Going Beyond the Ensemble Mean: Assessment of Future Floods From Global MultiModels

Giuntoli, Ignazio; Prosdocimi, Ilaria; Hannah, David M.

DOI:
10.1029/2020WR027897

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Giuntoli, I, Prosdocimi, I & Hannah, DM 2021, 'Going Beyond the Ensemble Mean: Assessment of Future Floods From Global MultiModels', *Water Resources Research*, vol. 57, no. 3, e2020WR027897. https://doi.org/10.1029/2020WR027897

AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR027897

Going Beyond the Ensemble Mean: Assessment of Future Floods From Global Multi-Models

Ignazio Giuntoli[1,2], Ilaria Prosdocimi[3], and David M. Hannah[1]

[1]School of Geography, Earth and Environment Sciences, University of Birmingham, Birmingham, UK, [2]Now at Istituto di Scienze dell'Atmosfera e del Clima, CNR-ISAC, Bologna, Italy, [3]Dipartimento di Scienze Ambientali, Informatica e Statistica, Ca' Foscari University of Venice, Venezia Mestre, Italy

Key Points:

- A Bayesian hierarchical model is developed to assess the changes in future flood magnitude and quantify uncertainty in a single step
- Future flood magnitude at selected sites over the eastern United States decreases in the south of the domain with varying uncertainty
- A constrained ensemble based on how well model runs replicate timing of observed peak flows yields similar results to the full ensemble

Supporting Information:

- Supporting Information S1

Correspondence to:
I. Giuntoli,
i.giuntoli@bham.ac.uk

Citation:
Giuntoli, I., Prosdocimi, I., & Hannah, D. M. (2021). Going beyond the ensemble mean: Assessment of future floods from global multi-models. *Water Resources Research*, *57*, e2020WR027897. https://doi.org/10.1029/2020WR027897

Received 7 MAY 2020
Accepted 31 DEC 2020

Abstract Future changes in the occurrence of flood events can be estimated using multi-model ensembles to inform adaption and mitigation strategies. In the near future, these estimates could be used to guide the updating of exceedance probabilities for flood control design and water resources management. However, the estimate of return levels from ensemble experiments represents a challenge: model runs are affected by biases and uncertainties and by inconsistencies in simulated peak flows when compared with observed data. Moreover, extreme value distributions are generally fit to ensemble members individually and then averaged to obtain the ensemble fit with loss of information. To overcome these limitations, we propose a Bayesian hierarchical model for assessing changes in future peak flows, and the uncertainty coming from global climate, global impact models and their interaction. The model we propose allows use of all members of the ensemble at once for estimating changes in the parameters of an extreme value distribution from historical to future peak flows. The approach is applied to a set of grid-cells in the eastern United States to the full and to a constrained version of the ensemble. We find that, while the dominant source of uncertainty in the changes varies across the domain, there is a consensus on a decrease in flood magnitudes toward the south. We conclude that projecting future flood magnitude under climate change remains elusive due to large uncertainty mostly coming from global models and from the intrinsic uncertain nature of extreme values.

1. Introduction

A warming climate is expected to intensify the global water cycle with changes in the occurrence and severity of extreme events like intense precipitations and floodings (Abbott et al., 2019; Lavell et al., 2012). In turn, the main components of flood risk (Crichton, 1999) are expected to increase: flood hazard (as a result of increased energy in the system and of an intensified water cycle), flood exposure of people and assets (owing to global population growth and cities becoming more urbanized) and flood vulnerability (especially in overpopulated regions with low preparedness and poor infrastructure; Oppenheimer et al., 2014). In this context, assessing changes in future floods is crucial to inform adaptation and mitigation strategies aimed at protecting human life, vulnerable ecosystems, human wellbeing, agricultural land, homes and other socio-economic assets.

Projected increases in temperature and heavy precipitation imply regional-scale changes in flood frequency and intensity (Seneviratne et al., 2012). The projected impacts of floods depend on the change in climatic characteristics and on the change in the magnitude and seasonal distribution of precipitation, temperature, and evaporation (Cisneros et al., 2014). Changes in land-use, water management and abstraction resulting from human activities are also factors that influence the terrestrial phase of the water cycle and, in turn, flood characteristics (Prosdocimi et al., 2015). Two practical examples are the likely increase in pluvial flooding, as a result of more frequent intense precipitation events under climate change (Pendergrass, 2018), and the reduction and shift in time of the annual spring flood in snow dominated catchments, as a result of reduced snow pack (Musselman et al., 2018).

Model-based climate change projections for different greenhouse gas emission scenarios are a valuable source of information about future extreme events (Goodess, 2012). Attempts to anticipate changes in future flood risk have come forth in recent years both at the catchment scale by statistically post-processing (e.g., downscaling) climate variables like rainfall and simulating runoff using a hydrological model

(Bosshard et al., 2013; Camici et al., 2014) and at continental to global scales, employing global model ensembles chains, usually using bias-corrected global climate model (GCM) runs feeding global impact models (GIMs) that simulate runoff at the land surface (e.g., Alfieri et al., 2015; Dankers et al., 2014; Hirabayashi et al., 2013; see François et al., 2019 for details on the two approaches). Regardless of scale, a consensus has grown in the hydrological community on the need to make the simulation of hydrologic processes less uncertain and consequently more useful for informing and guiding decisions (Clark et al., 2015; Merz et al., 2014). Concerning the focus of this study—global models—as the climate system is inherently chaotic, even using perfect models tuned with perfect observations we would still be dealing with uncertainty from natural variability (Deser et al., 2012). On top of natural variability, errors in model structure and parameterization undermine the estimate of future extreme events, notwithstanding the uncertainty coming from emission scenarios (Hawkins & Sutton, 2009; Lehner et al., 2020), although Giuntoli et al. (2018) report that this source accounts for very little uncertainty in runoff projections compared to that of GCMs and GIMs. The aim of improving the simulation of climate and land-surface systems through the increase of spatial and temporal resolution and the inclusion of physical processes that were until recently overlooked comes at a cost of increased complexity, likely to yield a wider spread of plausible outcomes, thus increased uncertainty. In this context, extremes should raise even more concern because of the catastrophic consequences of their occurrence and the difficulty in sampling and characterizing them even when using observed data. For flood hazard planning extreme value theory is generally employed (Goodess, 2012; Katz et al., 2013) to derive estimates of design events—that is, the flow magnitude that is expected to be exceeded on average with a certain fixed probability in any given year (under the assumption of independence between flows recorded in different years).

At the global scale, changes in mean flows from global models indicate an increase at high latitudes and in the wet tropics, and a decrease in most dry tropical regions, although some regions have high uncertainty in the magnitude and direction of change (e.g., Hagemann et al., 2013; Schewe et al., 2014). Conversely, changes in flood magnitude are less consistent, with contrasting results among studies depending on the region and the ensemble setup (Dankers et al., 2014; Giuntoli et al., 2015a; Hirabayashi et al., 2013). The lack of consistency in these changes is emphasized by Cisneros et al. (2014) reporting that studies of flood projections under different emission scenarios are still few, and highly uncertain, given the complexity of the mechanisms driving floods at the regional scale. In fact, studies using runoff projections have started trying, in addition to assessing future floods characteristics, to untangle the uncertainty originating from the different components of the modeling chain (e.g., Giuntoli et al., 2015a; Koirala et al., 2014).

The present work builds on Giuntoli et al., (2015a), who demonstrated the important role of GIMs in driving uncertainty in changes of future high flows globally (sometimes outweighing that of GCMs) and on Giuntoli et al. (2018), who highlighted the small role of scenario uncertainty compared to that of global models along with how the choice of GIMs affects overall uncertainty in peak flows projections. We combine findings from these works to go one step further overcoming the use of the ensemble mean (associated to e.g., the signal-to-noise to appraise model agreement) to characterize the signal of change of future floods and to quantify the uncertainty of the signal coming from GIMs and GCMs, provided that the RCP (Representative Concentration Pathways) contribution is negligible compared to the first two sources.

In light of these research gaps, the overarching aim of this study is to apply a novel Bayesian model to the eastern USA to estimate space-time changes in future flood magnitude from multi-model ensembles and so improve the overall signal/pattern of change and identify sources of uncertainty in projections. In particular, we:

1. Propose a statistical method for estimating changes in future flood magnitude that minimizes loss of information and allows for an interpretable partition of the sources of variability (uncertainty).
2. Test the method over the eastern USA on a full multi-model ensemble identifying spatial patterns of flood magnitude changes and uncertainty.
3. Compare simulated flood peaks to observed data for selecting more credible model runs for testing the method on a constrained ensemble and compare results.

For the first step, we propose an improved way to assess changes in flood magnitude using multi-model ensembles that goes beyond expressing changes through the ensemble mean (or median), which cancels

out information on model consensus (or lack thereof) and reduces the signal across multiple members to a single value. In fact, taking the mean of the ensemble, which is an approach commonly used to summarize the overwhelming amount of information from climate projections, serves only to conceal the uncertainty and negatively impact characterization of extremes, rather than actively incorporate that uncertainty into design (François et al., 2019). To this end, using a Bayesian hierarchical model, we consider all members at once within the same statistical model that provides not only the signal of the direction of change, but the entire distribution of the overall change, and therefore a comprehensive description of the uncertainty in the model outputs.

For the second step, using the ISIMIP multi-model ensemble—already employed in future high flows studies (Dankers et al., 2014; Dottori et al., 2018; Giuntoli et al., 2015a)—we focus on the eastern half of the United States where observed data (relatively free from anthropogenic disturbance) are available in catchments large enough to be compared to corresponding model grid-cells. On selected grid-cells over the domain of study, described in Section 2, we carry out an analysis of the annual maximum flow (extracted from daily data) using a Bayesian hierarchical model estimating changes in the future (2065–2099) flood peaks compared to the recent historical period (1971–2005) using the Gumbel distribution and expressing the uncertainty coming from the choice of GCMs or GIMs as the variation of the statistical model's random effects. It should be noted that the terminology "GIMs" used herein could also be referred to as "GHMs" that is, global hydrological models.

Lastly, for the third step, in addition to assessing changes in flood magnitude on all available runs of a multi-model ensemble experiment, we exploit model biases in present-day runoff peaks (against observed data) to constrain projected changes in flood design events (as in e.g., Yang et al., 2017). There is indeed a growing interest in the scientific community dealing with climate impact studies on the opportunity of going beyond the "one-model one-vote approach" (or "model democracy"; Knutti, 2010) and favoring model runs with a better historical performance in reproducing observations with the aim to reduce uncertainty (Padrón et al., 2019). The overall effort of model selection is to extract efficiently the information relevant to a given projection or impact question, beyond the naïve use of multi-model ensembles (e.g., CMIP5) in their entirety (Abramowitz et al., 2019). This approach is in line with the fact that, owing to different model performances against observations and the lack of independence among models, there is evidence now that giving equal weight to each available model projection is suboptimal (Eyring et al., 2019). Indeed, modeled data can show large discrepancies from observed data, especially in the tails of the distribution (Do et al., 2020). Thus, we apply this framework to the entire ensemble (oE) and to a constrained version (cE) in order to understand whether constraining model runs with observations can be considered beneficial to future peak flow changes analyses.

We present the data in Section 2 with an appraisal of how peak flow modeled data compares to observed data. In Section 3, we describe the statistical framework for estimating future changes in flood magnitude and then how the ensemble is constrained. Results are presented in Section 4 before discussing them in the final Section 5.

2. Data

Annual maximum flows (henceforth referred to as AMax) were extracted from 18 grid-cells daily runoff (simulated) and corresponding gauges' daily streamflow (observed) located in the eastern half of the United States (Figure 1).

Observed data were selected to match the size of model data grid-cells (0.5° × 0.5°, i.e., ∼50 km × 50 km at the equator), so those with catchment areas in the range of 2,000 – 2,500 (2,500 – 3,000) km^2 north (south) of 36°N latitude and with daily discharge data covering the models' control period (1971–2005). This choice follows the approach of Giuntoli et al., (2015b) of selecting pairs catchment/grid-cells of comparable size to deal with the misalignment between model and observational data. Because no land use changes or water management interventions are accounted for in the modeled data, the streamflow gauges were selected from the Hydro-Climatic Data Network, the reference set of streamflow gauges with historical data responsive to climatic variations, so relatively free of anthropogenic influences (Whitfield et al., 2012). The main characteristics of the streamflow gauges are presented in Table S1 in the Supporting information.

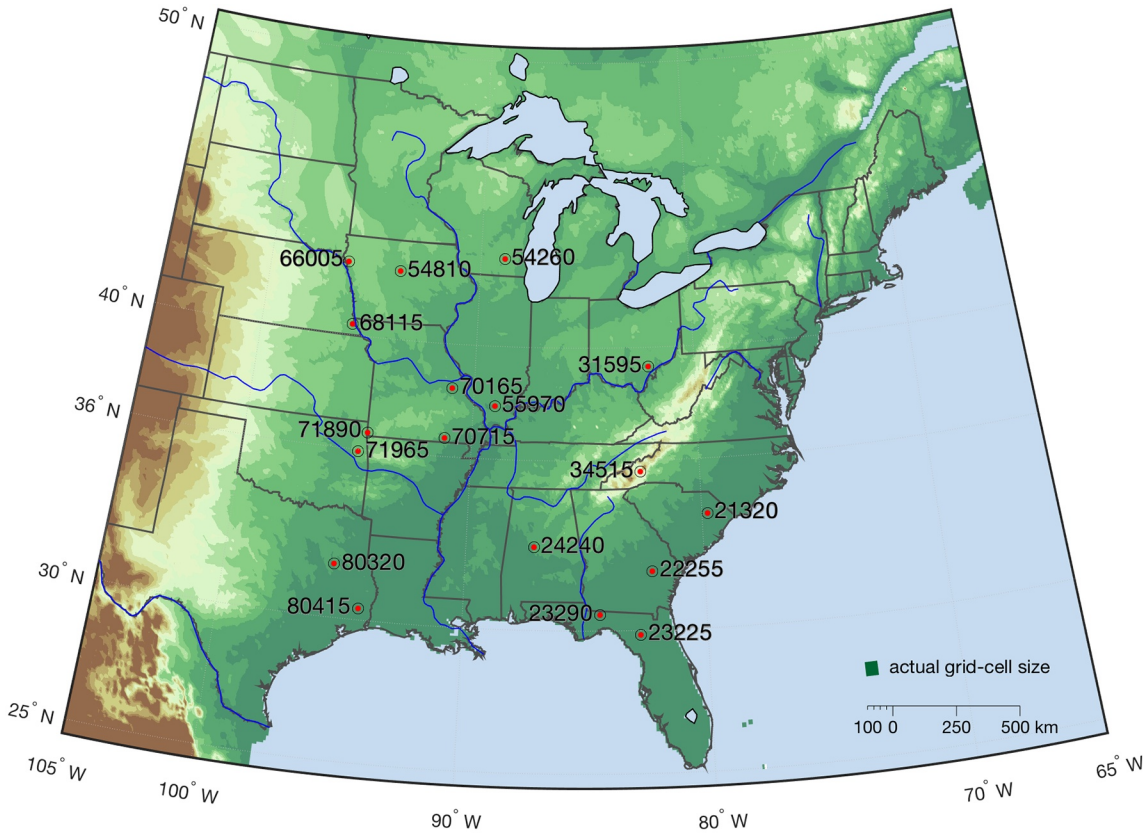

Figure 1. Map of the 18 streamflow gauges noted with their USGS code (eluding the last two digits 00). On lower right, above the scalebar, the actual grid-cell size (0.5°×0.5°) is shown in green.

For global models AMax, we use daily runoff outputs from the ISI-MIP Fast Track (Warszawski et al., 2014) comprising an ensemble of nine GIMs forced with five CMIP5 GCMs' bias-corrected climate (Hempel et al., 2013) in their control (1971–2005) and future (2065–2099) periods under the RCP8.5 scenario (i.e., 45 runs per grid-cell). The GCMs have been evaluated by McSweeney and Jones (2016). All GIMs were run at a spatial resolution of 0.5 decimal degrees (with the exception of JULES whose resolution is 1.25 × 1.875°). Models vary in structure (physical processes), parameterization, and time step; we provide a brief overview of the set of models and main characteristics in Table S2 of the Supporting information. Giuntoli et al. (2018) provide detailed information on model characteristics and evaluation.

2.1. Appraisal of Simulated versus Observed Peak Flows

We compare observed and modeled peak magnitude (AMax) and timing (AMaxDate) at the 18 locations highlighting discrepancies between observed and modeled data. Observed-modeled differences are to be expected and point to the nontrivial task of reconciling the two worlds, especially when dealing with extremes (Seneviratne et al., 2012).

2.1.1. Peak Flow Distributions

We compared raw peak flow time series from observed and modeled data using non-parametric tests (no assumption is made on the type of distribution) assessing: (1) same distribution (Kolmogorov-Smirnoff, noted KS, (Massey, 1952)), (2) equal median (Wilcoxon rank-sum, noted W, (Wilcoxon, 1945)), and (3) equal variance (Ansari-Bradley, noted AB, (Ansari & Bradley, 1960)). There is little overlap between observed and modeled peaks in terms of distribution (KS, 9.3% of runs) and medians (W, 11.9% of runs), while for the variance there is good agreement (AB, 84.4% of runs). Interestingly, testing modeled data from historical to

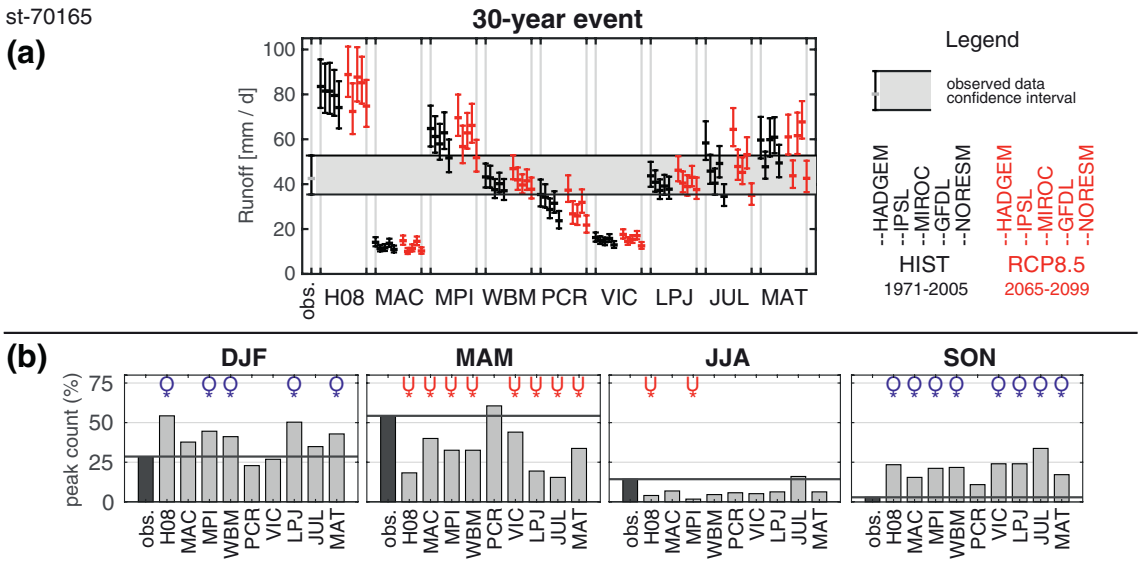

Figure 2. Comparison of observed-modeled magnitude (a) and timing (b) of annual maxima: (a) confidence intervals (95%) of observed data (gray band) and global climate-global impact model (GIMs-GCMs) combinations in their historical (black), and future (red) periods for the 30 years event; (b) Average peak flow occurrence per season. Bars indicate percentage of peak counts for observed (black) and modeled (gray) data. Horizontal black lines correspond to the observed peak counts (the reference). Each GIMs comprises five GCM runs. Blue (red) flags indicate over (under)—estimation of peak counts ≥ (≤) 20%.

future period (RCP 8.5) yields greater agreement across the three tests (KS 66%, W 69%, and AB 90%) than seen with the observed peak flows, as reported in Table S3 of the Supporting information.

2.1.2. Peak Flow Magnitude

In addition to testing raw peak flows we compared observed and modeled peak flows Gumbel fits—with location and scale parameters estimated via joint maximum likelihood and confidence intervals via profile likelihood (Coles, 2001). Figure 2a depicts, for one of the sites (Bourbeuse River at Union, MO), a plot of return levels for the one in 30 years event and corresponding 95% confidence intervals: the horizontal gray band shows the observed data, that is, the reference to which the historical period of the models (black lines) should tend to align, while the red colored lines correspond to the future period under scenario RCP8.5 (plots for all sites are in Supporting Information, Figures S1 and S2). While few models overlap the observed data confidence intervals, others lie well outside them (i.e., H08, MacPDM, and VIC combinations). Interestingly, the return levels resulting from the models tend to cluster per GIM, indicating that the GCMs tend to follow the peak magnitude described by the GIMs.

2.1.3. Peak Flow Timing

Peak flow timing in all sites tends to be overestimated in the winter and underestimated in the spring and to a smaller degree in the summer. This is noticeable when sorting peak counts into four seasons, winter DJF, spring MAM, summer JJA, and autumn SON, as shown in Figure 2b. Generally, in northern sites the autumn is overestimated too, while in southern sites SON peak counts are in line with observed data (Figure S3 in Supporting information). Overall, MacPDM, PCRGlob-WB, and VIC are the GIMs that capture timing of peak flows best, while, H08, LPJmL (north, especially), and MPI-HM (south, especially) struggle to replicate the right timing of peak occurrences. Furthermore, models generally anticipate peak occurrence (in Figure S4 of the Supporting information colored vectors, showing the median of the peak's date per GIM, are constantly indicating earlier dates than the observed peaks i.e., the black vector). In particular, in the north peaks occur from March to May, whereas models show a systematic shift of approximately one month earlier, with peaks occurring from February to April. In the south peaks occur from February to March-April, whereas models systematically anticipate occurrences to February with a few exceptions. In addition to clear time shifts of one or two months, at some sites modeled peaks occur in absence of corresponding observed peaks.

This modeled-observed comparison provides insight for creating a constrained ensemble version (*cE*)—detailed in Section 3.2—obtained by excluding models that capture poorly the timing of observed peak flows, which proved to be a suitable discriminant factor.

3. Methods

3.1. Statistical Analysis Framework

This section describes the statistical framework used to assess changes in future floods and their uncertainty. First (Section 3.1.1), we present the Bayesian hierarchical model used to analyze the flood peaks, and second (Section 3.1.2), we provide further detail on Bayesian inference and hierarchical models.

3.1.1. Modeling of Extreme Values

The relationship between the frequency and magnitude of high flows (Flood Frequency Analysis, FFA) is assessed often by estimating a statistical distribution for annual maxima. Although, extreme value theory indicates that the Generalized Extreme Value (GEV) distribution should be the limiting distribution of annual maxima (see Coles, 2001), the suitability of specific distributions for a given peak flow record is a topic of active research, and different distributions are recommended as standard in different countries: for example, LP-III for the United States (England Jr. et al., 2018), GLO for the UK (Institute of Hydrology, 1999), and more recently the Burr has been suggested for Canada (Zaghloul et al., 2020).

For the purpose of this investigation, runoff outputs of grid cells located at corresponding gauging stations are used as the variable of interest, thus mimicking an at-site analysis. For each grid-cell a Gumbel distribution with a specific time-dependent model presented below is employed. The Gumbel distribution, which corresponds to a GEV distribution when the shape parameter tends to 0, has a long history of application for the FFA and it is used routinely (Bertola et al., 2019; Castellarin et al., 2012). With the aim of identifying changes in the distribution of annual maxima, a simpler two-parameter distribution was preferred to avoid the hurdle of correctly estimating shape parameters, which are highly variable (Papalexiou & Koutsoyiannis, 2013) and arguably of little interest in the context of our analysis, especially considering that we do not wish to estimate actual design events of rare frequency. The Gumbel distribution was found to fit the data well (as in e.g., Hirabayashi et al., 2008; Lim et al., 2018) and was therefore adopted as the parent distribution for the grid runoff outputs. Its probability density function (pdf) is defined as:

$$\frac{1}{\theta}\exp\left\{-\frac{x-\xi}{\theta}-\exp\left\{-\frac{x-\xi}{\theta}\right\}\right\} \tag{1}$$

where $\xi \in R$ denotes the location parameter and $\theta \in R^{+}$ denotes the scale parameter.

Rather than fitting separate Gumbel distributions to each model run (as in e.g., Alfieri et al., 2018; Dankers et al. 2014), a hierarchical approach is employed in which data from all runs are modeled together. This allows for a clear partition of the variance of data into different components, thus highlighting the contribution from the GCM and the GIM components and their interaction to total variability: this gives an indication of the major source of uncertainty in future high flows. Moreover, by modeling all data together, it is possible to obtain an estimate of the overall difference between the future runs and the historic runs across all model runs. Figure 3 outlines the key components and steps of the statistical framework used in this study: for the 45 time series of historical and future flow (resulting from the combination nine GIMs and five GCMs) a unique model is estimated and measures of future changes and of the contribution of the GCM and GIM components to the overall variability are derived. The model assumes that the data (both present and future) follow a Gumbel distribution in which the scale parameter is the same in both time windows while the location parameter is allowed to take two different values: one for the historical and one for the future periods—while it is assumed to be constant within each time period. This is in line with the non-stationary extreme value analysis literature where models in which the location, rather than other parameters, is allowed to change are common—see Salas et al. (2018) and references therein. Indeed, models that attempt to explain changes in the distribution of extremes by allowing higher order parameters to vary are rarer than models in which the location is allowed to change: higher order parameters tend to be more variable and therefore harder to estimate accurately, especially when the samples under study are

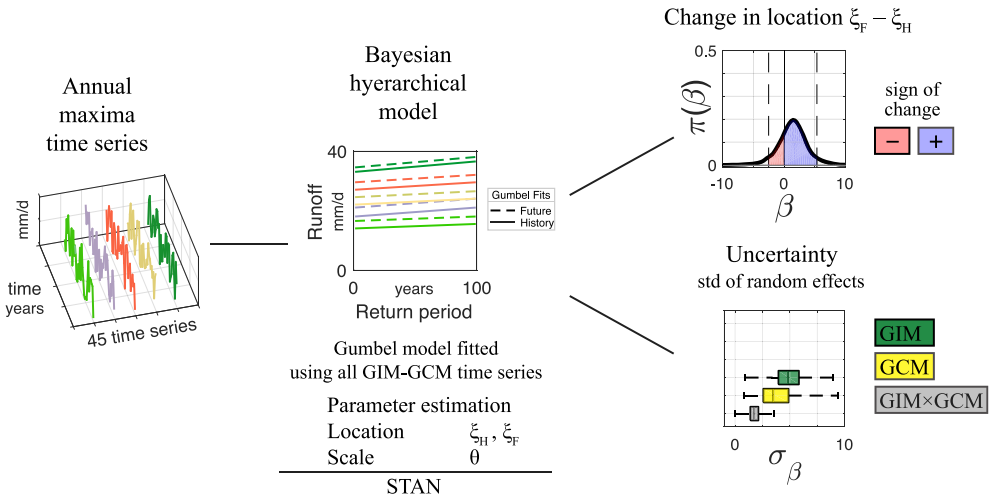

Figure 3. Flowchart of the statistical analysis framework. In the first two graphs from the left colours refer to the AMax time series of the GIM-GCM combinations (for explanatory purposes, five out of 45 are shown) and corresponding continuous (historical) and dashed (future) Gumbel fits.

not very large. The accurate estimation of models, which allow for more structure in the scale parameters, would require very large samples and very sizable changes in the scale parameters. The model structure was determined by a model selection procedure outlined in Section S3.1 following Vehtari et al. (2017): while models of increasing complexity were used for both the location and the scale parameter, the final model presented below adopts a more complex model for the location parameter and a relatively simple form for the scale parameter.

More formally, let $y_{i,j,k,h}$ be the h^{th} annual maximum flow value obtained from the i^{th} GCM combined with the j^{th} GIM, which results in the k^{th} GCM-GIM combination. Since all GCMs feed every GIM there are $5 \times 9 = 45$ combinations of GCM-GIM output.

It is assumed that $y_{i,j,k,h}$ follow a Gumbel distribution: $y_{i,j,k,h} \sim \text{Gumbel}\,(\xi_{i,j,k,h},\theta_{i,j})$ where the following model structures have been assumed for, respectively, the location and scale parameter:

$$\begin{aligned}\xi_{i,j,k,h} &= \alpha + \alpha_{\text{gcm},i} + \alpha_{\text{gim},j} + \alpha_{\text{comb},k} + \\ &\beta * \mathrm{I}_{[36,70]}(\mathrm{h}) + \beta_{\text{gcm},i} * \mathrm{I}_{[36,70]}(h) + \beta_{\text{gim},j} * \mathrm{I}_{[36,70]}(h) + \beta_{\text{comb},k} * \mathrm{I}_{[36,70]}(\mathrm{h})\end{aligned} \tag{2.a}$$

$$\theta_{i,j} = \exp\{\gamma + \gamma_{\text{gcm},i} + \gamma_{\text{gim},j}\} \tag{2.b}$$

with $i = 1,...,5, j = 1,...,9, k = 1,...,45$, and $h = 1,...,70$. $I_{[36,70]}$ (h) is an indicator variable that takes value 0 when the data point is in the historical period (i.e., $1 \le \mathrm{h} \le 35$) and 1 in the future period (i.e., $35 < \mathrm{h} \le 70$). The α parameters indicate the intercept for the location, the β parameters indicate the time-effect for the location and the γ parameters indicate the intercept for the scale.

The parameter α in Equation 2.a represents the overall population-level value for the intercept parameter of the location across all model combinations. To accommodate the variability across the different models three group-specific terms have been included: $\alpha_{\text{gcm},i}$ to allow for the variability across the GCMs; $\alpha_{\text{gim},i}$ to allow for the variability across the GIMs; and $\alpha_{\text{comb},k}$ to allow for the variability across each GCM and GIM combination. By comparing the different values of $\sigma^2_{\alpha,\text{gcm}}$, $\sigma^2_{\alpha,\text{gim}}$, and $\sigma^2_{\alpha,\text{comb}}$, it is possible to assess which grouping variable explains the largest proportion of variability (i.e., uncertainty) in the AMax values. Notice that the factor describing the combination of GCM and GIM is only included for the location parameter

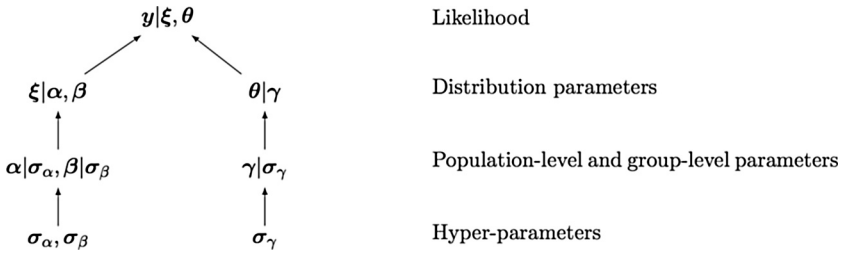

Figure 4. Structure of the Bayesian hierarchical model.

model. The inclusion of this factor has been found to improve the fit of the model prediction to the data, and was deemed useful to describe the interaction between different GIMs (applied to different areas of the continent and which might require different input variables) and the GCMs, which reproduce the different climate components in a very different fashion. The interaction between the two factors can be already guessed in Figure 2a, in which clusters of estimated design events are not fully explained by the GIM or the GCM under which the data were generated, but exhibit some further variability.

The parameter β represents the overall population-level change in location parameter when moving from the historic period time window to the future time window. The parameter quantifies the overall average difference between the location parameter in the two time periods across all model combinations. The $\beta_{gcm,i}$, $\beta_{gim,j}$, and $\beta_{comb,k}$ are group-specific effects that allow for each GCM and GIM and combination to have a different slope (i.e., a different location value in the two time windows) from the overall population-wide time-window effect β. The relative contribution of each component on the time effect for the location of the distribution is assessed by comparing the variance of the group-level slopes. The model structure for the scale parameter in equation 2.b is simpler than the one for the location parameter as it considers only the intercept (while the location also considers the slope) and two group-level parameters $\gamma_{gcm,i}$ and $\gamma_{gim,j}$ that allow for the group-wise variation around the overall population-level γ. Note that an exponential link function is employed in the scale parameter model to ensure that the function only takes positive values. The population-level parameters (in this model α, β, and γ) can be referred to as fixed effects, while the group-level parameters (in this model $\alpha_{gcm,i}$, $\alpha_{gim,j}$, $\alpha_{comb,k}$, $\beta_{gim,i}$, $\beta_{gcm,j}$, $\beta_{comb,k}$, $\gamma_{gcm,i}$, and $\gamma_{gim,j}$) can be referred to as random effects, assumed as normally distributed and with common variance. We use a Bayesian approach to the estimation of the model parameters (see Section 3.1), in which all model parameters are viewed as random variables therefore the terminology of population-level and group-level parameter is preferred (Gelman et al., 2013).

3.1.2. Bayesian Hierarchical Model

The model structure presented in Equations 2 is that of a multilevel model in which the annual maxima within a level (group) of a grouping variable (e.g., peak flows generated with the same underlying GIM) shares a common feature and have greater within-group similarity with respect to peak flows from the other groups. Thus, the variation in the data are decomposed into the individual observation variation and the variation of the levels of each grouping variable. These types of models are called hierarchical models, multilevel models or random-effect models and have enjoyed a great success in several fields of application (see Gelman & Hill, 2006). For instance, Northrop and Chandler (2014) proposed the use of multilevel model to quantify the sources of uncertainty in climate projections, highlighting the connection between the multilevel approach and the ANOVA approach used in for example, Yip et al. (2011).

A Bayesian approach allows for a straightforward estimation of multilevel models in which all uncertainties can be properly taken into account (see Gelman et al., 2013). A schematic form of the hierarchical structure of the statistical model employed is outlined in Figure 4.

Taking $y = (y_{1,1,1,1},\ldots,y_{5,9,45,70})$ to represent the vector of all annual maxima and $\eta = (\alpha,\beta,\gamma,\alpha_{gcm},\alpha_{gim},\alpha_{comb},\beta_{gim},\beta_{gcm},\beta_{comb},\gamma_{gcm},\gamma_{gim})$ to represent the vector of all model parameters, by virtue of Bayes' rule we have:

$$p(\eta \mid y) \propto p(y \mid \eta) * p(\eta), \tag{3}$$

where p (y|η) is the model for the distribution of the data conditional on the parameter η (i.e., the Gumbel distribution with a model structure specified in Equations 2.a and 2.b) and $p(\eta)$ is the prior distribution of η that needs to be specified and which encodes the beliefs about the distribution of the model parameters before any data is taken into account. Finally, $p(\eta|y)$ is the posterior distribution of η conditional on the annual maxima y: this represents the understanding of the distribution of the model parameters after the available data has been taken into account and is typically the quantity of interest in Bayesian inference.

Given the hierarchical multilevel structure of the model, a further layer of hyper-parameters (ϕ) that characterizes the prior distribution $p(\eta)$ needs to be specified so that $p(\eta) \propto p(\eta|\phi)*p(\phi)$. Here ϕ is the vector of the variances of the random effects: $\phi = (\sigma_\alpha, \sigma_\beta, \text{and } \sigma_\gamma)$. By applying again Bayes' rule we have that:

$$p(\eta,\phi \mid y) \propto p(y \mid \eta,\phi) * p(\eta \mid \phi) * p(\phi) \tag{4}$$

where $p(\eta,\phi|y)$ denotes the posterior joint distribution of the model parameter and the hyper-parameters, which is the quantity of interest in Bayesian multilevel models. The posterior distribution $p(\eta,\phi|y)$ cannot be obtained in a closed form and therefore needs to be estimated, typically using Montecarlo approaches in which the distribution is derived using a computer-simulation. In particular Stan (Stan Development Team, 2018), a state of the art probabilistic programming language for statistical modeling, was used to derive the posterior distribution for the parameters of the model presented in equation 2.a and 2.b and the hyperparameters defining their distributions. A sample Stan code employed in the estimation procedure is provided in Section S3.3 of the Supporting Information—the code was derived from the brms R package (Bürkner, 2017).

Following the recommendations in Gabry et al. (2019) informative priors were used for the hyper-parameters in the model and their suitability was verified via prior-predictive checks: using very wide, that is, uninformative, priors can results in excessively variable data. In particular, prior distributions were determined using information on the time series of each grid cell (i.e., sample mean and standard deviation). The sensitivity of the model estimates to the prior was investigated by attempting to estimate the models under study using several prior specifications. The model estimation was found to be mostly insensitive to different prior choices, provided that informative priors, which limit the potential variability of the data generating process, are used. The specification on the prior distributions can be found in Section S3.2.

Although, the use of multilevel models to partition the variability of modeled climate variables (Northrop & Chandler, 2014) has already been proposed, the uptake of these methods in the literature has been minor. In this work, we advocate that their use can deliver key information using a unified model: the overall direction of change and the information of which component of the modeling chain contributes the most to the signal variability. The computational burden connected to the implementation of these models has been greatly reduced by the availability of general purpose efficient probabilistic programming languages such as Stan, allowing for a fast and stable implementation of more informative models.

3.2. Constraining the Ensemble

As stated in Section 1, we create a constrained ensemble (*cE*) at each site by excluding models that simulate observed peak flow characteristics poorly. Forming this ensemble requires a level of informed subjectivity and is hindered by the striking discrepancies between observed and modeled values. Indeed, in Figure 2a, it would be expected that model data in the historical period (in black) overlaps the confidence interval (gray band) of the observed data, whereas in the majority of cases this hardly occurs (see Figures S1 and S2 in the Supporting Information). A model selection based on return levels rejects the vast majority of models and constitutes, perhaps, an overly stringent criterion. It should be noted that this ground-truthing effort is carried out on total (surface plus subsurface) unrouted runoff, so models cannot be expected to replicate accurately the actual quantities observed at the streamflow gauges (Giuntoli et al., 2015b; Gudmundsson et al., 2012). Furthermore, it has been emphasized how the model's capacity to simulate flood timing is an important metric to represent flood generation processes (Collins, 2019; Do et al., 2020). Therefore, we

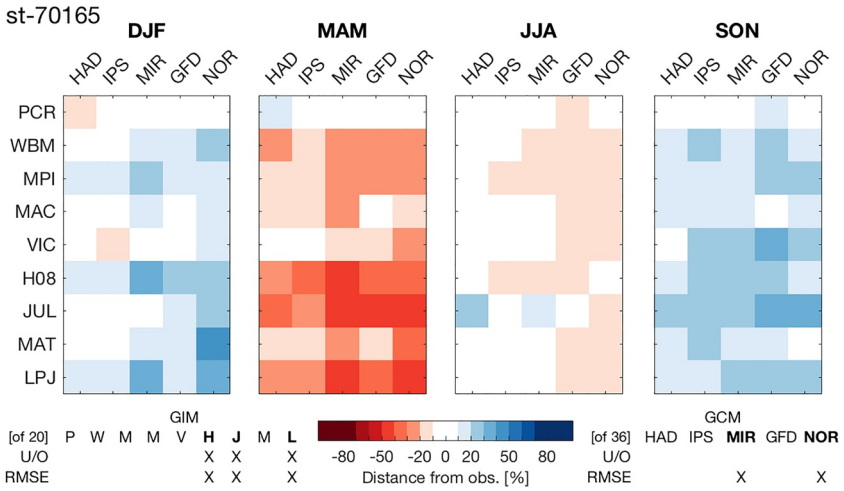

Figure 5. Departure (%) from average observed peak flow (AMax) occurrences per season. Individual global impact (GIMs) and global climate (GCMs) models are expressed in row and column, respectively. Red (blue) tones indicate under (over)–estimation ("U/O") of peak counts ≥ (≤) 10%. Model exclusions (GIMs lower left, GCMs lower right) are denoted with X.

constrain the ensemble on the basis of how well peak flow timings are simulated in the control period. To do this, we use two metrics to compare observed and modeled peak counts: (1) the distance between the proportion of seasonal counts of observed and modeled peaks (2) RMSE (root mean squared error) of counts. The steps for identifying and excluding GIM-GCM combinations (45) at each site are detailed below.

1. Observed peak timings are sorted into four seasons (DJF, MAM, JJA, and SON), and constitute the reference. For example, the site in Figure 5 over the 35 years the peaks amount to: 10 in DJF, 19 in MAM, five in JJA, one in SON.
2. Same as step 1 for simulated peak timings. For example, the site in Figure 5, for the JULES GIM fed by the HadGEM2-ES GCM peak counts are: seven in DJF, seven in MAM, 12 in JJA, nine in SON. Note that the comparison is done on the GIM-GCM combination output.
3. Counts in step 1 (observed) and step 2 (modeled) are expressed in percentage. A negative score is assigned to those GIM-GCM combinations whose proportion is more than 20% apart from the observed proportion. For example, counts of step 1 are: DJF = 28.6%, MAM = 54.3%, JJA = 14.3%, SON = 2.9%; while counts of step 2: DJF = 20%, MAM = 20%, JJA = 34.3%, SON = 25.7%. In this case there are three negative scores with distances above the 20% threshold: MAM-dist = |54.3–20| = 34.3, JJA-dist = |14.3–34.3| = 20, SON-dist = |2.9–25.7| = 22.8.
4. Negative scores described in step 3 are counted for all combinations (i) in row for excluding GIMs when the negative score is assigned to at least 10 out of 20 season count records (i.e., half of the cases); (ii) in column for excluding GCMs when the negative score is assigned to at least 18 out of 36 season count records (i.e., half the cases).
5. We consider the RMSE (root mean squared error) comparing the vector of seasonal peak counts (step 2) for each GIM in row (of length 5) and each GCM (of length 9) to a vector formed by the observed data counts (step 1) replicated to match the vector length to be compared to.
6. The threshold value of acceptance for the RMSE is set to the 90th percentile of all comparisons (11.1); model combinations above it in any of the seasons are thus excluded from the constrained ensemble.

Meeting any of the two conditions, that is, distance between the proportion of seasonal counts and RMSE, yields exclusion of the model from the ensemble.

In Figure 5 peak timing distances and exclusions are shown for station n. 70165: negative (positive) overshoots, denoted as "U/O" (under/over) are depicted in red (blue). Upon threshold crossing, model exclusions are denoted with "X" on the lower left the GIMs, on the lower right the GCMs. For instance, the JULES GIM is excluded because its series have seasonal proportion of peaks that are distant from that of observations more than 10 times (one time in DJF, five in MAM, one in JJA, and five in SON); it also crosses

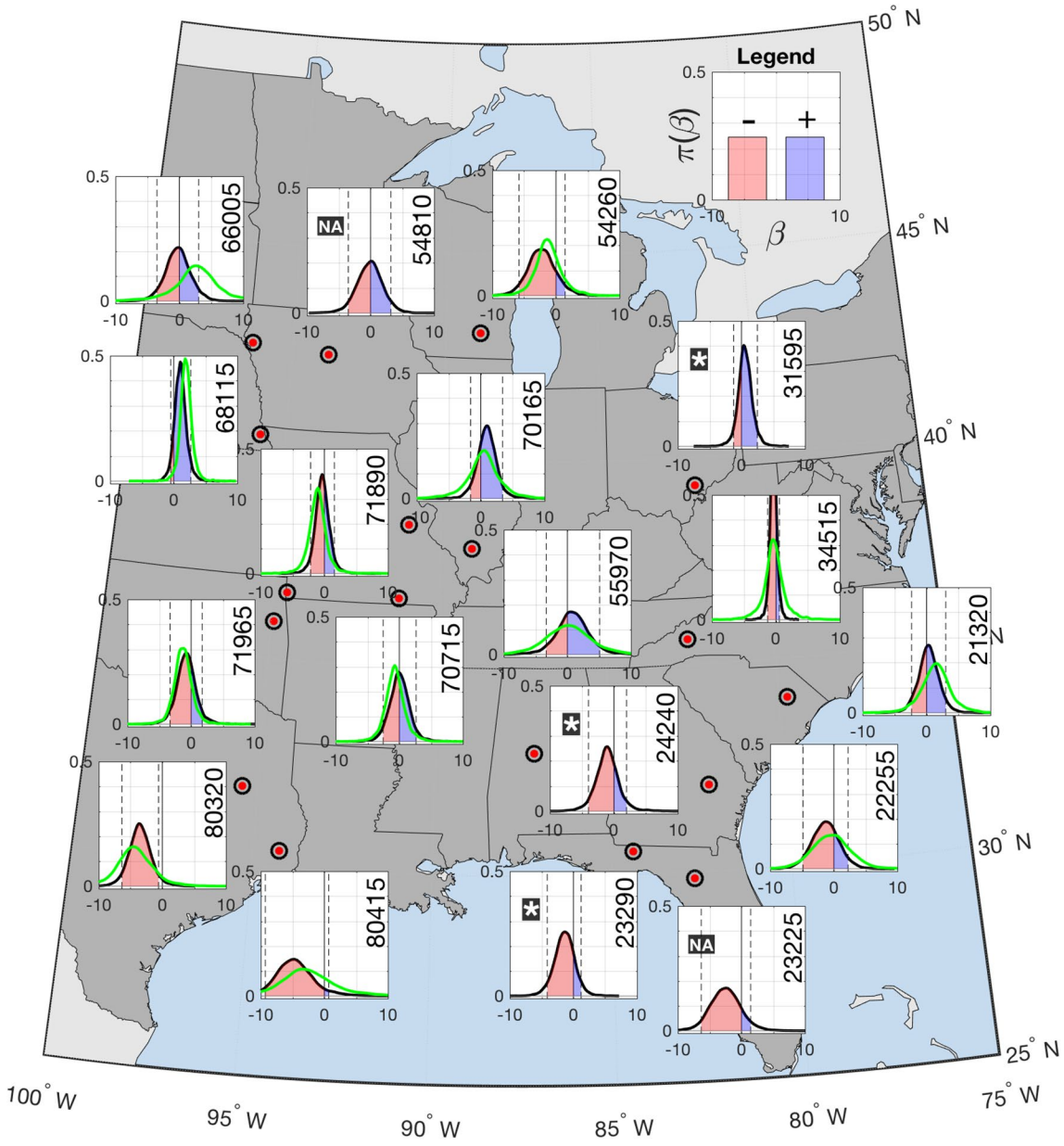

Figure 6. Posterior distribution of β (the parameter that describes the change in the location parameter in the future) of the full ensemble, *oE*. Shaded blue (red) depicts positive (negative) values; solid vertical line corresponds to 0, dashed lines correspond to the 95% credible intervals. The fluorescent green pdf refers to the constrained ensemble, *cE*. Inset plots with star "*" indicate same results as *oE*, while plots with "NA" indicate no *cE* results available.

the RMSE threshold in MAM and SON. At the same time, the MIROC-ESM-CHEM GCM, is not excluded for distance counts but because it has a RMSE above threshold in MAM. Plots for all sites are shown in Supporting Information Figures S5 (northern sites) Figure S6 (southern sites), and Figure S7 (two sites excluded), with the *cE* composition summarized in Table S4.

4. Results

The at-site changes in magnitude of future annual maxima (as outlined on the right-hand side of Figure 3) are illustrated in Figure 6 as changes in the estimate location parameter of the Gumbel distribution, that is, the difference between the future (2065–2099) and the historical (1971–2005) periods. Second, Figure 7 illustrates the corresponding uncertainty contribution coming from GIMs (green), GCMs (yellow),

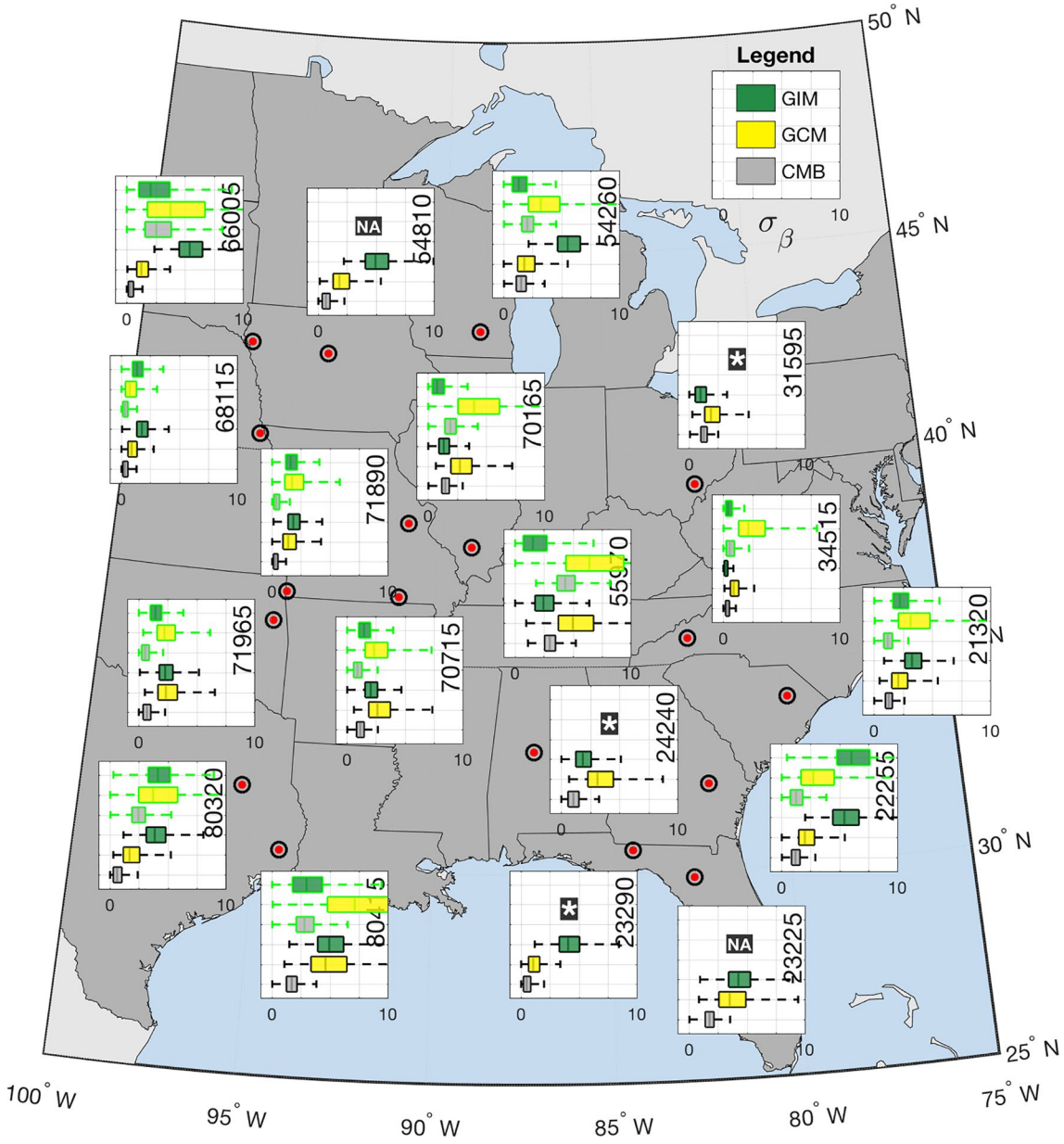

Figure 7. Standard deviation of the random effects expressing main contributions to uncertainty in the changes due to GCM (yellow), GIM (green), global climate-global impact model (GCM-GIM; gray) for the ß (time-window effect) of the location parameter. Lower three boxplot refer to the *oE*, while the upper three boxplot to the *cE* (fluorescent green). The higher the boxplot value, the higher the contribution to uncertainty. Inset plots with star "*" indicate same results as *oE*, while plots with "NA" indicate no *cE* results available.

and their interaction (gray), shown as boxplot of the random effects' standard deviation posterior sample. Table 1 summarizes overall direction of changes in magnitude and the corresponding dominant source of uncertainty (based on details in Figures 6 and 7). Finally, we discuss results using a constrained ensemble (*cE*) obtained by reducing the full ensemble (*oE*) having compared modeled and observed metrics—as detailed in the previous Section 3.2.

4.1. Full Ensemble

Our finding demonstrates clear spatial variability that characterizes changes in the annual maxima (Figure 6). As it is the case for other extremes like precipitations, changes in AMax are unlikely to be uniform

Table 1
Summary of the Changes in the Magnitude of AMax (Seen in Figure 6) and Corresponding Dominant Source of Uncertainty in the Full (oE) and the Constrained (cE) Ensemble

Station num (from N to S)	Location change (F–H)		Uncertainty major contributor		
	oE	cE	oE	cE (out of 9 GIM, 5 GCM)	
1. 54260			GIM	GCM	6 GIM 4 GCM
2. 66005			GIM	GCM	4 GIM 2 GCM
3. 54810		NA	GIM	NA	1 GIM 1 GCM
4. 68115		+	GIM	GIM	8 GIM 4 GCM
5. 31595[a]		[a]	GCM	[a]	All
6. 70165			GCM	GCM	6 GIM 3 GCM
7. 55970			GCM	GCM	6 GIM 3 GCM
8. 70715			GCM	GCM	7 GIM 4 GCM
9. 71890			GIM	GIM	8 GIM 4 GCM
10. 71965			GIM	GCM	8 GIM all GCM
11. 34515			GCM	GCM	6 GIM 3 GCM
12. 21320			GIM	GCM	7 GIM 3 GCM
13. 24240[a]		[a]	GCM	[a]	All
14. 22255			GIM	GIM	7 GIM 3 GCM
15. 80320	–		GIM	GIM	8 GIM 4 GCM
16. 23290[a]		[a]	GIM	[a]	All
17. 80415			GIM	GCM	7 GIM 3 GCM
18. 23225		NA	GIM	NA	None

Legend

Location change				Dominant uncertainty source				
	Negative	no ch.	Positive	Prominence	GIM		GCM	
Overall ch.				Low				
Significant ch.[b]	–		+	High				

[a] Full and constrained ensembles are the same. [b] The credible intervals of beta lie all below or above zero.

Notes. Changes are positive (negative) if the interquartile range, i.e. middle 50%, lies above (below) zero, and grey that is, no change otherwise. The dominant source of uncertainty, (seen in Figure 7) is coloured depending on the distance from the other sources, that is pale (bright) coloured when there is high (low) overlap—its interquartile range does (not) overlap that of the other sources of uncertainty.

across even small geographic areas (Schoof & Robeson, 2016). Nevertheless, the changes in flood magnitude (Figure 6) over the 18 sites considered herein do show some consistent regional patterns. Starting from the South, with the exception of one location (21320) with no predominant sign of change, all nine southern locations (south of parallel 36°N) show a negative change, with one that is significant (95% credible intervals all lie below zero). This indicates a consensus of the models on a general decrease in future flood magnitude over the southeast United States, a result that is consistent with other regional studies using global model projections (Naz et al., 2016). Conversely, for the other nine locations in the northern half of the domain, there is no clear pattern of change, although a consensus exists among models at some locations like sites 68115 in the west and 31595 in the east, which exhibit spiked pdfs with higher $\pi(\beta)$ values.

Wider pdfs in the southern and northernmost locations, may be the result of increased model spread that can be explained by the difficulties in simulating evaporation and recharge processes in semi-arid zones and wetlands of the south (Trigg et al., 2016); and by the high uncertainty in simulating ice and snowmelt processes, the GIMs especially, in the North (e.g., the sites in the northern Midwest; Giuntoli et al., 2015b).

The uncertainty in the changes coming from the GIMs, the GCMs or the interaction between both are shown in Figure 7, while in Table 1, as a summary, the major source of uncertainty is colored depending on the distance from the other sources, that is bright (pale) colored when there is low (high) overlap. A striking feature is that if there is a clearly dominant source (i.e., little overlap with a boxplot distinct from

the other two), this source is always the GIMs and it happens there where the changes have the largest spreads (i.e., wide pdf). This may be explained both by the aforementioned difficulties of the GIMs in simulating runoff and by the GCMs' uncertainty being at least partly attenuated by the bias correction they all underwent prior to feeding the GIMs (Hagemann et al., 2013). Also, the presence of a GCM uncertainty dominated southwest-northeast band indicates that the locations situated more inland, are less driven by GIM uncertainty, perhaps for being less exposed to ice-cold winters as in the north or atmospheric circulation patterns originating in the Atlantic as in the southeast. Overall, the major effects are mostly explained by the GCM and GIM sources while the remaining effects are explained, at least partly, by the combination between the two sources (in gray), which is smaller in the majority of cases. This is to be expected and points to the validity of the statistical model employed. In fact, with an inadequate model the combination source might explain most of the random effects, leaving little uncertainty to the main sources (GIMs and GCMs).

Given the complexity of the mechanisms driving floods at the regional scale, unraveling the causes of the different magnitudes or the directions of change in different models remains elusive. If on the one hand GCMs are responsible for regional runoff biases due to uncertainties in the representation of precipitation and sub-grid soil infiltration and flow; on the other hand the GIMs' total runoff include contributions from surface runoff—function of saturation (SE) and infiltration excess (IE)—and subsurface runoff—function of impermeable area and water table depth (Kooperman et al., 2018). For instance, throughout the domain of study portions of Texas, Louisiana, Kansas, Missouri, and Iowa are more likely dominated by IE runoff; on the other hand SE runoff is more likely in the southeast (e.g., Florida, south Georgia) and coastal areas of the Great Lakes region (Buchanan et al., 2018). The prevalence of IE or SE excess runoff depends on the type of soil and its capacity to become saturated/infiltrate. A sandy soil in the southeast will yield a higher flux (i.e., will transmit water faster) than a clayey soil under a given hydraulic gradient, reducing the effects of high-intensity precipitation. While runoff generation plays a role in flood generating processes and therefore in models simulation spread, it should be noted that all nine GIMs consider SE only, except three (PCRGlobWB, MATSIRO, and JULES) that also consider IE in their runoff schemes (as noted in Table S2 of the Supporting Information). Over the eastern half of the United States, this may represent a limitation provided that a considerable share of the area is IE dominated, therefore capturing the precipitation intensity dependence does matter in generating floods.

4.2. Constrained Ensemble

As seen in Section 2.1, runoff annual maxima from global models differ systematically from observed data in terms of distribution and medians. With only few exceptions, the majority of the models struggle to reproduce return period point and confidence estimates of observed AMax even at time spans for which extrapolations are relatively small, that is, return period of 30 years. For this reason, the constrained ensemble (*cE*) was based on model adequacy in simulating timing of peak flows throughout the year. Thus, model selection was carried out at-site excluding GIMs and GCMs with considerable departures from observed seasonal peak counts. This yields constrained sets that comprise on average 55% of the members of the full ensemble (see Table S4). It should be noted that while three sites have equal *oE* and *cE* configurations as they underwent no member exclusions, two sites have no *cE* version as they were left with too few members (zero or one, as shown in Figure S7).

In constraining the ensemble, the exclusion of GCMs is generally widespread across the domain of study, with the MIROC-ESM-CHEM and NorESM1-M models being excluded more often. GIMs are excluded more in the northern stations than in the southern ones (approximately 2 vs. 3 exclusions average, respectively out of 9), this can be explained by the increased difficulty in simulating cold climates processes like snowmelt and ice formation. More specifically, the H08 and JULES GIMs are the more often excluded across the whole domain, and LPJmL in the northern stations. Interestingly, H08 and JULES are GIMs that try to close the energy balance and have shown, under a different setup, larger temporal lags in timing of peak flows compared to GIMs that do not close the energy balance (Giuntoli et al., 2015b). Also, JULES and LPJmL simulate CO_2 dynamics while the other models do not (Davie et al., 2013) and their runs show a wet bias along with an over (under) -estimation of flood peaks in the winter (spring) period in the north

of the United States. Indeed, simulating plant physiological responses to rising CO_2 can yield considerably different results as higher CO_2 can reduce stomatal conductance and transpiration, which may lead to increased soil moisture and runoff in some regions, favoring flooding even without changes in precipitation (Kooperman et al., 2018).

Are results affected by the different composition in the GIM/GCM matrix of the *cE* with respect to the *oE*? Changes in flood magnitude obtained with the *cE* (Figure 6, in fluorescent green) are similar to those of *oE* with a consensus on negative change in the south of the domain, while the few positive changes actually increase (e.g., the stations in the northwest of the domain). Constraining the ensemble at-site yields essentially the same results as using the whole ensemble, although using almost half the runs. A slight change is noticeable in the shape of the pdfs, which tends to be less concentrated (smoother peaks), as if more members of the *oE* increase confidence in the estimate.

If the changes in magnitude remain similar in *oE* and *cE*, as the *cE* is composed by fewer members, this is reflected in the different contributions to uncertainty, with boxplot that tend to become wider, especially the GCM ones (Figure 7). In the *oE*, the northern and southern sites are GIM dominated (Figure 7 and Table 1); while for *cE*, this predominance tends to lose strength in favor of the GCM, especially in the very north of the domain, consistent with Giuntoli et al. (2018). Interestingly, never do GCM dominated sites become GIM dominated indicating that constraining the ensemble tends to reduce more the GIM than the GCM contribution to uncertainty, although the boxplot are often quite wide, resulting perhaps from fewer runs employed on average.

5. Discussion and Wider Implications

The inherent tendency to disagree on the absolute value or on the sign of projected changes of climate variables like precipitation and runoff in global model runs adds to the fact that generally these runs do not match observations well (Do et al., 2020). Therefore, estimates of future precipitation and runoff changes suffer from large uncertainty and from a signal that may be canceled out as different model simulations are averaged to generate a final value that is often taken as the ensemble mean (e.g., Dankers et al., 2014; Ragno et al., 2018; Wobus et al., 2017).

The aim of this paper was to propose a novel framework that allows for estimating the changes in future flood magnitude with the signal of the direction of change expressed as the distribution of the overall change rather than the ensemble mean. We quantified these changes by modeling the extreme values parameters using all multi-model ensemble simulations (GCM-GIM) at once and characterizing the uncertainty from both GCMs and GIMs as the variations of the random effects. Our approach was tested for selected locations of the eastern half of the United States: a region chosen to assess modeled and observed data effectively because catchments are relatively free from anthropogenic disturbances and basin sizes are comparable with those of the model grid-cells.

We revealed spatial patterns of change in future flood magnitudes over the eastern half of the U.S., showing a general decrease in the southeast. We found that with our data set the extreme value distribution's parameter that changes between historical and future periods is the location, while the scale can be left fixed.

Although an increase in flooding has been documented in parts of the Midwest and from the northern Appalachian Mountains to New England, overall there is no clear sign of change in the area of study over the last few decades (Archfield et al., 2016; Berghuijs et al., 2016; Hodgkins et al., 2017; Mallakpour & Villarini, 2015; Villarini & Smith, 2010). All the while, model projections indicate a reduction in flood magnitude toward the end of this century in the southeast of the United States. The signal remained the same even using fewer runs (~45%) deemed more credible, with the ensemble constrained using historical runoff, *cE* (as in e.g., Yang et al., 2017).

There is a clear pattern southwest-northeast in which GCMs dominate uncertainty, while in the northwest and the southeast GIMs are the predominant factor reflecting their increased challenge in reproducing runoff under more complex storage-release processes (like ice-cold conditions in the north and increased evaporation and aquifer dynamics in the south). The uncertainty depicted by our results indicates that the composition of multi-model ensembles should be tailored to the region of analysis, favor-

ing a rich set of GIMs while assessing floods in the south of the domain, and a rich set of GCMs in the central part of the domain. Constraining the ensemble produced similar partitions of uncertainty, with a few sites becoming GCM-dominated (from GIM-dominated in the full ensemble). Prioritizing better models does not necessarily reduce the uncertainty in the projections, but it does increase our confidence when results are based on models that simulate relevant aspects of the current climate more realistically (Knutti et al., 2017).

While global models are not expected to reach the same level of accuracy of for example catchment-calibrated models in reproducing flood characteristics, devising rules for selecting them helps to improve their credibility. Among the many possible rules, in this instance we opted to constrain the ensemble measuring the ability of models to reproduce the seasonality of flows. This choice was in part dictated by the fact that flow magnitude are mostly not well reproduced in the model outputs, therefore prioritizing models by this characteristic would yield an ensemble with too few members. In fact, we argue that global model evaluation against observed data is an essential step while carrying out continental to global scale studies. This is important because global models are increasingly challenged to provide information for planning and decision making, as reported by the EDgE Project (Samaniego et al., 2020), which has shown promise in the application of water-related climate services for decision making.

The difficulty of interpreting complex non-linear multi-model combinations in physical terms cannot be overemphasized. There are indeed multiple flood generating mechanisms in the domain of study and it is beyond the scope here to associate results in the occurrence of major floods at each site of the domain as seen with context-specific hydrological processes. Discerning which models simulate best which type of floods would require an in-depth study treating one model at a time and the validity of an assessment at a given catchment size may not apply to smaller or larger sizes (Wasko & Sharma, 2017).

Bayesian hierarchical models (like the one we apply herein) provide a valuable alternative to make use of numerous model runs in a robust and transparent way. Unlike previous studies, our methodology explicitly describes the overall signal of all runs, as opposed to the ensemble mean, thus minimizing loss of information and allowing at the same time a seamless partitioning of the uncertainty.

Work in the direction of making the best use of ensemble runs will benefit from exploiting newer runs from ensemble experiments and from assessing historical performance using additional observation data sets (i.e., ground measurements like streamflow data or satellite and reanalysis data). Improving projections of future flood risk will happen also through the improvement in the representation of plant processes like plant growth and stomatal conductance response to CO_2. Finally, a coveted step toward flood projections improvement—though a difficult step to implement everywhere due to lack of data—is the inclusion of water management and abstraction into global model simulations. An example of the importance of this aspect is the decrease over the last few decades in water retention capability (i.e., the fraction of precipitation lost by evapotranspiration decreased in favor of runoff) observed over eastern North America (among other regions of the world) that was not reflected in CMIP5 model runs, highlighting the importance of direct human intervention impacts, which strongly affects runoff estimates (Abbott et al., 2019; Yang et al., 2018). The inclusion in global models of human interventions on water resources like irrigation, new dam construction, and stream channeling is a necessary step to improve the simulation of current and future hydrological processes over a great portion of the planet and would certainly benefit the estimates of hydrological extremes.

Importantly, research efforts should go into finding ways to make the best use of the global model runs in order to produce the best possible estimates of future changes (Brunner et al., 2019), adopting statistical frameworks that retain effectively the information and the representativeness of all model runs employed.

Data Availability Statement

The ISI-MIP Fast-Track data set is available upon request following the instructions provided at the url www.isimip.org/gettingstarted/data-access/. The observed (streamflow gauges) data are openly available via the url: http://waterdata.usgs.gov/nwis/sw.

Acknowledgments
We thank the land-surface and hydrology modeling groups participating to the ISI-MIP Project, whose model output was used in this study. IG's contribution was funded by a postdoctoral research associateship at the University of Birmingham, UK. We thank the three anonimous reviewers and the associate editor for their helpful comments.

References

Abbott, B. W., Bishop, K., Zarnetske, J. P., Hannah, D. M., Frei, R. J., Minaudo, C., et al. (2019). A water cycle for the Anthropocene. *Hydrological Processes*, *33*(23), 3046–3052. https://doi.org/10.1002/hyp.13544

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., et al. (2019). ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth System Dynamics*, *10*(1), 91–105. https://doi.org/10.5194/esd-10-91-2019

Alfieri, L., Burek, P., Feyen, L., & Forzieri, G. (2015). Global warming increases the frequency of river floods in Europe. *Hydrology and Earth System Science*, *19*(5), 2247–2260. https://doi.org/10.5194/hess-19-2247-2015

Alfieri, L., Dottori, F., Betts, R., Salamon, P., & Feyen, L. (2018). Multi-model projections of river flood risk in Europe under global warming. *Climate*, *6*(1), 6. https://doi.org/10.3390/cli6010006

Ansari, A. R., & Bradley, R. A. (1960). Rank-sum tests for dispersions. *The Annals of Mathematical Statistics*, *31*(4), 1174–1189.

Archfield, S. A., Hirsch, R. M., Viglione, A., & Blöschl, G. (2016). Fragmented patterns of flood change across the United States. *Geophysical Research Letters*, *43*(19), 232–310. https://doi.org/10.1002/2016GL070590

Berghuijs, W. R., Woods, R. A., Hutton, C. J., & Sivapalan, M. (2016). Dominant flood generating mechanisms across the United States. *Geophysical Research Letters*, *43*, 1–9. https://doi.org/10.1002/2016GL068070

Bertola, M., Viglione, A., & Blöschl, G. (2019). Informed attribution of flood changes to decadal variation of atmospheric, catchment and river drivers in Upper Austria. *Journal of Hydrology*, *577*, 123919. https://doi.org/10.1016/j.jhydrol.2019.123919

Bosshard, T., Carambia, M., Goergen, K., Kotlarski, S., Krahe, P., Zappa, M., & Schär, C. (2013). Quantifying uncertainty sources in an ensemble of hydrological climate-impact projections. *Water Resources Research*, *49*, 1523–1536. https://doi.org/10.1029/2011WR011533

Brunner, L., Lorenz, R., Zumwald, M., & Knutti, R. (2019). Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters*, *14*(12), 124010. https://doi.org/10.1088/1748-9326/ab492f

Buchanan, B., Auerbach, D. A., Knighton, J., Evensen, D., Fuka, D. R., Easton, Z., et al. (2018). Estimating dominant runoff modes across the conterminous United States. *Hydrological Processes*, *32*(26), 1–10. https://doi.org/10.1002/hyp.13296

Bürkner, P.-C. (2017). An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Camici, S., Brocca, L., Melone, F., & Moramarco, T. (2014). Impact of climate change on flood frequency using different climate models and downscaling approaches. *Journal of Hydrologic Engineering*, *19*(8), 04014002. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000959

Castellarin, A., Kohnova, S., Gaal, L., Fleig, A., Salinas, J. L., Toumazis, A., et al. (2012). *Review of applied-statistical methods for flood-frequency analysis in Europe*–). Wallingford, UK: NERC/Centre for Ecology & Hydrology. ISBN: 9781906698324. https://nora.nerc.ac.uk/id/eprint/19286

Cisneros, J., Oki, T., Arnell, N. W., Benito, G., Cogley, J. G., Döll, P., et al. (2014). Freshwater resources. In C. B. Field, V. R. Barros, D. J. Dokken, K. J. Mach, & M. D. Mastrandrea (Eds.), *Climate change 2014 impacts, adaptation, and vulnerability* (pp. 229–269). Cambridge, UK: Cambridge University Press.

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., et al. (2015). Improving the representation of hydrologic processes in Earth System Models. *Water Resources Research*, *51*(8), 5929–5956. https://doi.org/10.1002/2015WR017096

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer.

Collins, M. J. (2019). River flood seasonality in the Northeast United States: Characterization and trends. *Hydrological Processes*, *33*(5), 687–698. https://doi.org/10.1002/hyp.13355

Crichton, D. (1999). The risk triangle. In J. Ingleton (Ed.), *Natural disaster management* (pp. 102–103). London, UK: Tudor Rose.

Dankers, R., Arnell, N. W., Clark, D. B., Falloon, P. D., Fekete, B. M., Gosling, S. N., et al. (2014). First look at changes in flood hazard in the Inter-Sectoral Impact model intercomparison project ensemble. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 3257–3261. https://doi.org/10.1073/pnas.1302078110

Davie, J. C. S., Falloon, P. D., Kahana, R., Dankers, R., Betts, R., Portmann, F. T., et al. (2013). Comparing projections of future changes in runoff from hydrological and biome models in ISI-MIP. *Earth System Dynamics*, *4*(2), 359–374. https://doi.org/10.5194/esd-4-359-2013

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: The role of internal variability. *Climate Dynamics*, *38*(3–4), 527–546. https://doi.org/10.1007/s00382-010-0977-x

Do, H. X., Zhao, F., Westra, S., Leonard, M., Gudmundsson, L., Boulange, J. E. S., et al. (2020). Historical and future changes in global flood magnitude—Evidence from a model--observation investigation. *Hydrology and Earth System Sciences*, *24*(3), 1543–1564. https://doi.org/10.5194/hess-24-1543-2020

Dottori, F., Szewczyk, W., Ciscar, J. C., Zhao, F., Alfieri, L., Hirabayashi, Y., et al. (2018). Increased human and economic losses from river flooding with anthropogenic warming. *Nature Climate Change*, *20*, 9039. https://doi.org/10.1038/s41558-018-0257-z

England, J. F., Jr, Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas, W. O., Jr, Veilleux, A. G., et al. (2018). Guidelines for determining flood flow frequency—Bulletin 17C (book 4, chap. B5, p. 148, ver. 1.1, May 2019). Reston, VA: U.S. Geological Survey Techniques and Methods. https://doi.org/10.3133/tm4B5

Eyring, V., Cox, P. M.. Flato, G. M.. Gleckler, P. J.. Abramowitz, G.. Caldwell, P.. et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

François, B., Schlef, K. E., Wi, S., & Brown, C. M. (2019). Design considerations for riverine floods in a changing climate–A review. *Journal of Hydrology*, *574*, 557–573. https://doi.org/10.1016/j.jhydrol.2019.04.068

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society A*, *182*(2), 389–402.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC. ISBN-13: 978-1439840955. http://www.stat.columbia.edu/~gelman/book/

Gelman, A., & Hill, J. (2006). Multilevel structures. In *Data analysis using regression and multilevel/hierarchical models* (pp. 237–250). Cambridge, UK: Cambridge University Press.

Giuntoli, I., Vidal, J.-P., Prudhomme, C., & Hannah, D. M. (2015a). Future hydrological extremes: The uncertainty from multiple global climate and global hydrological models. *Earth System Dynamics*, *6*(1), 267–285. https://doi.org/10.5194/esd-6-267-2015

Giuntoli, I., Villarini, G., Prudhomme, C., & Hannah, D. M. (2018). Uncertainties in projected runoff over the conterminous United States. *Climate Change*, *150*(3–4), 149–162. https://doi.org/10.1007/s10584-018-2280-5

Giuntoli, I., Villarini, G., Prudhomme, C., Mallakpour, I., & Hannah, D. M. (2015b). Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States. *Journal of Geophysical Research: Atmospheres*, *120*(18), 9138–9159. https://doi.org/10.1002/2015JD023401

Goodess, C. M. (2012). How is the frequency, location and severity of extreme events likely to change up to 2060 ? *Environmental Science & Policy*, *27*, S4–S14. https://doi.org/10.1016/j.envsci.2012.04.001

Gudmundsson, L., Wagener, T., Tallaksen, L. M., & Engeland, K. (2012). Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. *Water Resources Research*, *48*(11), W11504. https://doi.org/10.1029/2011WR010911

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., et al. (2013). Climate change impact on available water resources obtained using multiple global climate and hydrology models. *Earth System Dynamics*, *4*(1), 129–144. https://doi.org/10.5194/esd-4-129-2013

Hawkins, E., & Sutton, R. (2009). The Potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1107. https://doi.org/10.1175/2009BAMS2607.1

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., & Piontek, F. (2013). A trend-preserving bias correction—the ISI-MIP approach. *Earth System Dynamics*, *4*(2), 219–236. https://doi.org/10.5194/esd-4-219-2013

Hirabayashi, Y., Kanae, S., Emori, S., Oki, T., & Kimoto, M. (2008). Global projections of changing risks of floods and droughts in a changing climate. *Hydrological Sciences Journal*, *53*(4), 754–772. https://doi.org/10.1623/hysj.53.4.754

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nature Climate Change*, *3*(9), 816–821. https://doi.org/10.1038/nclimate1911

Hodgkins, G. A., Whitfield, P. H., Burn, D. H., Hannaford, J., Renard, B., Stahl, K., et al. (2017). Climate-driven variability in the occurrence of major floods across North America and Europe. *Journal of Hydrology*, *552*, 704–717. https://doi.org/10.1016/j.jhydrol.2017.07.027

Institute of Hydrology. (1999). *The flood estimation handbook, 5 volumes*. Wallingford, UK: Institute of Hydrology.

Katz, R. W., Craigmile, P. F., Guttorp, P., Haran, M., Sansó, B., & Stein, M. L. (2013). Uncertainty analysis in climate change assessments. *Nature Climate Change*, *3*(9), 769–771. https://doi.org/10.1038/nclimate1980

Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3–4), 395–404. https://doi.org/10.1007/s10584-010-9800-2

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*(4), 1909–1918. https://doi.org/10.1002/2016GL072012

Koirala, S., Yeh, P. J.-F., Hirabayashi, Y., Kanae, S., & Oki, T. (2014). Global-scale land surface hydrologic modeling with the representation of water table dynamics. *Journal of Geophysical Research: Atmospheres*, *119*(1), 75–89. https://doi.org/10.1002/2013JD020398

Kooperman, G. J., Fowler, M. D., Hoffman, F. M., Koven, C. D., Lindsay, K., Pritchard, M. S., et al. (2018). Plant physiological responses to rising CO_2 modify simulated daily runoff intensity with implications for global-scale flood risk assessment. *Geophysical Research Letters*, *45*(22), 457–512. https://doi.org/10.1029/2018GL079901

Lavell, A., Oppenheimer, M., Diop, C., Hess, J., Lempert, R., Li, J., et al. (2012). Climate change: New dimensions in disaster risk, exposure, vulnerability, and resilience. In M. Moser, & K. Takeuchi (Eds.), *Managing the risks of extreme events and disasters to advance climate change adaptation* (pp. 25–64). Cambridge University Press.

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., et al. (2020). Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth System Dynamics*, *11*(2), 491–508. https://doi.org/10.5194/esd-11-491-2020

Lim, W. H., Yamazaki, D., Koirala, S., Hirabayashi, Y., Kanae, S., Dadson, S. J., et al. (2018). Long-term changes in global socioeconomic benefits of flood defenses and residual risk based on CMIP5 climate models. *Earth's Future*, *6*(7), 938–954. https://doi.org/10.1002/2017EF000671

Mallakpour, I., & Villarini, G. (2015). The changing nature of flooding across the central United States. *Nature Climate Change*, *5*(3), 1–5. https://doi.org/10.1038/nclimate2516

Massey, F. J. (1952). Distribution table for the deviation between two sample cumulative. *The Annals of Mathematical Statistics*, *23*(3), 435–441.

McSweeney, C. F., & Jones, R. G. (2016). How representative is the spread of climate projections from the 5 CMIP5 GCMs used in ISI-MIP? *Climate Services*, *1*, 24–29. https://doi.org/10.1016/j.cliser.2016.02.001

Merz, B., Aerts, J. C. J. H., Arnbjerg-Nielsen, K., Baldi, M., Becker, A., Bichet, A., et al. (2014). Floods and climate: Emerging perspectives for flood risk assessment and management. *Natural Hazards and Earth System Sciences*, *14*(7), 1921–1942. https://doi.org/10.5194/nhess-14-1921-2014

Musselman, K. N., Lehner, F., Ikeda, K., Clark, M. P., Prein, A. F., Liu, C., et al. (2018). Projected increases and shifts in rain-on-snow flood risk over western North America. *Nature Climate Change*, *8*(9), 808–812. https://doi.org/10.1038/s41558-018-0236-4

Naz, B. S., Kao, S.-C., Ashfaq, M., Rastogi, D., Mei, R., & Bowling, L. C. (2016). Regional hydrologic response to climate change in the conterminous United States using high-resolution hydroclimate simulations. *Global and Planetary Change*, *143*, 100–117. https://doi.org/10.1016/j.gloplacha.2016.06.003

Northrop, P. J., & Chandler, R. E. (2014). Quantifying Sources of Uncertainty in Projections of Future Climate. *Journal of Climate*, *27*(23), 8793–8809. https://doi.org/10.1175/JCLI-D-14-00265.1

Oppenheimer, M., Campos, M., & Warren, R. (2014). Emergent risks and key vulnerabilities. In M. Brklacich, & S. Semenov (Eds.), *Climate change 2014: Impacts, adaptation, and vulnerability. Part A: Global and sectorial aspects. Contribution of working group II to the fifth assessment report of the IPCC* (pp. 1039–1099). Cambridge, UK and New York, NY: Cambridge University Press.

Padrón, R. S., Gudmundsson, L., & Seneviratne, S. I. (2019). Observational constraints reduce likelihood of extreme changes in multidecadal land water availability. *Geophysical Research Letters*, *46*(2), 736–744. https://doi.org/10.1029/2018GL080521

Papalexiou, S. M., & Koutsoyiannis, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, *49*(1), 187–201. https://doi.org/10.1029/2012WR012557

Pendergrass, A. G. (2018). What precipitation is extreme? *Science*, *84*(6393), 360. https://doi.org/10.1126/science.aat1871

Prosdocimi, I., Kjeldsen, T. R., & Miller, J. D. (2015). Detection and attribution of urbanization effect on flood extremes using nonstationary flood-frequency models. *Water Resources Research*, *51*(6), 4244–4262. https://doi.org/10.1002/2015WR017065

Ragno, E., AghaKouchak, A., Love, C. A., Cheng, L., Vahedifard, F., & Lima, C. H. R. (2018). Quantifying changes in future intensity-duration-frequency curves using multimodel ensemble simulations. *Water Resources Research*, *54*(3), 1751–1764. https://doi.org/10.1002/2017WR021975

Salas, J. D., Obeysekera, J., & Vogel, R. M. (2018). Techniques for assessing water infrastructure for nonstationary extreme events: A review. *Hydrological Sciences Journal*, *63*(3), 325–352. https://doi.org/10.1080/02626667.2018.1426858

Samaniego, L., Thober, S., Wanders, N., Pan, M., Rakovec, O., Sheffield, J., et al. (2020). Hydrological forecasts and projections for improved decision-making in the water sector in Europe. *Bulletin of the American Meteorological Society*, *100*(12), 2451–2472. https://doi.org/10.1175/BAMS-D-17-0274.1

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., et al. (2014). Multimodel assessment of water scarcity under climate change. *Proceedings of the National Academy of Sciences*, *111*(9), 3245–3250. https://doi.org/10.1073/pnas.1222460110

Schoof, J. T., & Robeson, S. M. (2016). Projecting changes in regional temperature and precipitation extremes in the United States. *Weather and Climate Extremes*, *11*, 28–40. https://doi.org/10.1016/j.wace.2015.09.004

Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., et al. (2012). Changes in climate extremes and their impacts on the natural physical environment. In M. Rusticucci, & V. Semenov (Eds.), *Managing the Risk Extreme Events Disasters to Advance Climate Change Adaptation A Spec. Rep. Work. Groups I II IPCC*, (pp. 109–230).

Stan Development Team. (2018). *Stan modeling language: User's guide and reference manual*. Version 2.17.1. Stan Development Team.

Trigg, M. A., Birch, C. E., Neal, J. C., Bates, P. D., Smith, A., Sampson, C. C., et al. (2016). The credibility challenge for global fluvial flood risk analysis. *Environmental Research Letters*, *11*(9), 094014. https://doi.org/10.1088/1748-9326/11/9/094014

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Villarini, G., & Smith, J. A. (2010). Flood peak distributions for the eastern United States. *Water Resources Research*, *46*(6), 1–17. https://doi.org/10.1029/2009WR008395

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The inter-sectoral impact model intercomparison project (ISI-MIP): Project framework. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(9), 3228–3232. https://doi.org/10.1073/pnas.1312330110

Wasko, C., & Sharma, A. (2017). Global assessment of flood and storm extremes with increased temperatures. *Scientific Reports*, *7*(1), 7945. https://doi.org/10.1038/s41598-017-08481-1

Whitfield, P. H., Burn, D. H., Hannaford, J., Higgins, H., Hodgkins, G. a., Marsh, T., & Looser, U. (2012). Reference hydrologic networks I. The status and potential future directions of national reference hydrologic networks for detecting trends. *Hydrological Sciences Journal*, *57*(8), 1562–1579. https://doi.org/10.1080/02626667.2012.728706

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometric Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Wobus, C., Gutmann, E., Jones, R., Rissing, M., Mizukami, N., Lorie, M., et al. (2017). Climate change impacts on flood risk and asset damages within mapped 100-year floodplains of the contiguous United States. *Nature Hazards Earth System Science*, *17*(12), 2199–2211. https://doi.org/10.5194/nhess-17-2199-2017

Yang, H., Piao, S., Huntingford, C., Ciais, P., Li, Y., Wang, T., et al. (2018). Changing the retention properties of catchments and their influence on runoff under climate change. *Environmental Research Letters*, *13*(9), 094019. https://doi.org/10.1088/1748-9326/aadd32

Yang, H., Zhou, F., Piao, S., Huang, M., Chen, A., Ciais, P., et al. (2017). Regional patterns of future runoff changes from Earth system models constrained by observation. *Geophysical Research Letters*, *44*(11), 5540–5549. https://doi.org/10.1002/2017GL073454

Yip, S., Ferro, C. a. T., Stephenson, D. B., & Hawkins, E. (2011). A simple, coherent framework for partitioning uncertainty in climate predictions. *Journal of Climate*, *24*(17), 4634–4643. https://doi.org/10.1175/2011JCLI4085.1

Zaghloul, M., Papalexiou, S. M., Elshorbagy, A., & Coulibaly, P. (2020). Revisiting flood peak distributions: A pan-Canadian investigation. *Advances in Water Resources*, *145*, 103720. https://doi.org/10.1016/j.advwatres.2020.103720