Going beyond the ensemble mean: assessment of future floods from global multi-models

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

[1]School of Geography, Earth and Environment Sciences, University of Birmingham, Birmingham, B15 2TT, United Kingdom.

[2]Dipartimento di Scienze Ambientali, Informatica e Statistica, Ca' Foscari University of Venice, Venezia Mestre, Italy.

Corresponding author: Ignazio Giuntoli, (ignazio.g@gmail.com)

†Current address: Institute of Atmospheric Sciences and Climate (ISAC) – CNR, Bologna, 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

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 (GCMs), global impact (GIMs) 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 towards 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 [*Lavell et al.*, 2012; *Abbott et al.*, 2019]. 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 [*Jiménez Cisneros et al.*, 2015]. 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 GCM runs feeding GIMs that simulate runoff at the land surface [e.g., *Hirabayashi et al.*, 2013; *Dankers et al.*, 2014; *Alfieri et al.*, 2015] (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 [*Merz et al.*, 2014; *Clark et al.*, 2015]. 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 and 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 global climate – GCMs and global impact – GIMs models. 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 characterising 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 – i.e. 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).

85 At the global scale, changes in mean flows from global models indicate an increase at high
86 latitudes and in the wet tropics, and a decrease in most dry tropical regions, although some
87 regions have high uncertainty in the magnitude and direction of change [e.g., *Hagemann et al.*,
88 2013; *Schewe et al.*, 2014]. Conversely, changes in flood magnitude are less consistent, with
89 contrasting results among studies depending on the region and the ensemble setup [*Hirabayashi*
90 *et al.*, 2013; *Dankers et al.*, 2014; *Giuntoli et al.*, 2015b]. The lack of consistency in these
91 changes is emphasized by *Jiménez Cisneros et al.*, [2015] reporting that studies of flood
92 projections under different emission scenarios are still few, and highly uncertain, given the
93 complexity of the mechanisms driving floods at the regional scale. In fact, studies using runoff
94 projections have started trying, in addition to assessing future floods characteristics, to untangle
95 the uncertainty originating from the different components of the modelling chain e.g. [*Koirala*
96 *et al.*, 2014; *Giuntoli et al.*, 2015b].

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

106 In light of these research gaps, the overarching aim of this study is to apply a novel Bayesian
107 model to the eastern USA to estimate space-time changes in future flood magnitude from multi-
108 model ensembles and so improve the overall signal/ pattern of change and identify sources of
109 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 oftentimes 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; *Giuntoli et al.*, 2015b; *Dottori et al.*, 2018] – 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

137 random effects. It should be noted that the terminology “GIMs” used herein could also be
138 referred to as “GHMs” i.e. global hydrological models.

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

156 We present the data in Section 2 with an appraisal of how peak flow modelled data
157 compares to observed data. In Section 3 we describe the statistical framework for estimating
158 future changes in flood magnitude and then how the ensemble is constrained. Results are
159 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).

Figure1

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 2000 to 2500 (2500 to 3000) km^2 north (south) of 36N latitude and with daily discharge data covering the models' control period (1971-2005). This choice follows the approach of *Giuntoli et al.*, [2015a] of carefully 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 modelled data, the streamflow gauges were selected from the Hydro-Climatic Data Network (HCDN), 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 (henceforth, SI).

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, i.e., ~50 km at the equator (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 SI. *Giuntoli et al.* [2018] provide detailed information on model characteristics and evaluation.

2.1. Appraisal of simulated vs observed peak flows

We compare observed and modelled peak magnitude (AMax) and timing (AMaxDate) at the 18 locations highlighting discrepancies between observed and modelled data. Observed-modelled 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 modelled data using non-parametric tests (no assumption is made on the type of distribution) assessing: i) same distribution (Kolmogorov-Smirnoff, noted KS, [*Massey*, 1952]), ii) equal median (Wilcoxon rank-sum, noted W, [*Wilcoxon*, 1945]), and iii) equal variance (Ansari-Bradley, noted AB, [*Ansari and Bradley*, 1960]). There is little overlap between observed and modelled 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 modelled data from historical to future period (RCP 8.5) yields greater agreement across the three tests (KS 66%, W 69%, AB 90%) than seen with the observed peak flows, as reported in Table S3 of the SI.

2.1.2. Peak flow magnitude

In addition to testing raw peak flows we compared observed and modelled 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 grey band shows the observed data, i.e. the reference to which the historical period of the models (black lines) should tend to align, while the red coloured lines correspond to the future period under scenario RCP8.5 (plots for all sites are in SI, Figure S1 and Figure 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 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 SI). Overall, MacPDM, PCRGlob-WB and VIC are the GIMs that capture timing of peak flows best, while, H08, LPJmL (north, especially), and MPI (south, especially) struggle to replicate the right timing of peak occurrences. Furthermore, models generally anticipate peak occurrence (in Figure S4 of the SI coloured 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 modelled peaks occur in absence of corresponding observed peaks.

Figure 2

This modelled-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. Firstly (Section 3.1.1), we present the Bayesian hierarchical model used to analyse the flood peaks, and secondly (Section 3.1.2), we provide further detail on Bayesian inference and hierarchical models.

3.1.1. Modelling 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: e.g., 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 [*Castellarin et al.*, 2012; *Bertola et al.*, 2019]. 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 and 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:

$$\frac{1}{\theta} exp\{-\frac{x-\xi}{\theta} - exp\{-\frac{x-\xi}{\theta}\}\} \quad [1]$$

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. *Dankers et al.* [2014]; *Alfieri et al.* [2018]), a hierarchical approach in which data from all runs are modelled together is employed. 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 the understanding of future high flows. Moreover, by modelling 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 9 GIMs and 5 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 not very large. The accurate estimation of models, which allow for more structure in the scale parameters, would require very large samples and very sizeable 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.

Figure 3

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:

$$\xi_{i,j,k,h} = \alpha + \alpha_{gcm,i} + \alpha_{gim,j} + \alpha_{comb,k} +$$

303 $$+ \beta * \boldsymbol{I}_{[36,70]}(h) + \beta_{gcm,i} * \boldsymbol{I}_{[36,70]}(h) + \beta_{gim,j} * \boldsymbol{I}_{[36,70]}(h) + \beta_{comb,k} * \boldsymbol{I}_{[36,70]}(h) \quad [2.a]$$

304 $$\theta_{i,j} = \exp\{\gamma + \gamma_{gcm,i} + \gamma_{gim,j}\} \quad [2.b]$$

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

309 The parameter α in equation [2.a] represents the overall population-level value for the
310 intercept parameter of the location across all model combinations. To accommodate the
311 variability across the different models three group-specific terms have been included: $\alpha_{gcm,i}$ to
312 allow for the variability across the GCMs; $\alpha_{gim,i}$ to allow for the variability across the GIMs;
313 and $\alpha_{comb,k}$ to allow for the variability across each GCM and GIM combination. By comparing
314 the different values of $\sigma^2_{\alpha,gcm}$, $\sigma^2_{\alpha,gim}$, and $\sigma^2_{\alpha,comb}$, it is possible to assess which grouping variable
315 explains the largest proportion of variability (i.e. uncertainty) in the runoff values. Notice that
316 the factor describing the combination of GCM and GIM is only included for the location
317 parameter model. The inclusion of this factor has been found to improve the fit of the model
318 prediction to the data, and was deemed useful to describe the interaction between different
319 GIMs (applied to different areas of the continent and which might require different input
320 variables) and the GCMs, which reproduce the different climate components in a very different
321 fashion. The interaction between the two factors can be already guessed in Figure 2a, in which
322 clusters of estimated design events are not fully explained by the GIM or the GCM under which
323 the data was generated, but exhibit some further variability.

324 The parameter β represents the overall population-level change in location parameter
325 when moving from the historic period time window to the future time window. The parameter
326 quantifies the overall average difference between the location parameter in the two time periods
327 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 $\alpha_{gcm,i}$, $\alpha_{gim,j}$, $\alpha_{comb,k}$, $\beta_{gim,i}$, $\beta_{gcm,j}$, $\beta_{comb,k}$, $\gamma_{gcm,i}$,$\gamma_{gim,j}$) 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 and 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 e.g., *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:

$$p(\eta|y) \propto p(y|\eta)*p(\eta) \qquad [3]$$

where $p(y|\eta)$ is the model for the distribution of the data conditional on the parameter η (i.e.

Figure 4

the Gumbel distribution with a model structure specified in equations [2.a] and [2.b]) and $p(\eta)$ is the prior distribution of η which 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: $\phi = (\sigma_\alpha , \sigma_\beta , \sigma_\gamma)$. By applying again Bayes' rule we have that:

$$p(\eta,\phi|y) \propto p(y|\eta,\phi) * p(\eta|\phi) * p(\phi) \quad [4]$$

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*, 2017], a state of the art probabilistic programming language for statistical modelling, 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 SI – 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, i.e. 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 modelled climate variables [*Northrop and 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 modelling 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 modelled values. Indeed, in Figure 2a, it would be expected that model data in the historical period (in black) overlaps the confidence interval (grey band) of the observed data, whereas in the majority of cases this hardly occurs (see Figure S1 and Figure S2 in the SI). 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 [*Gudmundsson et al.*, 2012; *Giuntoli et al.*, 2015a]. 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 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 modelled peak counts: i) the distance between the proportion of seasonal counts of observed and modelled peaks ii) 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, 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, 5 in JJA, 1 in SON.

2. Same as step 1 for simulated peak timings. For example, the site in Figure 5, for the JUL GIM fed by the HAD GCM peak counts are: 7 in DJF, 7 in MAM, 12 in JJA, 9 in SON. Note that the comparison is done on the GIM-GCM combination output.
3. Counts in step 1 (observed) and step 2 (modelled) 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 90^{th} 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, i.e. 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 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 the RMSE threshold in MAM and SON. At the same time, the MIROC GCM, is not excluded for distance counts but because it has a RMSE above threshold in MAM. Plots for all sites are shown in SI Figures S5 (northern sites) S6 (southern sites) S7 (two sites excluded), with the *cE* composition summarised in Table S4.

Figure 5

4. Results

The at-site change 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, i.e. the difference between the future (2065-2099) and the historical (1971-2005) periods. Secondly, Figure 7 illustrates the corresponding uncertainty contribution coming from GIMs (green), GCMs (yellow), and their interaction (grey), shown as boxplots 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 Figure 6 and Figure 7). Finally, we discuss results using a constrained ensemble (*cE*) obtained by reducing the full ensemble (*oE*) having compared modelled 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 across even small geographic areas [*Schoof and 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 36N) show a negative change, with one that is significant (95% credibility 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.

Figure 6

Wider pdf 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.*, 2015a].

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

Figure 7

Table 1

uncertainty is coloured depending on the distance from the other sources, that is bright (pale) coloured 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 grey), 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, unravelling 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 and infiltration excess – 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 infiltration excess runoff; on the other hand saturation excess 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 infiltration (IE) or saturation (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 SI). Over the eastern half of the United States, this may represent a limitation provided that a considerable share of the area is IE dominated, and 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, i.e. 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 is 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.*, 2015a]. 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 CO2 can reduce stomatal conductance and transpiration, which may lead to increased soil moisture and runoff in some regions, favouring 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.

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

579 # 5. Discussion and wider implications

580 The inherent tendency to disagree on the absolute value or on the sign of projected changes
581 of climate variables like precipitation and runoff in global model runs adds to the fact that
582 generally these runs do not match observations well [*Do et al.*, 2020]. Therefore, estimates of
583 future precipitation and runoff changes suffer from large uncertainty and from a signal that may
584 be cancelled out as different model simulations are averaged to generate a final value that is
585 often taken as the ensemble mean (e.g. [*Dankers et al.*, 2014; *Wobus et al.*, 2017; *Ragno et al.*,
586 2018]).

587 The aim of this paper was to propose a novel framework that allows for estimating the
588 changes in future flood magnitude with the signal of the direction of change expressed as the
589 distribution of the overall change rather than the ensemble mean. We quantified these changes
590 modelling the extreme values parameters using all multi-model ensemble simulations (GCM-
591 GIM) at once, and characterizing the uncertainty from both GCMs and GIMs as the variations
592 of the random effects. Our approach was tested for selected locations of the eastern half of the
593 United States of America: a region chosen to assess modelled 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 USA, 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 [*Villarini and Smith*, 2010; *Mallakpour and Villarini*, 2015; *Archfield et al.*, 2016; *Berghuijs et al.*, 2016; *Hodgkins et al.*, 2017]. All the while, model projections indicate a reduction in flood magnitude towards 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, favouring 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

618 when results are based on models that simulate relevant aspects of the current climate more
619 realistically [*Knutti et al.*, 2017].

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

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

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

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

659 Importantly, research efforts should go into finding ways to make the best use of the global
660 model runs in order to produce the best possible estimates of future changes [*Brunner et al.*,
661 2019], adopting statistical frameworks that retain effectively the information and the
662 representativeness of all model runs employed.

663 6. Acknowledgements

664 We thank the land-surface and hydrology modelling groups participating to the ISI-MIP
665 Project, whose model output was used in this study. The ISI-MIP Fast-Track dataset is available
666 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. IG's contribution was funded by a postdoctoral research associateship at the University of Birmingham, UK.

References

Abbott, B. W. et al. (2019), A water cycle for the Anthropocene, *Hydrol. Process.*, *33*(23), 3046–3052, doi:10.1002/hyp.13544.

Abramowitz, G., N. Herger, E. Gutmann, D. Hammerling, R. Knutti, M. Leduc, R. Lorenz, R. Pincus, and G. A. Schmidt (2019), ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, *Earth Syst. Dyn.*, *10*(1), 91–105, doi:10.5194/esd-10-91-2019.

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

Alfieri, L., F. Dottori, R. Betts, P. Salamon, and L. Feyen (2018), Multi-Model Projections of River Flood Risk in Europe under Global Warming, *Climate*, *6*(1), 6, doi:10.3390/cli6010006.

Ansari, A. R., and R. A. Bradley (1960), Rank-Sum Tests for Dispersions, *Ann. Math. Stat.*, *31*(4), 1174–1189.

Archfield, S. A., R. M. Hirsch, A. Viglione, and G. Blöschl (2016), Fragmented patterns of flood change across the United States, *Geophys. Res. Lett.*, *43*(19), 10,232-10,239, doi:10.1002/2016GL070590.

Berghuijs, W. R., R. A. Woods, C. J. Hutton, and M. Sivapalan (2016), Dominant flood generating mechanisms across the United States, *Geophys. Res. Lett.*, 1–9, doi:10.1002/2016GL068070.

691 Bertola, M., A. Viglione, and G. Blöschl (2019), Informed attribution of flood changes to
692 decadal variation of atmospheric, catchment and river drivers in Upper Austria, *J. Hydrol.*,
693 *577*, 123919, doi:10.1016/j.jhydrol.2019.123919.

694 Bosshard, T., M. Carambia, K. Goergen, S. Kotlarski, P. Krahe, M. Zappa, and C. Schär (2013),
695 Quantifying uncertainty sources in an ensemble of hydrological climate-impact
696 projections, *Water Resour. Res.*, *49*, n/a-n/a, doi:10.1029/2011WR011533.

697 Brunner, L., R. Lorenz, M. Zumwald, and R. Knutti (2019), Quantifying uncertainty in
698 European climate projections using combined performance-independence weighting,
699 *Environ. Res. Lett.*, *14*(12), 124010, doi:10.1088/1748-9326/ab492f.

700 Buchanan, B., D. A. Auerbach, J. Knighton, D. Evensen, D. R. Fuka, Z. Easton, M. Wieczorek,
701 J. A. Archibald, B. McWilliams, and T. Walter (2018), Estimating dominant runoff modes
702 across the conterminous United States, *Hydrol. Process.*, (September), 1–10,
703 doi:10.1002/hyp.13296.

704 Bürkner, P.-C. (2017), brms : An R Package for Bayesian Multilevel Models Using Stan, *J.*
705 *Stat. Softw.*, *80*(1), doi:10.18637/jss.v080.i01.

706 Camici, S., L. Brocca, F. Melone, and T. Moramarco (2014), Impact of Climate Change on
707 Flood Frequency Using Different Climate Models and Downscaling Approaches, *J.*
708 *Hydrol. Eng.*, *19*(8), 04014002, doi:10.1061/(ASCE)HE.1943-5584.0000959.

709 Castellarin, A., S. Kohnova, L. Gaal, A. Fleig, J. L. Salinas, A. Toumazis, T. R. Kjeldsen, and
710 N. Macdonald (2012), *Review of applied-statistical methods for flood-frequency analysis*
711 *in Europe*, NERC/Centre for Ecology & Hydrology, Wallingford.

712 Clark, M. P. et al. (2015), Improving the representation of hydrologic processes in Earth System
713 Models, *Water Resour. Res.*, *51*(8), 5929–5956, doi:10.1002/2015WR017096.

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

715 Collins, M. J. (2019), River flood seasonality in the Northeast United States: Characterization
716 and trends, *Hydrol. Process.*, *33*(5), 687–698, doi:10.1002/hyp.13355.
717 Crichton, D. (1999), The risk triangle, in *Natural Disaster Management*, edited by J. Ingleton,
718 pp. 102–103, Tudor Rose, London.
719 Dankers, R. et al. (2014), First look at changes in flood hazard in the Inter-Sectoral Impact
720 Model Intercomparison Project ensemble., *Proc. Natl. Acad. Sci. U. S. A.*, *111*, 3257–3261,
721 doi:10.1073/pnas.1302078110.
722 Davie, J. C. S. et al. (2013), Comparing projections of future changes in runoff from
723 hydrological and biome models in ISI-MIP, *Earth Syst. Dyn.*, *4*(2), 359–374,
724 doi:10.5194/esd-4-359-2013.
725 Deser, C., A. Phillips, V. Bourdette, and H. Teng (2012), Uncertainty in climate change
726 projections: the role of internal variability, *Clim. Dyn.*, *38*(3–4), 527–546,
727 doi:10.1007/s00382-010-0977-x.
728 Do, H. X. et al. (2020), Historical and future changes in global flood magnitude -- evidence
729 from a model--observation investigation, *Hydrol. Earth Syst. Sci.*, *24*(3), 1543–1564,
730 doi:10.5194/hess-24-1543-2020.
731 Dottori, F. et al. (2018), Increased human and economic losses from river flooding with
732 anthropogenic warming, *Nat. Clim. Chang.*, *20*, 9039, doi:10.1038/s41558-018-0257-z.
733 England Jr., J. F., T. A. Cohn, B. A. Faber, J. R. Stedinger, W. O. Thomas Jr., A. G. Veilleux,
734 J. E. Kiang, and R. R. Mason Jr. (2018), Guidelines for determining flood flow
735 frequency—Bulletin 17C, in *4*, p. 168, USGS, Reston, VA.
736 Eyring, V. et al. (2019), Taking climate model evaluation to the next level, *Nat. Clim. Chang.*,
737 *9*(February), doi:10.1038/s41558-018-0355-y.
738 François, B., K. E. Schlef, S. Wi, and C. M. Brown (2019), Design considerations for riverine

floods in a changing climate – A review, *J. Hydrol.*, *574*, 557–573, doi:10.1016/j.jhydrol.2019.04.068.

Gabry, J., D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman (2019), Visualization in Bayesian workflow, *J. R. Stat. Soc. A*, *182*(Part 2), 389–402.

Gelman, A., and J. Hill (2006), Multilevel structures, in *Data Analysis Using Regression and Multilevel/Hierarchical Models*, pp. 237–250, Cambridge University Press, Cambridge.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013), *Bayesian Data Analysis Third Edition*, Chapman and Hall/CRC.

Giuntoli, I., G. Villarini, C. Prudhomme, I. Mallakpour, and D. M. Hannah (2015a), Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States, *J. Geophys. Res. Atmos.*, *120*(18), 9138–9159, doi:10.1002/2015JD023401.

Giuntoli, I., J.-P. Vidal, C. Prudhomme, and D. M. Hannah (2015b), Future hydrological extremes: the uncertainty from multiple global climate and global hydrological models, *Earth Syst. Dyn.*, *6*(1), 267–285, doi:10.5194/esd-6-267-2015.

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

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

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

Hagemann, S. et al. (2013), Climate change impact on available water resources obtained using

763 multiple global climate and hydrology models, *Earth Syst. Dyn.*, *4*(1), 129–144,
764 doi:10.5194/esd-4-129-2013.

765 Hawkins, E., and R. Sutton (2009), The Potential to Narrow Uncertainty in Regional Climate
766 Predictions, *Bull. Am. Meteorol. Soc.*, *90*(8), 1095–1107, doi:10.1175/2009BAMS2607.1.

767 Hempel, S., K. Frieler, L. Warszawski, J. Schewe, and F. Piontek (2013), A trend-preserving
768 bias correction – the ISI-MIP approach, *Earth Syst. Dyn.*, *4*(2), 219–236, doi:10.5194/esd-
769 4-219-2013.

770 Hirabayashi, Y., S. Kanae, S. Emori, T. Oki, and M. Kimoto (2008), Global projections of
771 changing risks of floods and droughts in a changing climate, *Hydrol. Sci. J.*, *53*(4), 754–
772 772, doi:10.1623/hysj.53.4.754.

773 Hirabayashi, Y., R. Mahendran, S. Koirala, L. Konoshima, D. Yamazaki, S. Watanabe, H. Kim,
774 and S. Kanae (2013), Global flood risk under climate change, *Nat. Clim. Chang.*, *3*(9),
775 816–821, doi:10.1038/nclimate1911.

776 Hodgkins, G. A. et al. (2017), Climate-driven variability in the occurrence of major floods
777 across North America and Europe, *J. Hydrol.*, doi:10.1016/j.jhydrol.2017.07.027.

778 Institute of Hydrology (1999), *The Flood Estimation Handbook, 5 Volumes*, Centre for Ecology
779 and Hydrology, Wallingford.

780 Jiménez Cisneros, B. E., T. Oki, N. W. Arnell, G. Benito, J. G. Cogley, P. Döll, T. Jiang, S. S.
781 Mwakalila, Z. Kundzewicz, and A. Nishijima (2015), Freshwater Resources, in *Climate*
782 *Change 2014 Impacts, Adaptation, and Vulnerability*, edited by C. B. Field, V. R. Barros,
783 D. J. Dokken, K. J. Mach, and M. D. Mastrandrea, pp. 229–270, Cambridge University
784 Press, Cambridge.

785 Katz, R. W., P. F. Craigmile, P. Guttorp, M. Haran, B. Sansó, and M. L. Stein (2013),
786 Uncertainty analysis in climate change assessments, *Nat. Clim. Chang.*, *3*(9), 769–771,
787 doi:10.1038/nclimate1980.

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

790 Knutti, R., J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring (2017), A
791 climate model projection weighting scheme accounting for performance and
792 interdependence, *Geophys. Res. Lett.*, *44*(4), 1909–1918, doi:10.1002/2016GL072012.

793 Koirala, S., P. J.-F. Yeh, Y. Hirabayashi, S. Kanae, and T. Oki (2014), Global-scale land surface
794 hydrologic modeling with the representation of water table dynamics, *J. Geophys. Res.*
795 *Atmos.*, *119*(1), 75–89, doi:10.1002/2013JD020398.

796 Kooperman, G. J., M. D. Fowler, F. M. Hoffman, C. D. Koven, K. Lindsay, M. S. Pritchard, A.
797 L. S. Swann, and J. T. Randerson (2018), Plant Physiological Responses to Rising CO 2
798 Modify Simulated Daily Runoff Intensity With Implications for Global-Scale Flood Risk
799 Assessment, *Geophys. Res. Lett.*, *45*(22), 12,457-12,466, doi:10.1029/2018GL079901.

800 Lavell, A., M. Oppenheimer, C. Diop, J. Hess, R. Lempert, J. Li, R. Muir-Wood, and S. Myeong
801 (2012), Climate change: new dimensions in disaster risk, exposure, vulnerability, and
802 resilience, in *Managing the Risks of Extreme Events and Disasters to Advance Climate*
803 *Change Adaptation*, pp. 25–64.

804 Lehner, F., C. Deser, N. Maher, J. Marotzke, E. M. Fischer, L. Brunner, R. Knutti, and E.
805 Hawkins (2020), Partitioning climate projection uncertainty with multiple large ensembles
806 and CMIP5/6, *Earth Syst. Dyn.*, *11*(2), 491–508, doi:10.5194/esd-11-491-2020.

807 Lim, W. H., D. Yamazaki, S. Koirala, Y. Hirabayashi, S. Kanae, S. J. Dadson, J. W. Hall, and
808 F. Sun (2018), Long-Term Changes in Global Socioeconomic Benefits of Flood Defenses
809 and Residual Risk Based on CMIP5 Climate Models, *Earth's Futur.*, *6*(7), 938–954,
810 doi:10.1002/2017EF000671.

811 Mallakpour, I., and G. Villarini (2015), The changing nature of flooding across the central
812 United States, *Nat. Clim. Chang.*, (February), 1–5, doi:10.1038/nclimate2516.

Massey, F. J. (1952), Distribution Table for the Deviation Between two Sample Cumulatives, *Ann. Math. Stat.*, *23*(3), 435–441.

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

Merz, B. et al. (2014), Floods and climate: emerging perspectives for flood risk assessment and management, *Nat. Hazards Earth Syst. Sci.*, *14*(7), 1921–1942, doi:10.5194/nhess-14-1921-2014.

Musselman, K. N., F. Lehner, K. Ikeda, M. P. Clark, A. F. Prein, C. Liu, M. Barlage, and R. Rasmussen (2018), Projected increases and shifts in rain-on-snow flood risk over western North America, *Nat. Clim. Chang.*, *8*(9), 808–812, doi:10.1038/s41558-018-0236-4.

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

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

Oppenheimer, M., M. Campos, and R. Warren (2014), Emergent risks and key vulnerabilities. In: 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, *Cambridge Univ. Press. Cambridge, UK New York, USA*, 1039–1099.

Padrón, R. S., L. Gudmundsson, and S. I. Seneviratne (2019), Observational Constraints Reduce Likelihood of Extreme Changes in Multidecadal Land Water Availability, *Geophys. Res. Lett.*, *46*(2), 736–744, doi:10.1029/2018GL080521.

Papalexiou, S. M., and D. Koutsoyiannis (2013), Battle of extreme value distributions: A global

838 survey on extreme daily rainfall, *Water Resour. Res.*, *49*(1), 187–201,
839 doi:10.1029/2012WR012557.
840 Pendergrass, A. G. (2018), What precipitation is extreme?, *Science (80-.).*, *360*(6393), 1072
841 LP – 1073, doi:10.1126/science.aat1871.
842 Prosdocimi, I., T. R. Kjeldsen, and J. D. Miller (2015), Detection and attribution of urbanization
843 effect on flood extremes using nonstationary flood-frequency models, *Water Resour. Res.*,
844 *51*(6), 4244–4262, doi:10.1002/2015WR017065.
845 Ragno, E., A. AghaKouchak, C. A. Love, L. Cheng, F. Vahedifard, and C. H. R. Lima (2018),
846 Quantifying Changes in Future Intensity-Duration-Frequency Curves Using Multimodel
847 Ensemble Simulations, *Water Resour. Res.*, *54*(3), 1751–1764,
848 doi:10.1002/2017WR021975.
849 Salas, J. D., J. Obeysekera, and R. M. Vogel (2018), Techniques for assessing water
850 infrastructure for nonstationary extreme events: a review, *Hydrol. Sci. J.*, *63*(3), 325–352,
851 doi:10.1080/02626667.2018.1426858.
852 Samaniego, L. et al. (2020), Hydrological Forecasts and Projections for Improved Decision-
853 Making in the Water Sector in Europe, *Bull. Am. Meteorol. Soc.*, *100*(12), 2451–2472,
854 doi:10.1175/BAMS-D-17-0274.1.
855 Schewe, J. et al. (2014), Multimodel assessment of water scarcity under climate change, *Proc.*
856 *Natl. Acad. Sci.*, *111*(9), 3245–3250, doi:10.1073/pnas.1222460110.
857 Schoof, J. T., and S. M. Robeson (2016), Projecting changes in regional temperature and
858 precipitation extremes in the United States, *Weather Clim. Extrem.*, *11*, 28–40,
859 doi:10.1016/j.wace.2015.09.004.
860 Seneviratne, S. et al. (2012), Changes in climate extremes and their impacts on the natural
861 physical environment, *Manag. Risk Extrem. Events Disasters to Adv. Clim. Chang. Adapt.*
862 *A Spec. Rep. Work. Groups I II IPCC*, 109–230.

Stan Development Team (2017), *Stan Modeling Language: User's Guide and Reference Manual. Version 2.17.1*.

Trigg, M. A. et al. (2016), The credibility challenge for global fluvial flood risk analysis, *Environ. Res. Lett.*, *11*(9), 094014, doi:10.1088/1748-9326/11/9/094014.

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

Villarini, G., and J. A. Smith (2010), Flood peak distributions for the eastern United States, *Water Resour. Res.*, *46*(6), 1–17, doi:10.1029/2009WR008395.

Warszawski, L., K. Frieler, V. Huber, F. Piontek, O. Serdeczny, and J. Schewe (2014), The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): project framework., *Proc. Natl. Acad. Sci. U. S. A.*, *111*(9), 3228–32, doi:10.1073/pnas.1312330110.

Wasko, C., and R. Nathan (2019), Influence of changes in rainfall and soil moisture on trends in flooding, *J. Hydrol.*, *575*, 432–441, doi:https://doi.org/10.1016/j.jhydrol.2019.05.054.

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

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

Wilcoxon, F. (1945), Individual Comparisons by Ranking Methods, *Biometrics Bull.*, *1*(6), 80–83, doi:10.2307/3001968.

Wobus, C., E. Gutmann, R. Jones, M. Rissing, N. Mizukami, M. Lorie, H. Mahoney, A. W. Wood, D. Mills, and J. Martinich (2017), Climate change impacts on flood risk and asset

887 damages within mapped 100-year floodplains of the contiguous United States, *Nat.*
888 *Hazards Earth Syst. Sci.*, *17*(12), 2199–2211, doi:10.5194/nhess-17-2199-2017.
889 Yang, H., F. Zhou, S. Piao, M. Huang, A. Chen, P. Ciais, Y. Li, X. Lian, S. Peng, and Z. Zeng
890 (2017), Regional patterns of future runoff changes from Earth system models constrained
891 by observation, *Geophys. Res. Lett.*, *44*(11), 5540–5549, doi:10.1002/2017GL073454.
892 Yang, H., S. Piao, C. Huntingford, P. Ciais, Y. Li, T. Wang, S. Peng, Y. Yang, D. Yang, and J.
893 Chang (2018), Changing the retention properties of catchments and their influence on
894 runoff under climate change, *Environ. Res. Lett.*, *13*(9), 094019, doi:10.1088/1748-
895 9326/aadd32.
896 Yip, S., C. a. T. Ferro, D. B. Stephenson, and E. Hawkins (2011), A Simple, Coherent
897 Framework for Partitioning Uncertainty in Climate Predictions, *J. Clim.*, *24*(17), 4634–
898 4643, doi:10.1175/2011JCLI4085.1.
899 Zaghloul, M., S. M. Papalexiou, A. Elshorbagy, and P. Coulibaly (2020), Revisiting flood peak
900 distributions: A pan-Canadian investigation, *Adv. Water Resour.*, *145*, 103720,
901 doi:https://doi.org/10.1016/j.advwatres.2020.103720.
902

List of captions

Tables

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

Figures

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.

Figure 2 – Comparison of observed-modelled magnitude (a) and timing (b) of annual maxima: a) confidence intervals (95%) of observed data (grey band) and 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 modelled (grey) 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%.

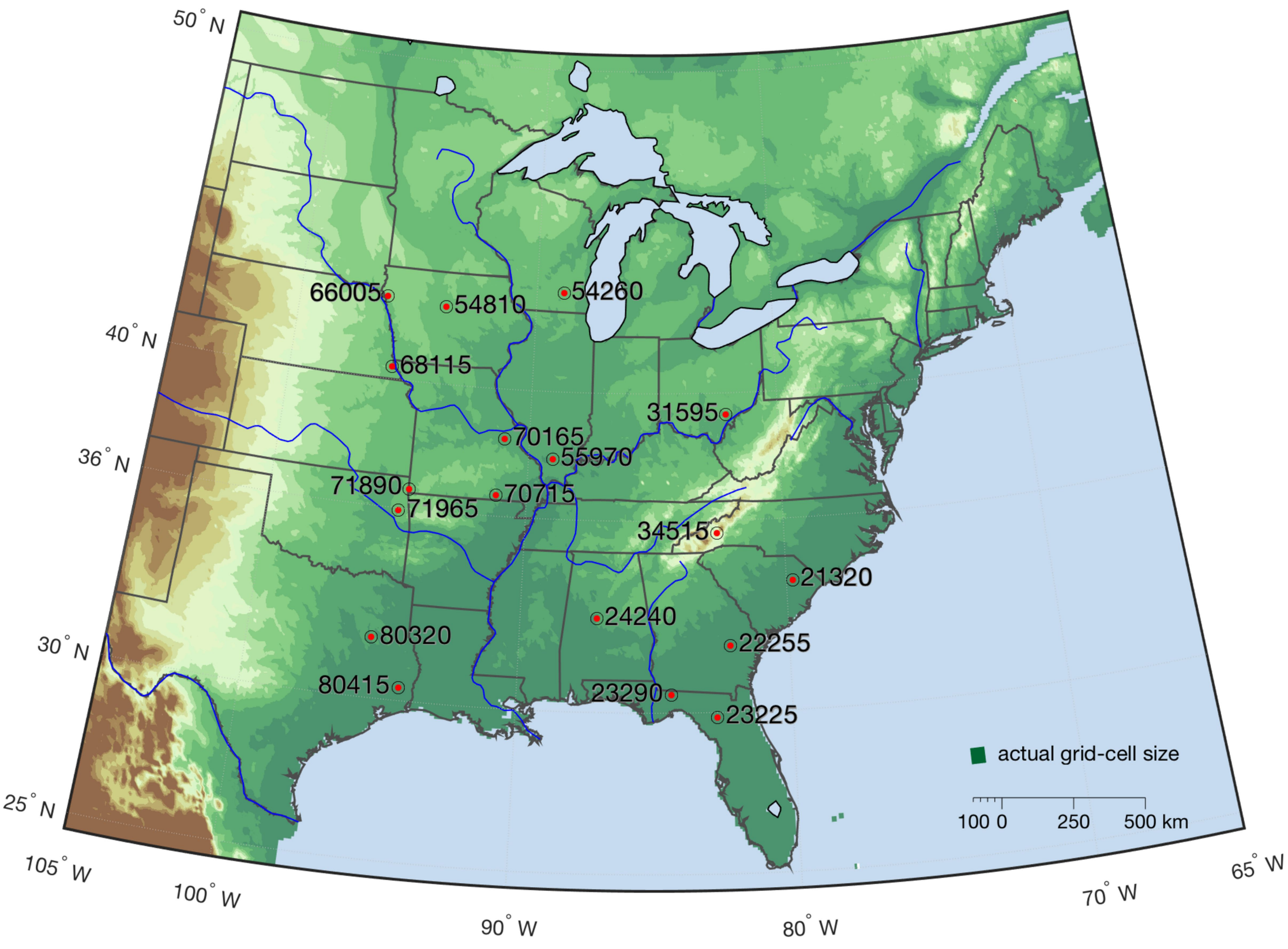
Figure 3 – Flowchart of the statistical analysis framework.

Figure 4 – Structure of the Bayesian hierarchical models.

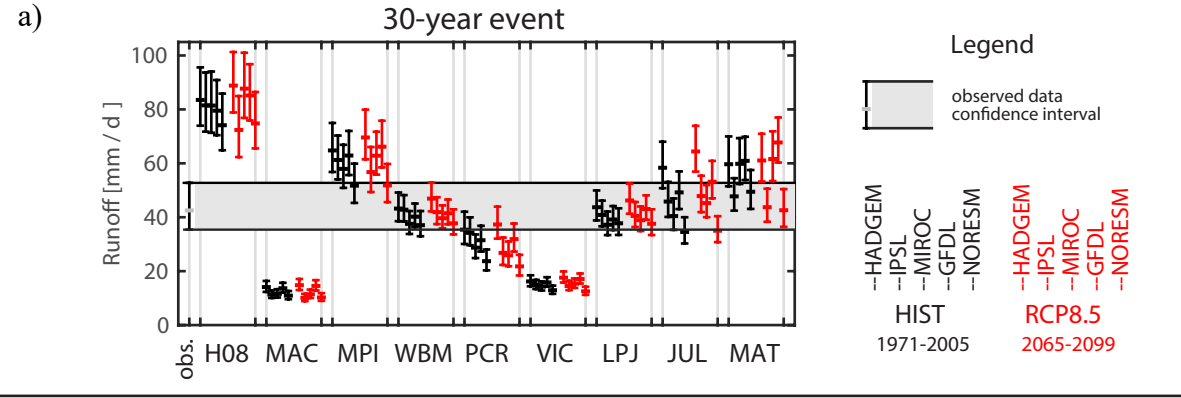
Figure 5 –Departure (%) from average observed peak flow (AMax) occurrences per season. Individual GIMs (GCMs) are expressed in row (column). Red (blue) tones indicate under (over) – estimation ("U/O") of peak counts ≥ (≤) 10%. Model exclusions (GIMs lower left, GCMs lower right) are denoted with X.

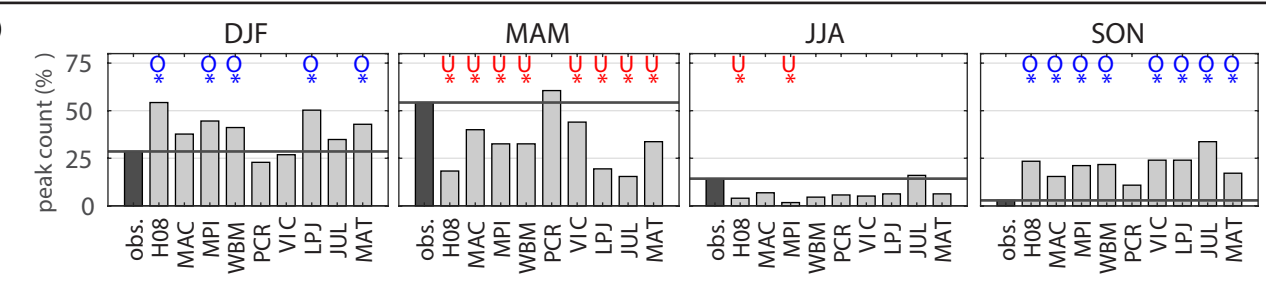
Figure 6 – Posterior distribution of β (the parameter that describes the change in the location parameter in the future) of the full ensemble, oE. Shaded blue (red) depicts positive (negative) values; solid vertical line corresponds to 0, dashed lines correspond to the 95% credible intervals. The fluorescent 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.

Figure 7 – Standard deviation of the random effects expressing main contributions to uncertainty in the changes due to GCM (yellow), GIM (green), GCM-GIM (gray) for the ß (time-window effect) of the location parameter. Lower three boxplots refer to the *oE*, while the upper three boxplots 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.

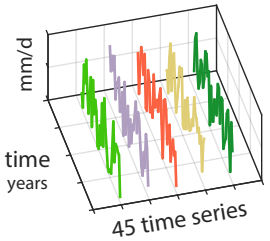

50° N
40° N
36° N
30° N
25° N
105° W
100° W
90° W
80° W
70° W
65° W
66005
54810
54260
68115
31595
70165
55970
71890
71965
70715
34515
21320
24240
80320
22255
80415
23290
23225
actual grid-cell size
100 0 250 500 km

st-70165

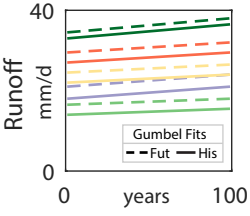

a)
30-year event
Runoff [mm / d]
100
80
60
40
20
0
obs.
H08
MAC
MPI
WBM
PCR
VIC
LPJ
JUL
MAT
Legend
observed data
confidence interval
--HADGEM
--IPSL
--MIROC
--GFDL
--NORESM
HIST
1971-2005
--HADGEM
--IPSL
--MIROC
--GFDL
--NORESM
RCP8.5
2065-2099


b)
DJF
MAM
JJA
SON
peak count (%)
75
50
25
0
obs.
H08
MAC
MPI
WBM
PCR
VIC
LPJ
JUL
MAT

Annual maxima time series



Bayesian hyerarchical model

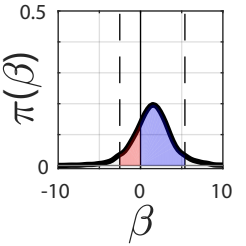


Gumbel model fitted using all GIM-GCM time series

Parameter estimation
Location ξ_H, ξ_F
Scale θ

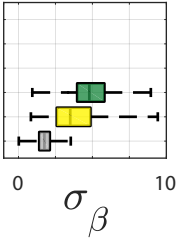
STAN

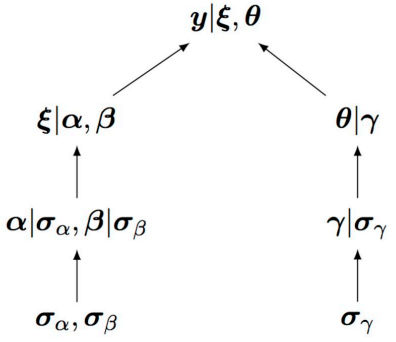
Change in location $\xi_F - \xi_H$



sign of change

Uncertainty
std of random effects





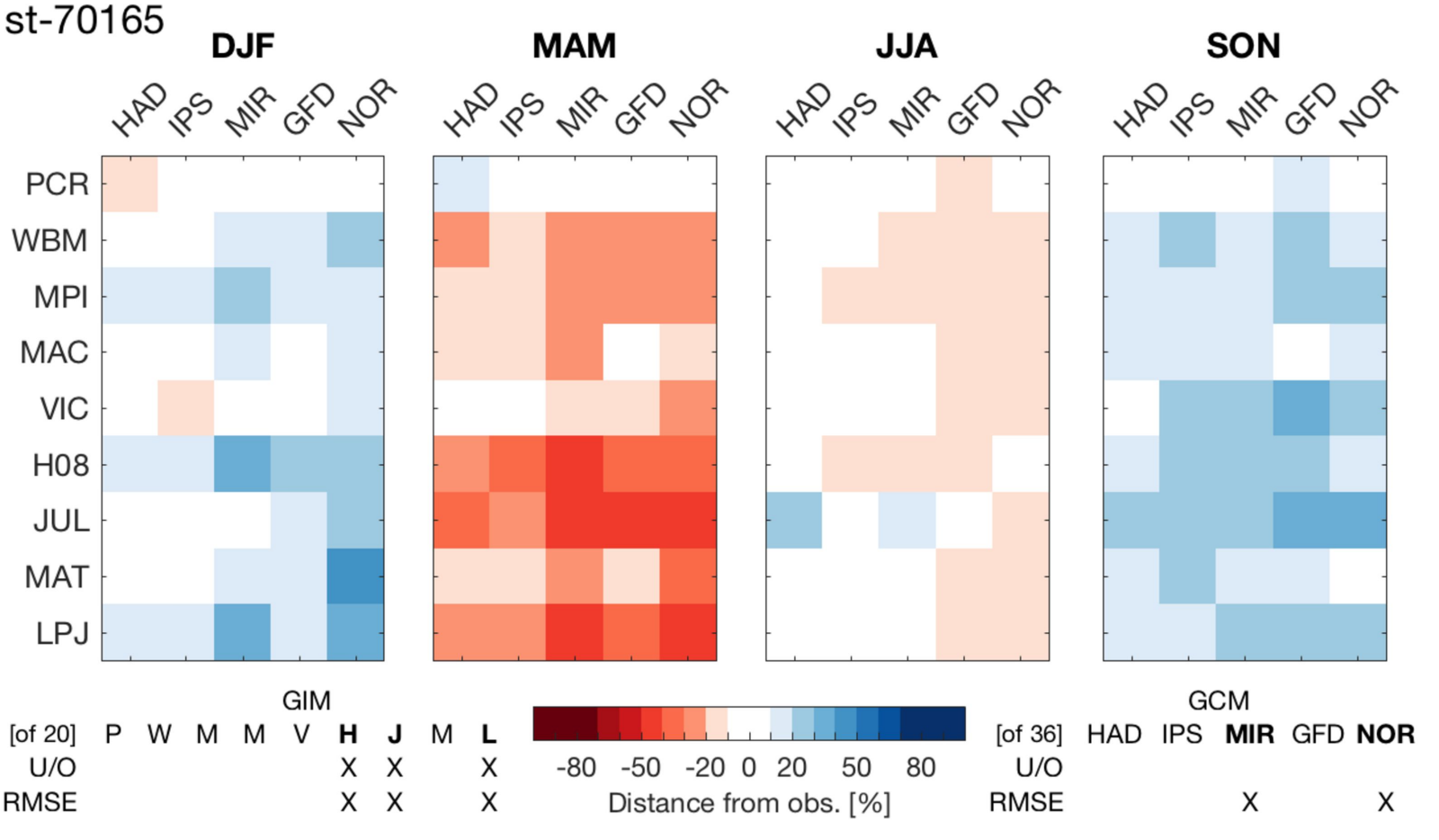
Likelihood

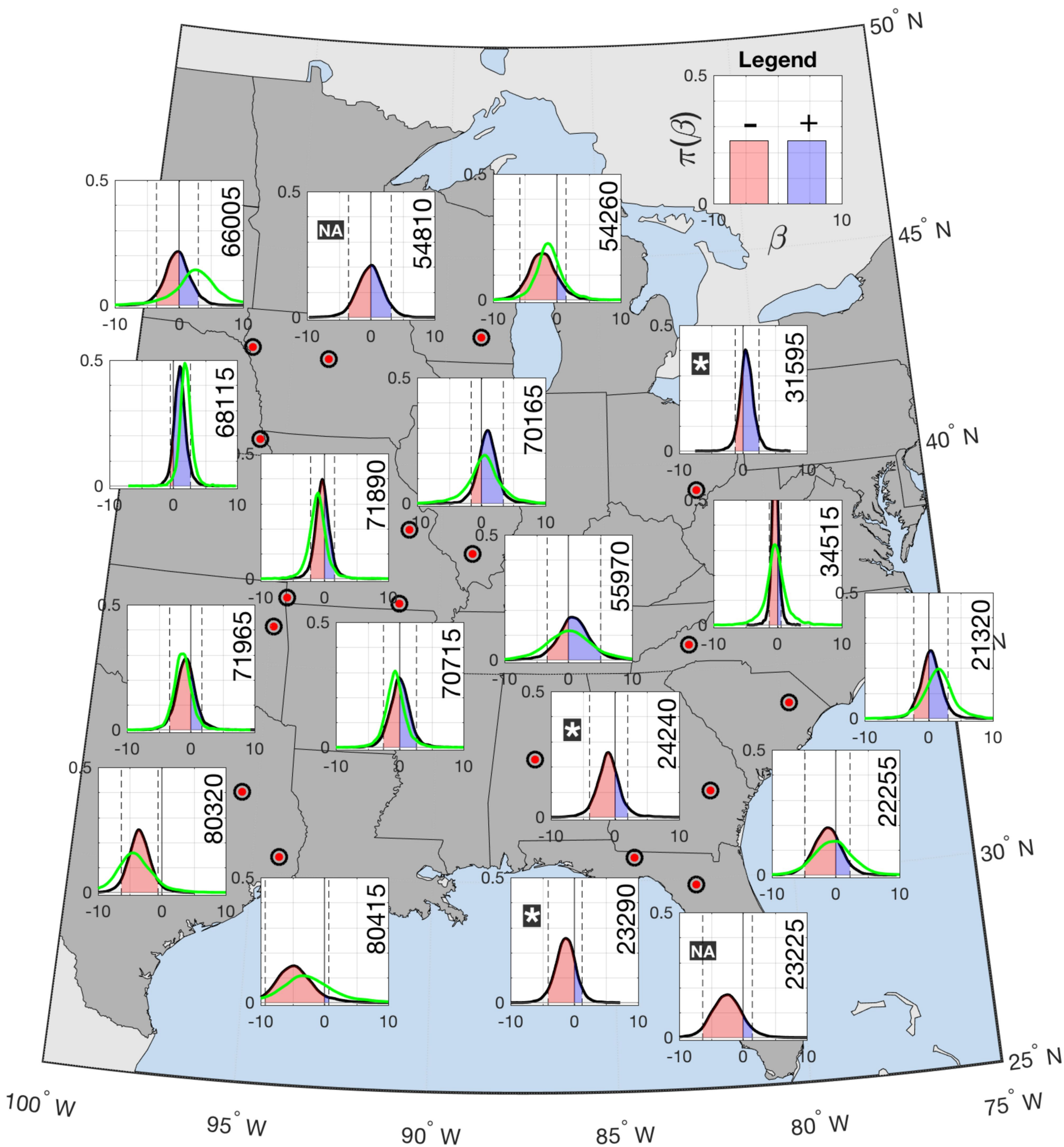
Distribution parameters

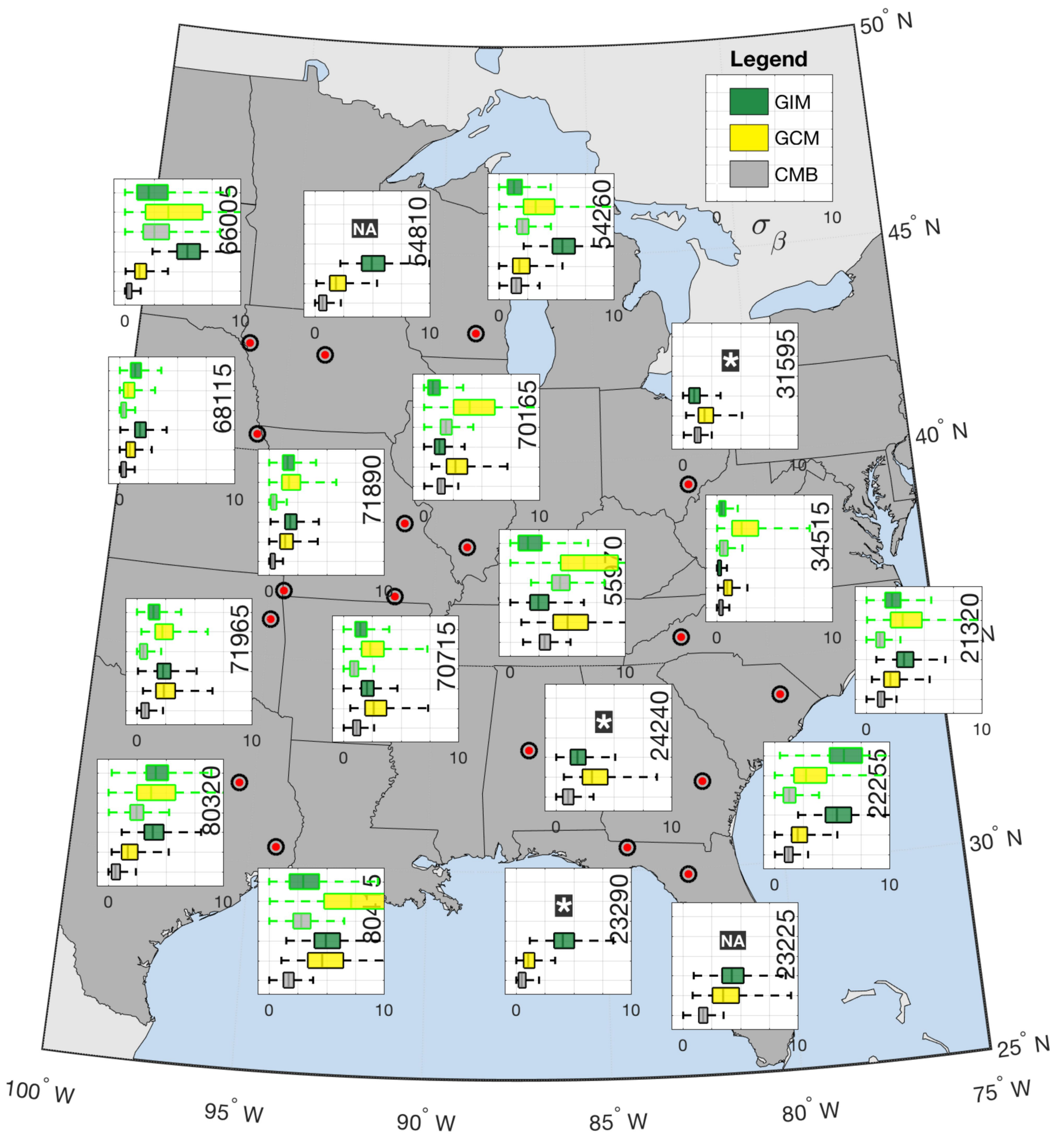
Population-level and group-level parameters

Hyper-parameters

st-70165


DJF
MAM
JJA
SON
HAD
IPS
MIR
GFD
NOR
PCR
WBM
MPI
MAC
VIC
H08
JUL
MAT
LPJ
GIM
[of 20] P W M M V H J M L
U/O X X X
RMSE X X X
-80 -50 -20 0 20 50 80
Distance from obs. [%]
GCM
[of 36] HAD IPS MIR GFD NOR
U/O
RMSE X X


Legend
$\pi(\beta)$
-
+
0.5
0
-10
10
β
66005
NA
54810
54260
68115
70165
31595
71890
71965
70715
65970
34515
21320
80320
24240
22255
80415
23290
NA
23225
50° N
45° N
40° N
30° N
25° N
100° W
95° W
90° W
85° W
80° W
75° W


Legend
GIM
GCM
CMB
σ
β
66005
NA
64810
54260
68115
70165
31595
71890
71965
70715
24240
80320
23290
NA
23225
22255
21320
34515
50° N
45° N
40° N
30° N
25° N
100° W
95° W
90° W
85° W
80° W
75° W