most in the basins with the lowest elevation, again except for HYPE, with the mean elevation of the basins in the %bias < −50% group 367 m for GloFAS, 458 m for PCR-GLOBWB, 342 m for CaMa-Flood, and 596 m for HYPE.

The RFFA methods show a different relationship between Q100%bias and topography. Both the Fathom RFFA and Zhao RFFA tend to underpredict the most in basins with higher mean elevations, with mean elevation of these basins 1244 m for Fathom RFFA and 984 m for Zhao RFFA. However, there is a contrasting relationship with overpredicting Q100, with the average elevation of the gauges overestimating by more than 100% is 604 m for Fathom RFFA and 938 m for Zhao RFFA.

### 3.3. Do ensemble predictions perform better than any individual prediction?

Figure 4 shows that there is not a single GHM which performs best in each of the regions. This is an expected result, as each of the models has different meteorological forcing datasets and process representations. However, it does show that some of the GHMs perform better in some settings in comparison to others. For example, GloFAS consistently underpredicts in all regions, except for the semi-arid mountains. Figure 4 also highlights that Zhao RFFA is skilful in all regions, except for arid plains. However, as mentioned previously, many of the gauges that are used in this study to compare the flood magnitude estimates against are used in the training dataset used in Zhao et al. (2020) to develop the sub-region regression models. This gives this dataset an unfair advantage when assessing its predictive capabilities, as it has been calibrated against the GSIM discharge dataset, which incorporates many of the USGS gauges that are used for performance evaluation here.

Figure 5 shows the per cent of GHM bootstrap samples at the gauges in each of the regions where the GHM is the best performing model. This analysis shows that HYPE and CaMa-Flood are the best performing models in the majority of regions, where they account for the best model of approximately 60% of gauges in all regions, except for arid plains, semi-arid plateaus and semi-arid mountains. However, when looking at the percentage of gauges where each of the GHMs is the best performing in each of the regions, there is an almost equal number in most regions. For example, in the subhumid plains, the percentage of gauges in the region where each GHM is the best performing is 13% for GloFAS, 28% for PCR-GLOBWB, 24% for CaMa-Flood and 35% for HYPE. Although Figure 4 shows GloFAS as being the least skilful at estimating Q100, in the arid plains, semi-arid plateaus and humid mountains it performs best at 32%, 20% and 32% of gauges in each region, respectively.

### 4. Discussion and conclusion

#### 4.1. Spatial variability in predictive performance and contributing factors

Our results have shown that there is a clear variability in the predictive performance of extreme discharge magnitudes from the GHMs and RFFA datasets used in this study over the conterminous US. GHMs tend to overpredict Q100 magnitudes in the western and central regions of the US. This is a common finding in hydrological modelling studies due to the overestimation of discharge in basins with large (semi-) arid regions (Müller Schmied et al. 2014, Bierkens 2015, Beck et al. 2017b, Mathews et al. 2020), as found in the Central Great plains and Southwestern US. This problem is likely caused by a combination of uncertainty in precipitation data inputs and poor representation of complex rainfall-runoff processes in dry climates (Yatheendradas et al. 2008, Zhang et al. 2009).
In a GHM evaluation and comparison study, Mathews et al. (2020) found that the forcing data uncertainty was often augmented by the model structure in arid regions, as small variations in precipitation can drive much higher variations in simulated variables in dry systems. The RFMA methods also have similar poor predictive performance in the southwest (seen in figure 1). Smith et al. (2015) attribute this to the variability of rainfall and the temporal variability of runoff responses tend to be higher in drier catchments (Di Baldassarre et al. 2006, Merz and Bäuschl 2009b, Padi et al. 2011), and this is not reflected in average annual precipitation, which is used as a proxy for extreme rainfall in RFMA methods. The GHMs, except for HYPE, show a pattern of overpredicting Q100 in high elevation basins (figures 2 and 3). This is likely due to the error introduced into estimates in global precipitation datasets in mountainous regions where precipitation variability is very high and gauge density (for ground truthing) is very low (Gudmundsson et al. 2012, Sperna Weiland et al. 2015), and where possible high resolution forcing data should be used in regions with complex topography (Paiva et al. 2012, Towner et al. 2019). For HYPE, Arheimer et al. (2020) found that the precipitation datasets have more robust estimates at low altitudes with smooth orography, and where the seasonality is more regular and easier to capture. Also, high altitude basins are likely affected by snow dynamics and therefore there may be an issue with the representation of snow in the hydrological model structures (Gudmundsson et al. 2012). CaMa-Flood mostly overpredicts in the high elevation basins and this might be because of the snow scheme that is used in HITESSEL which is the LSM that is used to produce runoff. Previous studies (e.g. Zhang et al. 2016, Beck et al. 2017b) have found that LSMs on average perform better in rainfall-dominated catchments, and GHMs tend to achieve better results in snow-dominated catchments due to the use of complex energy balance equations used in LSMs introducing additional uncertainties.

We have attempted to address the uncertainty of extrapolation of short flow time series to extreme events by bootstrapping the AMAX series of the observed records and model simulations. This is done to allow us to assess the sensitivity of the Q100 estimates to the underlying flow data. We compared the bootstrap sample results with the deterministic results (i.e. the Q100 calculated from the observed and modelled flow time series), and the overall spatial patterns were found to be the same. These deterministic plots can be found in the supplementary information (figure S3). This leads us to believe that the results that we have presented here are robust.

4.2. Global flood hazard predictions should explore the use of a weighted ensemble of extreme flood magnitude estimates

Our analysis in this study has shown that there is no single method or model that performs best everywhere, and therefore the standard practice of the global flood hazard modelling community of using a single estimate of input flood magnitude is not representative of the hydrological uncertainties. Figure 4 and figure 5 show that CaMa-Flood and HYPE are the best performing models in the majority of regions, however, the other models perform well in other settings. For example, GloFAS provides the best Q100 predictions in the semi-arid mountains. These
results suggest that a combination of models in a weighted ensemble should be explored for future iterations of GFHM development in order to represent the uncertainty of model selection for the generation of extreme flow magnitudes. Multi-model ensembles are a way of combining the outputs from multiple models and typically improve the predictive capabilities, and are a widely used approach in atmospheric, climate and hydrological sciences (Wandishin et al 2001, Tebaldi and Knutti 2007, Breuer et al 2009, Viney et al 2009).

In the context of global flood hazard modelling, ideally an ensemble of flood flow magnitudes would be used for the conditioning of river channel sizes and for the simulation of inundation extent and flood hazard. Multi-model ensembles are often formed of the mean of the ensemble members, but this is often not the most meaningful approach (Winter and Nychka 2010). By using a weighted ensemble, this gives the modeller the opportunity to give ensemble members with better skill higher weights in regions where they are the ‘best’, and give a weight of zero to ensemble members who have no skill in a region. This approach is technically very feasible with the data we currently have. This would allow for different process representations and an array of different meteorological forcing datasets—which together may compensate for each other’s biases—and lead to a more robust input for GFHMs to produce flood maps that represent input flood magnitude uncertainty. Open questions remain regarding how to strategically select ensemble members that maximize ensemble performance (Winter and Nychka 2010).

4.3. Limitations and emerging research questions

All of the GHMs use different forcing datasets and have different process representations in their model structures. This makes it difficult to attribute the differences that we have presented in our results to any specific processes or datasets used in the models (as shown previously, e.g. Gudmundsson et al 2012). In order to correct this source of difference, we would need to force each of the GHM simulations with the same forcing data and homogenise the workflows used to set-up the different models—to avoid differences due to parameter estimation or model initialisation strategies (Ceola et al 2015), which is out of the scope of this paper and an extremely difficult task to tackle and likely the project of an international consortium. The aim here was to provide general insight into the contributing factors to the differences between the Q100 estimates. However, it is widely recognised that an end-to-end uncertainty propagation study would help advance our understanding of the errors and uncertainties in global scale hydrological studies (Bernhofen et al 2018, Mathews et al 2020). Another drawback is that we cannot quantify the influence on flood hazard assessment and estimates that these different estimates of Q100 have within a GFHM framework, but this does raise an interesting research question for future work. Recent work by Aerts et al (2020) compared estimates of flood hazard and impacts on the population in China, finding variations exist up to a factor of four between the flood hazard map outputs of GFHMs in terms of inundated area and exposed GDP. However, they only used the deterministic final outputs of these models. An end-to-end flood risk assessment using a weighted ensemble of flood magnitude estimates, as produced in this study, in a GFHM would be beneficial to the flood risk research community as it may identify regions where hydrological uncertainties have large impacts on estimated inundation extents.

Data availability statement

USGS river gauge data are available from https://waterdata.usgs.gov/nwis/. The CaMa-Flood model output used here can be accessed via http://hydro.iis.u-tokyo.ac.jp/~yamadai/cama-flood/under ‘Additional Data (external)’ (file: E2O_WRR2 ECMWF HITESSEL). PCR-GLOBWB output files from the model version used in this work are available on https://geo.data.uu.nl/research-pcrglobwb-kinematic-wave/global_daily_discharge_1958-2015/ (access requested from Edwin Sutanudjaja). GloFAS model output can be accessed from the Copernicus Climate Change Service Climate Data Store: https://cds.climate.copernicus.eu/cdsapp#!/dataset/ecmcs-glofas-historical?tab=overview. Time series and maps from the WorldWide HYPE model for single catchments are available for free at https://hypeweb.smhi.se/explore-water/ and documentation and open-source code of the HYPE model are available at https://hypeweb.smhi.se/model-water/. Data was extracted and delivered by SMHI. The parameters for the Fathom RFFA can be found in Smith et al (2015) and the flood magnitudes estimates can be generated from this. The flood frequency estimates used here were provided to us by Andrew Smith. The Zhao et al (2020) dataset is available on request from the lead author at gang.zhao@bristol.ac.uk.

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

Acknowledgments

This work is funded as part of the Water Informatics Science and Engineering Centre for Doctoral Training (WISE CDT) under a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant EP/L016214/1. Partial support to TW was provided by a Royal Society Wolfson Research Merit Award [WM170042] and by the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research. We would like to thank those at SMHI that extracted the
flow timeseries from WorldWide HYPE for all of the US basins and shared this data with us. We would also like to thank Dai Yamazaki for providing us with the flow timeseries for the US from CaMa-Flood, and Edwin Sutanudjaja for providing us with the model outputs from PCR-GLOBWB, Andrew Smith for providing us with the Fathom RFFA data at the gauges used in this work, and Gang Zhao for providing us with his RFFA data in the US. We thank the two anonymous reviewers for their constructive criticism of the manuscript.

**ORCID iD**

Laura Devitt [https://orcid.org/0000-0001-8800-429X](https://orcid.org/0000-0001-8800-429X)

**References**

Aerts J P M, Uhlemann-Elmer S, Fialander D and Ward P J 2020 Comparison of estimates of global flood models for flood hazard and exposed gross domestic product: a China case study Nat. Hazards Earth Syst. Sci. 20 3245–60

Andreas K M, Schumann G J P and Paveisky T 2013 A simple global river bankfull width and depth database Water Resour. Res. 49 7164–8

Archfield S A et al 2015 Accelerating advances in continental domain hydrological modelling Water Resour. Res. 51 10978–91

Arheimer B, Pimentel R, Isberg K, Crochemore L, Andersson J C M, Hasan A and Pineda L 2020 Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation Hydrol. Earth Syst. Sci. 24 535–59

Balsamo G, Beljaars A, Scipal K, Viterbo P, van der Hurk B, Hirschi M and Betts A K 2009 A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the integrated forecast system J. Hydrometeorol. 10 623–43

Beck H E, van Dijk A J M, de Roo A, Dutra E, Fink G, Orth R and Schellekens J 2017b Global evaluation of runoff from 10 state-of-the-art hydrological models Hydrology and Earth System Sciences 21 2881–903

Beck H E, van Dijk A J M, de Roo A, Miralles D G, McVicar T R, Schellekens J and Bruinjze L A 2016 Global-scale regionalization of hydrologic model parameters Water Resour. Res. 52 3599–622

Beck H E, Vergopolan N, Pan M, Levizzani V, van Dijk A J M, Weedon G P, Brocca L, Pappenberger F, Huffman G J and Wood E F 2017a Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modelling Hydrol. Earth Syst. Sci. 21 6201–17

Berg P D, Donnelly C and Gustafsson D 2018 Near-real-time adjusted reanalysis forcing data for hydrology Hydrol. Earth Syst. Sci. 22 989–1000

Bernhoven 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

Beven K J and Cloke H L 2012 Comment on ‘Hyperresolution global land surface modelling: meeting a grand challenge for monitoring Earth’s terrestrial water’ by Eric F. Wood et al Water Resour. Res. 48 W191801

Bierkens M F P 2015 Global hydrology 2015: state, trends, and directions Water Resour. Res. 51 4923–47

Blöschl G et al (eds) 2013 Prediction of floods in ungauged basins Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales (Cambridge: Cambridge University Press) 189–225

Breuer L et al 2009 Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM): I: model intercomparison with current land use Adv. Water Resour. 32 129–46

Burek P, van der Knijff J M and de Roo A P J D 2013 LISFLOOD—distributed water balance and flood simulation model—revised user manual, Publications Office of the European Union (https://doi.org/10.2788/24719)

Cecilia S et al 2015 Virtual laboratories: new opportunities for collaborative water science Hydrol. Earth Syst. Sci. 19 2101–17

Chen S et al 2014 Evaluation of high-resolution precipitation estimates from satellites during July 2012 Beijing flood event using dense rain gauge observations PLoS One 9 e89681

Coxon G, Freer J, Westerberg J K, Wagener T, Woods R and Smith P J 2015 A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations Water Resour. Res. 51 3531–46

CRED and UNISDR 2015 The Human Cost of Weather Related Disasters 1995–2015, United Nations Office 497 for Disaster Risk Reduction (UNISDR) and Centre for Research on the Epidemiology of Disasters (CRED)

de Cicco L A, Lorenz D, Hirsch R M and Watkins W 2018 dataRetrieval: R packages for discovering and retrieving water data available from U.S. federal hydrologic web services (https://doi.org/10.5066/P9XL5GE) (https://code.usgs.gov/water/dataRetrieval)

Dee D P et al 2011 The ERA-Interim reanalysis: configuration and performance of the data assimilation system Q.J.R. Meteorol. Soc. 137 533–97

Di Baldassarre G, Castellarin A and Brath A 2006 Relationships between statistics of rainfall extremes and mean annual precipitation: an application for design-storm estimation in northern central Italy Hydrol. Earth Syst. Sci. 10 589–610

Di Baldassarre G and Montanari A 2009 Uncertainty in river discharge observations: a quantitative analysis Hydrol. Earth Syst. Sci. 13 913–21

Do H X, Gudmundsson L, Leonard M and Westra S 2018 The global streamflow indices and metadata archive—part I: station catalog and catchment boundary Earth Syst. Data 10 765–85

Döll P, Kaspar F and Lehrner B 2003 A global hydrological model for deriving water availability indicators: model tuning and validation J. Hydrol. 270 105–34

Dottori F, Salamono 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

Engeland K, Xu C- Y and Gottschalk I 2005 Assessing uncertainties in a conceptual water balance model using Bayesian methodology Hydroli. Sci. J. 50 45–60

Engeland J F Jr., Cohn T A, Faber B A, Steding J R, Thomas W O Jr, Veilleux A G, Kiang J E and Mason R R Jr 2018 Guidelines for determining flood flow frequency—bulletin 17 C (ver. 1.1, May 2019): U.S. Geological Survey Techniques and Methods, book 4, chap. B5 148

Engeland J F, Salas J D and Jarrett R D 2003 Comparisons of two moment-based estimators that utilize historical and paleoflood data for the log Pearson type III distribution Water Resour. Res. 39 1243

Falcone J A 2011 GAGES-II: geospatial attributes of gages for evaluating streamflow, digital spatial data set 2011, (available at: http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sep2011.xml) (Accessed 15 May 2020)

Griffis T W and Stedinger J R 2007 Log-Pearson type 3 distribution and its application in flood frequency analysis: I: distribution characteristics J. Hydrol. Eng. 12 482–91

Gudmundsson L, Wagener T, Tallaksen L M and Engeland K 2012 Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe Water Resour. Res. 48 W11504
balance and flood simulation *International Journal of Geographical Information Science* 24 189–212

Viney N R et al 2009 Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: ensemble combinations and predictions *Adv. Water Resour.* 32 147–58

Wagener T and Wheater H S 2006 Parameter estimation and regionalisation for continuous rainfall-runoff models including uncertainty *J. Hydrol.* 320 132–54

Wandishin M S, Mullen S L, Stensrud D J and Brooks H E 2001 Evaluation of a short-range multimodel ensemble system *Mon. Weather Rev.* 129 729–47

Ward P J et al 2015 Usefulness and limitations of global flood risk models *Nat. Clim. Change* 5 712–5

Weedon G P, Balsamo G, Bellouin N, Gornes 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

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

Winter C L and Nychka D 2010 Forecasting skill of model averages *Stoch Environ Res Risk Assess* 24 633–38

Wolock D M, Winter T C and McMahon G 2004 Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analyses *Environ. Manage.* 34 S71–88

Yamazaki D, Ikeshima D, Tawatari R, Yamaguchi T, O’Loughlin F, Neal J C, Sampson C C, Kanae S and Bates P D 2017 A high-accuracy map of global terrain elevations *Geophys. Res. Lett.* 44 5844–53

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 W04501

Yamazaki D, O’Loughlin F, Trigg M A, Miller Z F, Pavelsky T M and Bates P D 2014 Development of the global width database for large rivers *Water Resour. Res.* 50 3467–80

Yatheelendradas S, Wagener T, Gupta H, Unkrich C, Goodrich D, Schaffner M and Stewart A 2008 Understanding uncertainty in distributed flash flood forecasting for semi-arid regions *Water Resour. Res.* 44 W05519

Zajac Z., Revilla-Romero B, Salamon P, Burek P, Hirpa F A and Beck H 2017 The impact of lake and reservoir parameterisation on global streamflow simulation *J. Hydrol.* 548 522–68

Zhang Y, Zheng H, Chiew F H S, Pena-Arancibia J and Zhou X 2016 Evaluating regional and global hydrological models against streamflow and evapotranspiration measurements *J. Hydrometeorol.* 17 995–1010

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

Zhao G, Bates P, Neal J and Pung B 2020 in review Design flood estimation for global river networks based on machine learning models *Hydrol. Earth Syst. Sci.* (https://doi.org/10.5194/hess-2020-594)