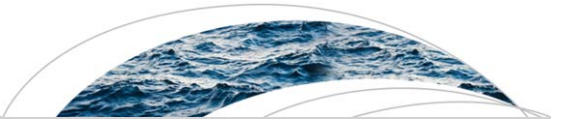
AGU PUBLICATIONS

Water Resources Research



RESEARCH ARTICLE

10.1002/2015WR018143

Combining information from multiple flood projections in a hierarchical Bayesian framework

Nataliya Le Vine[1]

[1]Department of Civil and Environmental Engineering, Imperial College London, London, UK

Key Points:
- Hierarchical Bayesian approach is used to combine flood projections from multiple models
- The approach explicitly accommodates shared multimodel discrepancy
- The approach ensures adequate baseline conditions and reduces flood estimate uncertainty

Correspondence to:
N. Le Vine,
nlevine@imperial.ac.uk

Citation:
Le Vine, N. (2016), Combining information from multiple flood projections in a hierarchical Bayesian framework, *Water Resour. Res.*, *52*, doi:10.1002/2015WR018143.

Received 25 SEP 2015
Accepted 13 MAR 2016
Accepted article online 18 MAR 2016

Abstract This study demonstrates, in the context of flood frequency analysis, the potential of a recently proposed hierarchical Bayesian approach to combine information from multiple models. The approach explicitly accommodates shared multimodel discrepancy as well as the probabilistic nature of the flood estimates, and treats the available models as a sample from a hypothetical complete (but unobserved) set of models. The methodology is applied to flood estimates from multiple hydrological projections (the Future Flows Hydrology data set) for 135 catchments in the UK. The advantages of the approach are shown to be: (1) to ensure adequate "baseline" with which to compare future changes; (2) to reduce flood estimate uncertainty; (3) to maximize use of statistical information in circumstances where multiple weak predictions individually lack power, but collectively provide meaningful information; (4) to diminish the importance of model consistency when model biases are large; and (5) to explicitly consider the influence of the (model performance) stationarity assumption. Moreover, the analysis indicates that reducing shared model discrepancy is the key to further reduction of uncertainty in the flood frequency analysis. The findings are of value regarding how conclusions about changing exposure to flooding are drawn, and to flood frequency change attribution studies.

1. Introduction

Multimodel ensembles (MMEs)—a set of predictions made by different models—are becoming widely accepted for flood frequency change analysis [e.g., *Arnell and Gosling*, 2014; *Hirabayashi et al*., 2013; *Kay and Jones*, 2012]. The use of multiple models results in large uncertainty around estimates of flood magnitudes, due to both uncertainty in model selection and natural variability of (simulated) river flow [*Madsen et al*., 1997]. The challenge is therefore to extract the most meaningful signal from the multimodel predictions, accounting for both model quality and uncertainties in individual model estimates.

Multimodel ensembles became increasingly widespread at the beginning of the 21st century, with a notable development being the IPCC reporting ensemble averages in the IPCC Third Assessment Report [*Penner et al*., 2001]. Concurrently, a probabilistic view on ensemble predictions was introduced by *Räisänen and Palmer* [2001], who assigned equal weights to each ensemble member. This was refined in the reliability ensemble average (REA) approach developed by *Giorgi and Mearns* [2002]; the authors proposed assigning weights to the ensemble member models based on each model's track record of bias and its distance from the ensemble's consensus; this was extended via a Bayesian analysis for the unknown model weights by *Tebaldi et al*. [2005]. Similarly, Bayesian Model Averaging method [*Hoeting et al*., 1999*; Raftery et al*., 2005] provides an estimate for a quantity of interest as a weighted average of predictive density functions provided by multiple models that are centered on the bias-corrected forecasts, with the weights reflecting the models' relative predictive skill. Other examples of MME analysis include hierarchical linear model fitting to aggregated observations using similarly aggregated MME projections [*Greene et al*., 2006], and use of spatial statistics to model geographical patterns [*Furrer et al*., 2007]. The typical challenge with model averaging is that the chosen weighting scheme depends on an arbitrary selection of the performance metric, which itself becomes an additional source of uncertainty. Further, a set of weights, derived via unavoidably arbitrary criteria and for a limited subset of possible models, cannot be reliably interpreted either as faithfully defining a probability distribution or as representative of the full model uncertainty range. Moreover, shared model components and information result in interdependencies in model outputs that potentially lead to persistent biases in MME, and to underestimation of the uncertainty in predictions [*Tebaldi and Knutti*, 2007]. Last,

© 2016. The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

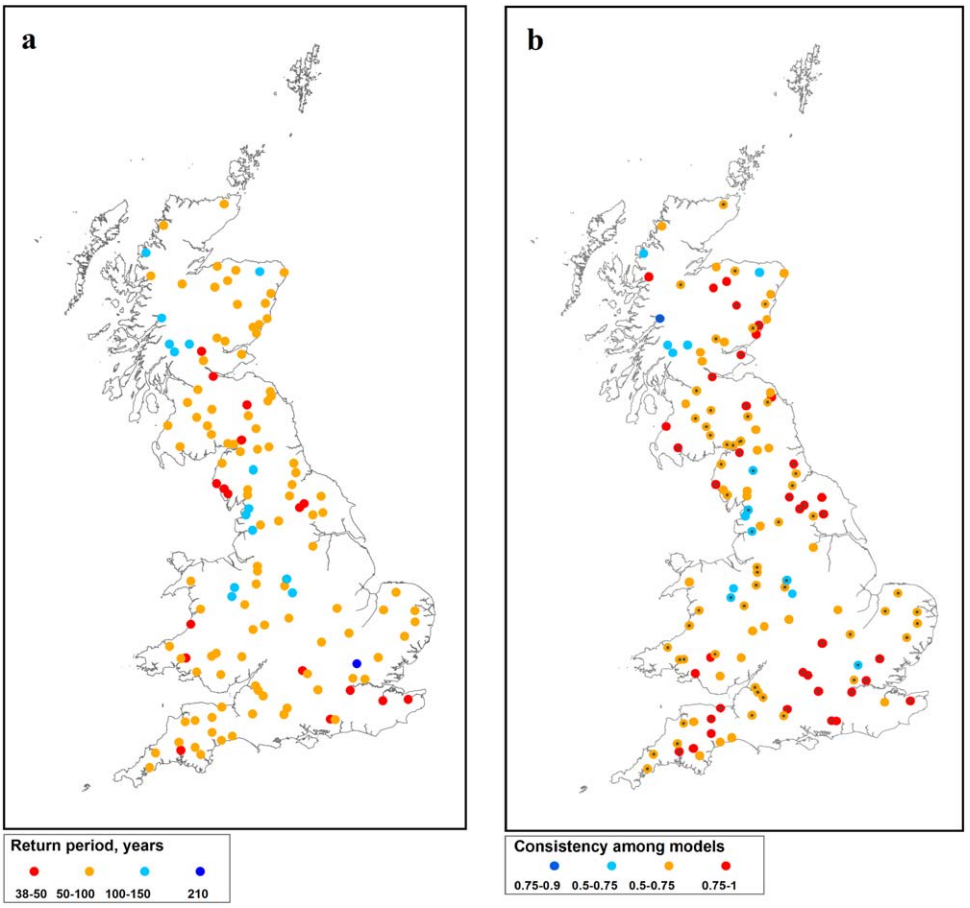


Figure 1. Projected change in *Q*100 based on MME. (a) Multimodel median return period (years) at the end of 21st century for discharge corresponding to the historic 100 year flood; (b) Model consistency—dark and light blue correspond to future decrease in flood frequency, red, and orange correspond to future increase in flood frequency; basins with at least one statistically significant change in *Q*100 are denoted by dots.

typical approaches suffer from a strong assumption of stationarity of the relation between observed and modeled trends that are estimated in the past and are applied to future projections [*Tebaldi and Knutti*, 2007]. For further particulars regarding standard practices, interested readers are referred to the excellent comprehensive reviews by *Tebaldi and Knutti* [2007] and *Knutti* [2010].

When combining models for flood frequency analysis, a further problem with averaging is that the arithmetic average is sensitive to extreme values. Current applications therefore focus on the median change (median is defined based on the set of multimodel predictions of change) as well as model "consistency," which refers to the fraction of models that show an increase/decrease in flood frequency [*Arnell and Gosling*, 2014; *Hirabayashi et al*., 2013; *Ward et al*., 2014]. A typical quantity of interest in such studies is the 100 year return period flood (*Q*100) estimated for past and future horizons (often taken as 30 year periods). To motivate the method developed in the paper, Figure 1 exemplifies such an analysis for 135 British catchments included into the *Future Flow Hydrology* data set [*Prudhomme et al*., 2013]. The data set provides simulated flows for years 1951–2098 based on 11 different *Regional Climate Model* variants and three hydrological models, and aims to support decision-making under changing environmental conditions. Further details about the data set and the models used are not required at this point, and are provided in section 2. Figure 1 compares *Q*100 ensemble estimates for the 1983–2012 historic period with *Q*100 ensemble estimates for the 2069–2098 future period. Based on the statistical properties of the extreme value distribution parameter estimates (discussed in detail in Appendix A), an analysis of the multimodel outputs shows that there is no statistically significant change in *Q*100 given by a model providing median prediction of change in the ensemble ($p = 0.05$) for 135 basins considered, and that there is no basin for which the majority of its MME

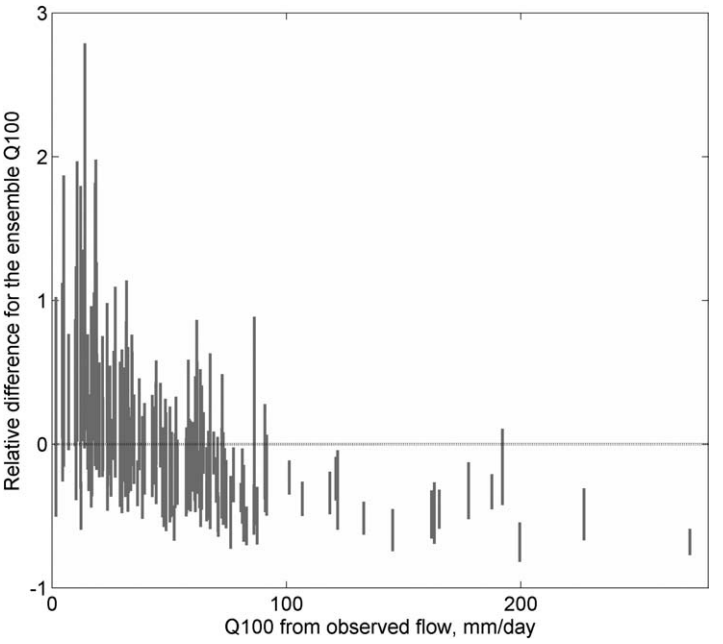


Figure 2. Relative differences for the historic multimodel 100 year flood estimates with respect to the estimates based on observations (the relative difference is defined as the difference between the model-based estimate and observation-based estimate of *Q100* divided by the observation-based estimate of *Q100*). The vertical grey lines connect minimum and maximum of the relative differences for each catchment. The horizontal dotted line shows a relative difference of 0 that corresponds to a model estimate exactly matching the observation-based estimate.

projected changes are statistically significant ($p = 0.05$), and the maximum number of statistically significant changes is 4 out of 11 projections. Furthermore, the multimodel historic *Q*100 estimates exhibit much disparity (Figure 2) and also fail to provide reliable bounds (lower and upper) for the observation-based *Q*100 estimates for 40% of the considered basins (54 out of 135). The inability to extract a meaningful signal out of highly uncertain multimodel estimates raises questions regarding whether meaningful conclusions can be drawn about changes in the exposure of assets to flooding [*Arnell and Gosling*, 2014; *Arnell et al*., 2013; *Hallegatte et al*., 2013; *Hirabayashi et al*., 2013] and the attribution of changes in flood frequency [*Ward et al*., 2014].

As described in section 3, this study demonstrates the potential of a recently proposed hierarchical Bayesian approach to combine information from multiple models [*Chandler*, 2013], in the context of flood frequency estimation. The methodology employs MME in a manner that explicitly accounts for (1) uncertainty in individual model estimates; (2) model-to-reality discrepancy, and (3) the available ensemble of models being only a subset from the complete (but hypothetical and therefore unobserved) set of models. In a hydrological context, hierarchical Bayesian approach has been previously applied to transfer information from multiple donor catchments for flood frequency analysis [*Kuczera*, 1982; *Renard*, 2011] and to define model parameter distributions for ungauged basins [*Smith et al*., 2014], as well as to assimilate multisource observations into hydrological model parameter and state distributions [*Wu et al*., 2010].

The objective of the proposed technique is to extract maximum useful statistical information from the divergent predictions of multimodel flood projections. The method requires specification of a "summary" metric for the behavior of interest that can be extracted from the MME and observational time series. In the context of air temperatures, regression coefficients for ambient air temperature are used as such a "summary" in *Chandler* [2013], facilitating the detection of warming/cooling trends. The present study proposes, in the context of flood frequency analysis, to treat parameters of probability distributions defining flood frequency as a summary for flow time series. This is preferred over an individual T-year flood magnitude, due to the added efficiency and consistency from analyzing an entire flood frequency curve.

2. Data

The analysis uses both simulated and observed daily flows for 135 basins in the UK. The simulated flows are part of Future Flows Hydrology data set [*Haxton et al*., 2012; *Prudhomme et al*., 2013], which is obtained from *the Future Flows Climate data set* [*Prudhomme et al*., 2012b]. Future Flows Climate is a gridded daily precipitation and monthly potential evapotranspiration time series ensemble derived from HadRM3-PPE that has been bias-corrected and downscaled at 1 and 5 km for hydrological applications [*Prudhomme et al*., 2012b; *Prudhomme et al*., 2013]. Future Flows Climate contains a set of 11 plausible time series from 1950 to 2098, capturing natural climate variability and climate change uncertainty, as characterized by 11

regional climate model variants for the "A1B" emission scenario. Further details regarding the Future Flows Climate data set can be found in *Prudhomme et al*. [2012b] and *Prudhomme et al*. [2012a].

Three conceptual hydrological models are used to derive Future Flows Hydrology, so that flow for the majority of basins (251) is simulated by one hydrological model, and flow at the remaining 30 sites is simulated by two hydrological models. The hydrological models used are CERF [*Griffiths et al*., 2006], PDM [*Moore*, 2007], and CLASSIC [*Crooks and Naden*, 2007]. CERF parameterization emphasizes water resources (water balance and low flows), while PDM and CLASSIC parameterizations place priority on the upper part of the flow regime and peak flows. Details regarding the models and their parameters can be found in *Prudhomme et al*. [2012a] and *Prudhomme et al*. [2013]. Consequently, Future Flows Hydrology contains an 11 member (or 22 member, for the 30 basins for which two hydrological models are employed) ensemble of projections from January 1951 to December 2098, each associated with a single realization from a different variant of HadRM3 and a single hydrological model.

The observed flows are obtained from the National River Flow Archive (CEH, UK; http://www.ceh.ac.uk/data/nrfa/data/search.html) as daily time series. Based on the data availability, a subset of 135 (out of 251) Future Flows Hydrology catchments is used in the analysis; the criterion for selection is that a site must have at least 30 years of measured flow records for a common period (1983–2012) (Figure 1). For the selected subset, flows in 113 catchments are simulated by one hydrological model (that corresponds to 11 ensemble members); and flows in the other 22 catchments (out of the 30 catchments) are simulated by two hydrological models (that corresponds to 22 ensemble members). Following the standard practice [*Arnell and Gosling*, 2014; *Hirabayashi et al*., 2013], the study considers two 30 year periods: (1) a historic period (for which observed data are available) between 1983 and 2012, and (2) a future period between 2069 and 2098.

3. Method

3.1. Conceptual Framework

One aim of hydrological projections is to reproduce observed statistical properties (e.g., flood and drought frequencies) over extended time periods, as opposed to an alternative objective of reproducing the system's temporal trajectory in detail. Parameters of a statistical model describing system outputs can be considered as such statistical properties/summaries to facilitate the characterization of MME outputs. In the context of flood frequency estimation, the Partial Duration Series (PDS) method is employed to statistically describe extreme hydrological events. It requires specification of the frequency of peak flows that exceed a pre-defined threshold, and the characteristics of exceedance flow magnitudes.

PDS assumes that the occurrence of peak flows follows a Poisson process with parameter λ, while exceedance flow magnitude is specified to follow a generalized Pareto (GP) distribution with parameters α and κ [*Madsen et al*., 1997]. One option is to use a fixed flow threshold q_0 (specific to each basin) for both observed and the modeled flow time series. This allows comparison of flood frequency parameters between the historic and future periods, as well as between the observed and simulated flow time series. For example, the flow threshold can be selected based on the observed time series for the historic period. However, using a fixed threshold may result in either too small (with less than one occurrence per year) or too large a number of threshold-exceedance peak flows. The former may propagate into a high degree of uncertainty regarding flood frequency, whereas the latter may violate the assumption that peak flows are characterized by a GP distribution. For example, if the threshold is chosen based on the historic observed flow records, and is consequently used with flows generated by a model underestimating high flows, the number of threshold exceedances will be small; this may result in unduly large uncertainty in estimated extremes.

Therefore, rather than using a fixed flow threshold, the present study employs a fixed average number of annual threshold exceedances (e.g., five exceedances per year, on average) to define individual flow thresholds for each time series [*Coles*, 2001, chap. 4; *Madsen et al*., 1997]. To allow the flood frequency parameter intercomparison, the PDS/GP parameters are converted into the annual maximum series/generalized extreme value distribution (AMS/GEV) parameters: location parameter ξ, scale parameter α^*, and shape parameter κ (the same shape parameter as in the GP distribution) [*Madsen et al*., 1997]. This establishes a direct correspondence between the three independent PDS/GP parameters (q_0 is defined by λ) and three

Table 1. Summary of Notation

Notation	Description
$\rho,\ \sigma$	Degrees of freedom for Wishart distributions
$\boldsymbol{\theta_i}$	Descriptor of the actual hydrological response for $i=0$; descriptor of the modeled hydrological response for $i=1,m$
$\hat{\boldsymbol{\theta}}_i$	Estimate of the underlying descriptor $\boldsymbol{\theta_i}$
$\boldsymbol{\delta_i}$	Model discrepancy $\boldsymbol{\delta_i}=\boldsymbol{\theta_i}-\boldsymbol{\theta_0}$
$\boldsymbol{\omega}$	Model ensemble discrepancy
$\boldsymbol{\mu_0}$	Expected value of the prior for $\boldsymbol{\theta_0}$
$\boldsymbol{J_i}$	Covariance matrix of $\boldsymbol{\theta}_i,\ i=1,\mathrm{m}$
$\boldsymbol{C_i}$	Conditional on $\boldsymbol{\omega}$ covariance matrix of discrepancy $\boldsymbol{\delta_i}$
$\boldsymbol{D_i}$	Defined as $\boldsymbol{C_i}+\boldsymbol{J_i}$
$\boldsymbol{\Lambda}$	Covariance matrix of $\boldsymbol{\omega}$
$\boldsymbol{\Sigma_0}$	Covariance matrix of prior distribution for $\boldsymbol{\theta_0}$
$\boldsymbol{R}$	Expected prior precisions of $\boldsymbol{\delta}_i,\ i=1,m$
$\boldsymbol{L}$	Expected prior precision on $\boldsymbol{\omega}$

AMS/GEV parameters. Further particulars regarding the annual maximum distribution corresponding to the PDS/GP model are discussed in Appendix A.

The six parameters (three each from one historic and one future period) can be combined into a vector $\boldsymbol{\theta}=\left(ln\ \xi_h,\ ln\alpha_h^*,\kappa_h,\ ln\xi_f,\ ln\alpha_f^*,\kappa_f\right)'$, which is termed a descriptor. Here, the subindex "*h*" indicates the parameters of the historic period, while the subscript "*f*" indicates the parameters of the future period. Due to the subsequent reliance on Gaussian distributions, the descriptor $\boldsymbol{\theta}$ is defined using log-transformed values of the location parameter ξ and scale parameter α^*, both of which are strictly positive. The "actual flood frequency" will be summarized using a descriptor $\boldsymbol{\theta}_0$, and outputs from m models will be summarized using descriptors $\boldsymbol{\theta_i},\ i=1,m$. For ease of exposition, the notation used is consistent with that of *Chandler* [2013], and is summarized in Table 1.

The uncertainty in an MME can be described by considering the descriptors $\boldsymbol{\theta_i},\ i=1,m$ to be drawn from some underlying probability distribution. Following *Chandler* [2013], Figure 3 illustrates how these model descriptors relate to each other and to (imperfectly observed) reality. The model descriptors are centered on $\boldsymbol{\theta}_0+\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is an ensemble discrepancy shared by all models that in principle could have been added to the ensemble. The grey-dashed lines in Figure 3 represent estimation errors from internal variability in the real and simulated flow time series; and the quantities $\hat{\boldsymbol{\theta}}_i,\ i=0,m$ are the maximum likelihood (ML) estimates of the underlying descriptors $\boldsymbol{\theta_i},\ i=0,m$, so that minimal information is lost [*Casella and Berger*, 1990, chap.6; *Chandler*, 2013; *Leith and Chandler*, 2010; *Pawitan*, 2001, chap. 3].

3.2. The Posterior Distribution: Gaussian Specification

3.2.1. Hierarchical Model

A Hierarchical Bayesian approach is employed to provide information regarding the actual hydrological response descriptor $\boldsymbol{\theta}_0$ using information contained in the descriptor estimates $\hat{\boldsymbol{\theta}}_i,\ i=0,m$ [*Chandler*, 2013].

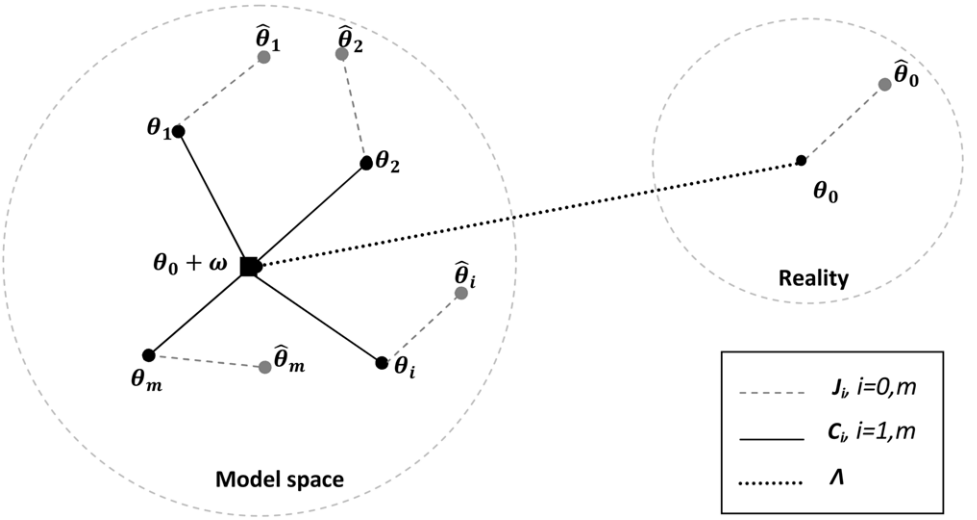


Figure 3. Schematic representation of the MME combining framework proposed by *Chandler* [2013]. Black-dotted line represents shared model discrepancy $\boldsymbol{\omega}$ characterized by a covariance matrix Λ in the Gaussian formulation, with simulator output centered on $\boldsymbol{\theta}_0+\boldsymbol{\omega}$. Grey-dashed straight lines represent estimation errors characterized by covariance matrices $\boldsymbol{J_i}$, $i=0,m$ in the Gaussian formulation. Black solid lines represent the propensity of an individual model to deviate from the model consensus with covariance matrices $\boldsymbol{C_i}$, $i=0,m$ in the Gaussian formulation.

A set of assumptions regarding the distributions of various quantities in Figure 3 are required to specify the posterior distribution of θ_0. The case when all of the required distributions are multivariate normal (Gaussian) is examined here, and section 4 discusses the applicability/validity of the assumptions. This can be expressed as the following hierarchical model:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{i}} \sim N(\boldsymbol{\theta}_{\boldsymbol{i}}, \boldsymbol{J}_{\boldsymbol{i}}), \;\; i=0, m \tag{1}$$

$$\boldsymbol{\delta}_{\boldsymbol{i}}=\boldsymbol{\theta}_{\boldsymbol{i}}-\boldsymbol{\theta}_{\mathbf{0}} \sim N(\boldsymbol{\omega}, \boldsymbol{C}_{\boldsymbol{i}}), \;\; i=1, m \tag{2}$$

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \boldsymbol{\Lambda}) \tag{3}$$

$\boldsymbol{J}_{\boldsymbol{i}}$ denotes a covariance matrix of the ML estimate for the ith data source, matrix $\boldsymbol{C}_{\boldsymbol{i}}$ measures the propensity of the ith model to deviate from the model consensus and $\boldsymbol{\Lambda}$ is covariance matrix of shared discrepancy $\boldsymbol{\omega}$. The discrepancies $\boldsymbol{\delta}_{\boldsymbol{i}}, \;\; i=1, m$ are assumed to be independent given $\boldsymbol{\omega}$ (see section 4). The assumption can be relaxed by extending the general framework in Figure 3 to include families of models, which are centered on their own consensuses, with those centered on the ensemble consensus.

For a given catchment and a given hydrological model (but different climate variants), the descriptors derived from simulations are considered to be equally credible, and the corresponding covariance matrices $\boldsymbol{C}_{\boldsymbol{i}}$ are set to be equal. Consequently, this results in either one or two unique covariance matrices $\boldsymbol{C}_{\boldsymbol{i}}$, when, correspondingly, one or two hydrological models are used to represent the catchment hydrological response. For simplicity, a case with one hydrological model is described below, i.e., $\boldsymbol{C}=\boldsymbol{C}_{\mathbf{1}}=\ldots=\boldsymbol{C}_{\boldsymbol{m}}$; similar derivations can be obtained for a case with two distinct matrices $\boldsymbol{C}_{\boldsymbol{i}}$.

The hierarchical structure of the model is completed with prior distributions for $\boldsymbol{\theta}_{\mathbf{0}}$, $\boldsymbol{C}$, and $\boldsymbol{\Lambda}$. This is done by choosing conjugate forms of these distributions, namely,

$$\boldsymbol{\theta}_{\mathbf{0}} \sim N(\boldsymbol{\mu}_{\mathbf{0}}, \boldsymbol{\Sigma}_{\mathbf{0}}) \tag{4}$$

$$\boldsymbol{C}^{-\mathbf{1}} \sim W\left((\rho \boldsymbol{R})^{-1}, \rho\right) \tag{5}$$

$$\boldsymbol{\Lambda}^{-1} \sim W\left((\sigma \boldsymbol{L})^{-1}, \sigma\right) \tag{6}$$

W denotes a Wishart distribution [*Carlin and Louis*, 2000, chap.5]; $\boldsymbol{R}^{-\mathbf{1}}$ is an expected prior precision (precision matrix is an inverse of covariance matrix) of $\boldsymbol{\delta}_{\boldsymbol{i}}, \;\; i=1, m$; $\boldsymbol{L}^{-\mathbf{1}}$ is an expected prior precision on $\boldsymbol{\omega}$; and ρ and σ are degrees of freedom for the Wishart distributions.

The marginal posterior for $\boldsymbol{\theta}_{\mathbf{0}}$ is sought given only the data $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}, \;\; i=0, m$. The hyperparameters (parameters of a prior distribution) $\rho, \;\; \boldsymbol{L}, \;\; \sigma, \boldsymbol{\mu}_{\mathbf{0}}, \;\; \boldsymbol{\Sigma}_{\mathbf{0}}$ are assumed to be known; and the covariance matrices $\boldsymbol{J}_{\boldsymbol{i}}, \;\; i=0, m$ are calculated (fixed) using the Fisher information matrix (Appendix A). However, only the "historic" part of $\boldsymbol{J}_{\mathbf{0}}$ can be defined in this way due to future observations not being knowable. Therefore the "future" components of $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$ are considered as estimated with zero precision giving $\boldsymbol{J}_{\mathbf{0}}^{-\mathbf{1}}=\begin{pmatrix} \boldsymbol{J}_{\boldsymbol{h}}^{-\mathbf{1}} & 0 \\ 0 & \boldsymbol{J}_{\boldsymbol{f}}^{-\mathbf{1}} \end{pmatrix}=\begin{pmatrix} \boldsymbol{J}_{\boldsymbol{h}}^{-\mathbf{1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. For convenience, the fixed parameters are excluded from the notation while specifying the conditional distribution functions.

3.2.2. Gibbs Sampling

Gibbs sampling is used here to approximate the posterior marginal distribution for $\boldsymbol{\theta}_{\mathbf{0}}$ by sampling from the full conditional distributions while treating the remaining parameters as known (fixed) [*Carlin and Louis*, 2000, chap. 5], and proceeds as follows:

0. Initialize all unknown quantities, set the iteration counter to 0.
1. Increment the iteration counter by 1; store the current values of all unknown quantities.
2. Calculate the conditional distribution of $\{\boldsymbol{\theta}_i\}_{i=0}^{m}$ given the current values of the other unknown quantities in the model, and replace the current values of these $\{\boldsymbol{\theta}_i\}_{i=0}^{m}$ with a sample drawn from this conditional distribution.
3. Calculate the conditional distribution of $\boldsymbol{C}^{-\mathbf{1}}$ given the current values of the other unknown quantities, and replace the current value of $\boldsymbol{C}^{-\mathbf{1}}$ with a sample from this distribution.
4. Replace the current value of $\boldsymbol{\Lambda}^{-1}$ in a similar way.
5. Replace the current value of $\boldsymbol{\omega}$ in a similar way.

6. If the iteration limit is reached, stop; otherwise return to step 1.

The full conditional distribution of the descriptors $\{\theta_i\}_{i=0}^m$ can be expressed as follows

$$p\left(\{\theta_i\}_{i=0}^m \middle| \left\{\hat{\theta}_i\right\}_{i=0}^m, \ \mathbf{\Lambda}, \ \boldsymbol{C}, \ \boldsymbol{\omega}\right) = p\left(\{\theta_i\}_{i=1}^m \middle| \left\{\hat{\theta}_i\right\}_{i=1}^m\right) * p\left(\theta_0 \middle| \left\{\hat{\theta}_i\right\}_{i=0}^m, \ \mathbf{\Lambda}, \ \boldsymbol{C}, \ \boldsymbol{\omega}\right) \tag{7}$$

The first term on the left-hand side is proportional to a product of normal distributions from (1), and the second term is provided by *Chandler* [2013] as

$$\begin{gathered} p\left(\theta_0 \middle| \left\{\hat{\theta}_i\right\}_{i=0}^m, \ \mathbf{\Lambda}, \ \boldsymbol{C}, \ \boldsymbol{\omega}\right) \sim N(\boldsymbol{\tau}, \boldsymbol{S}), \ \text{where} \\ \boldsymbol{S}^{-1} = \mathbf{\Sigma}_0{}^{-1} + \mathbf{J}_0{}^{-1} + \left[\mathbf{\Lambda} + \left(\sum_{i=1}^m \boldsymbol{D}_i^{-1}\right)^{-1}\right]^{-1} \\ \boldsymbol{\tau} = \boldsymbol{S}\left[\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{J}_0^{-1}\hat{\theta}_0 + \left(\boldsymbol{I} + \sum_{k=1}^m \mathbf{D}_k^{-1}\mathbf{\Lambda}\right)^{-1} \sum_{i=1}^m \mathbf{D}_i^{-1}\hat{\boldsymbol{\theta}}_i\right] \end{gathered} \tag{8}$$

$\boldsymbol{I}$ is the *6 x 6* identity matrix, and $\boldsymbol{D_i} = \boldsymbol{C} + \boldsymbol{J_i}$.

Further, combining the priors (5) and (6) with the likelihood for the random effects (2) and (3) produces the updated Wishart full conditional distributions for the covariance matrices:

$$p\left(\boldsymbol{C}^{-1} \middle| \ \{\theta_i\}_{i=0}^m, \ \left\{\hat{\theta}_i\right\}_{i=0}^m, \ \mathbf{\Lambda}, \ \boldsymbol{\omega}\right) \sim W\left(\left(\sum_{i=1}^m (\theta_i - \bar{\theta}_i)' \cdot (\theta_i - \bar{\theta}_i) + \rho \boldsymbol{R}\right)^{-1}, \rho + m - 1\right) \tag{9}$$

$$p\left(\mathbf{\Lambda}^{-1} \ | \ \{\theta_i\}_{i=0}^m, \ \left\{\hat{\theta}_i\right\}_{i=0}^m, \ \mathbf{C}, \ \boldsymbol{\omega}\right) \sim W\left((\boldsymbol{\omega}' * \boldsymbol{\omega} + \sigma \boldsymbol{L})^{-1}, \sigma + 1\right) \tag{10}$$

Last, using (2) and (3) it can be shown [*Lindley and Smith*, 1972] that the full conditional of $\boldsymbol{\omega}$ is

$$p\left(\boldsymbol{\omega} \ | \ \{\theta_i\}_{i=0}^m, \ \left\{\hat{\theta}_i\right\}_{i=0}^m, \ \mathbf{\Lambda}, \ \boldsymbol{C}\right) \sim N\left(\boldsymbol{V} * m\boldsymbol{C}^{-1} * \bar{\boldsymbol{\delta}}, \boldsymbol{V}\right) \tag{11}$$

where $\boldsymbol{V} = \left(m\boldsymbol{C}^{-1} + \mathbf{\Lambda}^{-1}\right)^{-1}$.

Sampling sequentially from the full conditional distributions (7) and (9)–(11) while the remaining parameters are kept fixed and equal to their most recent sampled values approximates the joint distribution $p\left(\{\theta_i\}_{i=0}^m, \ \mathbf{\Lambda}, \ \boldsymbol{C}, \ \boldsymbol{\omega} \ | \ \left\{\hat{\theta}_i\right\}_{i=0}^m\right)$, and the marginal posterior $p\left(\theta_0 \middle| \ \left\{\hat{\theta}_i\right\}_{i=0}^m\right)$, in particular. The starting point $\mathbf{\Lambda}$, $\boldsymbol{C}$ for a Gibbs sampling chain is obtained from the ML estimates of the descriptors $\left\{\hat{\theta}_i\right\}_{i=0}^m$ as proposed by *Chandler* [2013] (Appendix B), and a single chain is run for 20,000 iterations. While diagnosing the Gibbs sampling convergence is a nontrivial problem [*Carlin and Louis*, 2000, chap. 5], the first 10,000 iterations are used here as a warm-up sample and discarded from the further analysis. As argued by *Geyer* [1992], only the number of iterations for the autocovariances to decay to a negligible level is to be discarded, and that less than 1% of the run will normally be sufficient for a burn-in period (with a sample size of 10,000 order). Further, a visual inspection shows that trace plots of the iteration history for the elements of θ_0 (the primary quantities of interest) have been stabilized after 10,000 first (discarded) iterations.

3.2.3. Selection of Hyperparameters

The hyperparameter values are chosen to determine very vague priors (4)–(6), namely $\rho = \sigma = 6$ (minimum degrees of freedom for the Wishart distributions), and $\mathbf{\Sigma_0^{-1}} = \mathbf{0}$ (complete prior ignorance on θ_0). Hyperparameter $\boldsymbol{R}^{-1}$ specified as the expected values of the precision matrix $\boldsymbol{C}^{-1}$ in (5) is selected using information from "donor" catchments. The donor catchments are defined among the catchments that are represented by the same (as the "target" catchment) hydrological model in the data set (e.g., PDM, CLASSIC, or CERF). Further, the catchment similarity is based on the catchment similarity measure used by the *Institute of Hydrology* [1999].The measure includes the primary factors categorizing rural catchments in the UK, namely, catchment area, standardized annual average rainfall, and estimated base flow index (details can be found in *McIntyre et al.* [2005]). The hyperparameter $\boldsymbol{R}$ is selected as an average of matrix $\boldsymbol{C}$ estimates ((B1) in Appendix B) from three most similar catchments.

The same regionalization procedure is used to specify the hyperparameter $\boldsymbol{L}$ as an average of $\boldsymbol{\Lambda}$ estimates ((B2) in Appendix B) from three most similar catchments. For each of the donor catchments, due to unavailability of future empirical data, the shared discrepancy covariance matrix $\boldsymbol{\Lambda}$ can be estimated only for the historic period (Appendix B). To reflect uncertainty about the magnitude of future shared discrepancy, a constant $K \geq 0$ is specified to represent prior knowledge (Appendix B). The value $K=0$ represents the shared discrepancy spread remaining the same between the past and the future; this would reflect a circumstance in which model performance remains broadly the same between two different (sufficiently long) time periods (i.e., the model calibration period and forecasting period) [e.g., *Beven*, 2011; *Tebaldi and Knutti*, 2007]. Values of K larger than zero represent a condition in which the shared discrepancy spread increases in the future. For example, when $K=1$, the formulation in Appendix B suggests that the additional discrepancy in the future (term $\boldsymbol{\eta}$ in the discussion following equation (B2) is of the same magnitude as the historic discrepancy $\boldsymbol{\omega_h}$; this means that the future discrepancy could be as small as zero or as large as double the magnitude of the historical discrepancy. When significantly longer observation records (longer than the 30 years used in the analysis) are available, it is possible to use the remaining observations (i.e., the years that are not used in the estimation process) to specify the prior on the discrepancy spread for both historic and future periods. However, due to the limited time series of data availability in this study, as is typically the case, the choice of K is unavoidably subjective and there is no theoretical or empirical support for any specific non-zero value of K. We therefore adopted the standard strategy of specifying that $K=0$, and then subsequently quantifying the sensitivity of outputs to this arbitrary selection of its numerical value.

The hyperparameters $\boldsymbol{R}$ and $\boldsymbol{L}$ reflect the scale of elements in the covariance matrices $\boldsymbol{C}$ and $\boldsymbol{\Lambda}$, respectively. While the posterior in (9) shows that there is substantial evidence available to update an estimate for $\boldsymbol{C}$, there is only one data point available to update an estimate for $\boldsymbol{\Lambda}$ as shown in the posterior (10). Hence, an estimator of $\boldsymbol{\Lambda}$ will be mainly based on the regionalized information used in the procedure.

4. Applicability of the Assumptions

The Gaussian specification in the previous sections relies on several assumptions. In this section, we set out the assumptions and report the results of tests that have been carried out to assess their validity for the data used in this study. The assumptions (A0–A4) are as follows:

A0. The descriptors $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}},\ i=0,m$ summarize flow time series for flood frequency analysis.

The descriptors $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}},\ i=0,m$ represent transformed PDS/GP parameters. For each catchment, multiple hypotheses that the occurrence of peak flows follows Poisson process and that exceedance flow magnitude are drawn from a generalized Pareto distribution are tested for both observed and modeled flows, and for both historic and future time periods. The suitability of the Poisson assumption has been tested by means of the dispersion index [*Cunnane*, 1979], while the suitability of the Generalized Pareto assumption has been tested by means of the Cramér-von Mises statistic [*Choulakian and Stephens*, 2001]. As the test requires testing multiple catchment-specific hypotheses (46 or 90 hypotheses per catchment, depending on the number of models in the ensemble), the likelihood of a rare event increases, and therefore, the likelihood to incorrectly reject a null hypothesis (Type I error) becomes greater. The Holm-Bonferroni method [*Holm*, 1979] is used here to control the family-wise error rate (the probability of witnessing one or more Type I errors) by adjusting the rejection criteria of each of the individual hypotheses. All catchment-specific hypotheses cannot be rejected at 0.05 family-wise error rate for 114 catchments out of 135 catchments considered; and are rejected for 1, 2, 3, and 6 individual hypotheses (out of 46 or 90) for 5, 13, 2, and 1 catchments, correspondingly. These results are all consistent with what is expected if assumption A0 holds.

A1. The descriptor estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}},\ i=0,m$ are unbiased and Gaussian.

While the descriptor estimates are approximately multivariate normal for large samples [*Millar*, 2011, chap. 12.2], it can be shown (Appendix C) that it still provides a reasonable approximation for the study sample size of 150 flow peaks. It is to be noted, that for low shape parameter (e.g., $\kappa = -1$), the descriptor distributions have "fatter" tails and sharper "peaks" than the corresponding asymptotic multivariate normal distributions.

A2. The model discrepancies $\boldsymbol{\delta_i},\ i=1,m$ are Gaussian.

The use of simulator-specific covariance matrices $\boldsymbol{C_i}$ provides flexibility to accommodate outlying simulators via distributions that are highly dispersed rather than heavy-tailed.

A3. The model discrepancies $\boldsymbol{\delta_i}$, $i=1,m$ are independent.

A necessary condition for independence, when all cross-covariances are zero, is examined. For this purpose, sample cross-covariances of the standardized model discrepancies (Appendix D) are calculated based on information from all catchments in the data set; and a family of hypotheses about the lack of cross-correlation is tested. It is found that the hypotheses cannot be rejected at a family-wise error rate of 0.05 [*Holm*, 1979].

A4. The shared discrepancy ω is Gaussian.

It is impossible to verify the appropriateness of the assumption, as a multimodel ensemble provides only a single realization of ω; the assumption is merely a convenient device to incorporate the shared discrepancy formally into the analysis [e.g., *Chandler*, 2013; *Kennedy and O'Hagan*, 2001].

5. Results and Discussion

A T-year flood magnitude (*QT*) is estimated from the (1) MME projections, (2) observed flow time series, and (3) combining the MME and observational information, for both historic and future periods, for each of the 135 catchments. The flood estimates for the observations and MME are derived individually using the AMS/GEV assumption and (A3), while the flood estimates based on the combined information are estimated using the proposed method and (A3). Confidence intervals for flood estimates based on only observations or only MME information are estimated using Monte-Carlo sampling of the corresponding asymptotically normal AMS/GEV parameter distributions (Appendix A); while confidence intervals for the flood estimates based on the combined information are approximated using the AMS/GEV parameter values drawn using a Gibbs sampling scheme (described in section 3). While *Q*100 is the focus for the discussion, due to its wide usage in the literature; a set of findings for the other return periods (i.e., $T=10$ and $T=25$) are also shown. For flood projections alone, flood change median estimates and model consistency are shown in Figure 1, and model prediction spread is shown in Figure 2. The *Q*100 estimates from the flow projections are widely spread; nevertheless, 40% (54 out of 135) of observation-based *Q*100 estimates are outside of the range of the ensemble-based *Q*100 estimates. Therefore, historic "baseline" conditions (*Q*100) estimated by the MME for the purposes of flood frequency change assessment can be seen to be unreliable.

In contrast, the Bayesian analysis strategy of combining information results in estimates of *Q*100 that are consistent with *Q*100 estimated from the observational evidence (for the historic period); thus yielding reliable approximation of "baseline" (historic) conditions. Figure 4a shows the projected changes in flood frequency at the end of the 21st century based on combined information. Assuming no change in model discrepancy spread between the historic and future periods ($K=0$), there are 4 out of 135 basins whose posterior 95% credible intervals for the change in *Q100* exclude zero (Appendix A describes the procedure for the interval estimation). If spread in the future model discrepancy magnitude increases (the case of $K=1$ is considered here), all posterior 95% credible intervals for the change in *Q100* contain zero. This shows that the conclusions are very sensitive to the unverifiable (but necessary) prior assumption regarding the future magnitude of model discrepancy (see section 3.2.3). And while this study does not provide guidance on how to treat or resolve this sensitivity given the data limitations, the study raises awareness that there is a significant problem that requires a solution. Figure 4b shows maximum sizes (defined via probability mass contained in the interval) of the posterior credible intervals to exclude zero in *Q100* change ($K=0$), so that the larger the interval size the stronger the support for change in *Q100*. It shows, for example, that there are four river sites with zero change in *Q100* outside of the posterior 95% credible intervals, and a further thirteen when the credible interval size is decreased to 90%.

Further, Figure 5 shows the posterior 95% credible intervals for the relative change in *Q100* with respect to *Q100* estimated from the observed (historic) flow records, arranged in the observation-based *Q100* increasing order (*Q100* is in mm/d); so that the first set of intervals corresponds to a site with the smallest observation-based *Q100* estimate, and the last set corresponds to the largest observation-based *Q100 estimate*. A positive value of the relative change indicates a decrease in future *Q100*, and a negative value indicates an increase in future *Q100*. The figure shows 95% credibility intervals for the relative change in *Q100* given by all models in the ensemble, a model that provides a median change in *Q100*, and by combining information with $K=0$ and $K=1$. The uncertainty in the combined estimator change for $K=0$ is generally smaller than the uncertainty in all-model or median estimator changes, and is somewhat bigger than that

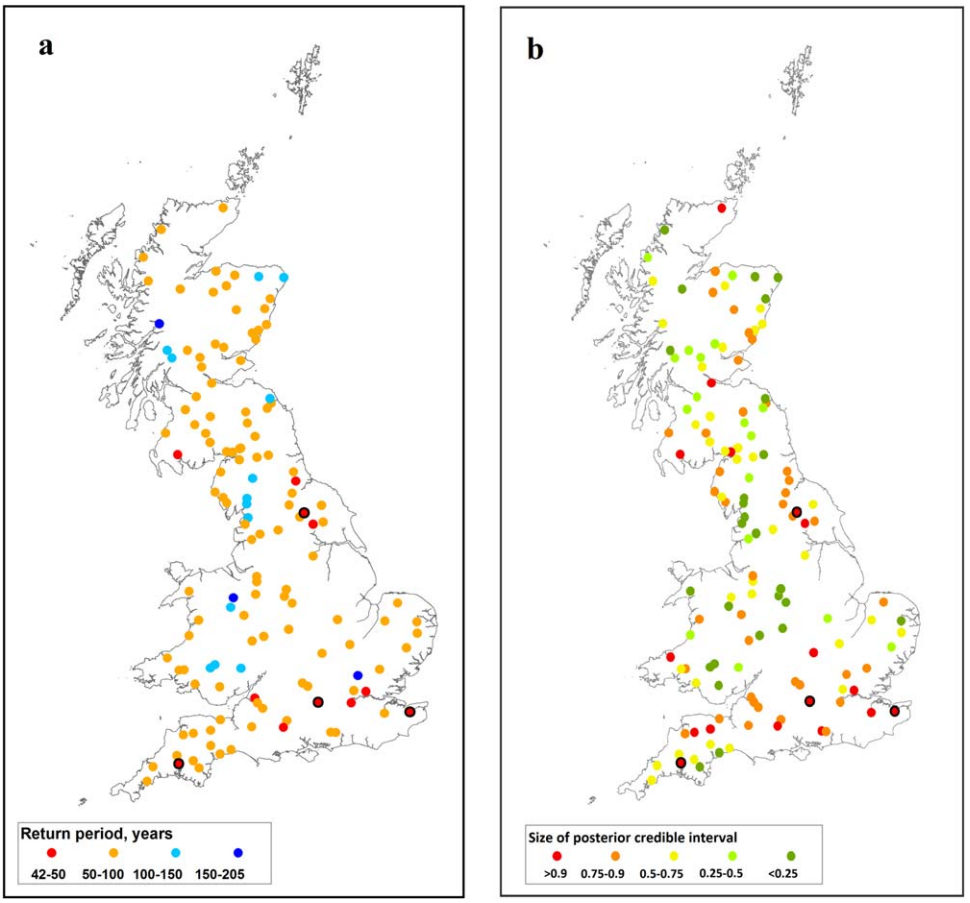


Figure 4. Projected change in flood frequency based on combined information. (a) Median return period (years) at the end of 21st century for discharge corresponding to the historic 100 year flood; (b) maximum size of posterior credible intervals (defined via probability mass contained in the interval) that exclude zero in *Q100* change (K = 0)—the larger the size the stronger the support for change in *Q100*. Basins with the posterior 95% credible intervals that exclude zero change in *Q100* are denoted by black circles (K = 0).

for the sites with the largest *Q100* estimates (mm/d). The latter is attributed to the model underpredictive bias for the sites with large *Q100*, and is discussed later in the section. The comparison of the posterior credible intervals for K = 0 and K = 1 in Figure 5 further illustrates that the estimates are very sensitive to the prior assumption regarding the future magnitude of model discrepancy.

By comparing the current protocol (Figure 1) and the combined information approach (Figure 4), three cases can be highlighted to examine details on the inner working of the considered method. The first case (Figure 6a) exemplifies the reduction in individual model uncertainty when information is combined, so that the posterior 95% credible interval for the change in *Q100* based on the combined information excludes zero (*K* = 0), while 18 out of 22 posterior 95% credible intervals based only on the MME projections include zero. This case is characterized by a low shared model discrepancy, such that the proposed method reduces the uncertainty around *Q*100 estimates (as compared with uncertainty in *Q*100 derived from observations and MME). The second case illustrates a situation when individual ensemble members are consistent on the direction of change (11 out of 11 models estimate an increase in flood frequency), but the combined information does not support change in *Q*100, as the corresponding posterior 95% credible interval for the change in *Q100* includes zero (Figure 6b). Here, the discrepancy between the models and observations is much larger than for the case in Figure 5a, which inflates the uncertainty associated with the MME estimates (see the discussion for Figure 6 below), and does not allow a larger reduction in the uncertainty of the combined flood estimate with respect to the uncertainty in the observation-based estimate (see (8)). And the third case shows a situation when the posterior 95% credible interval for the median change in *Q25* excludes zero, suggesting a change in *Q25* (this example is considered as all posterior 95%

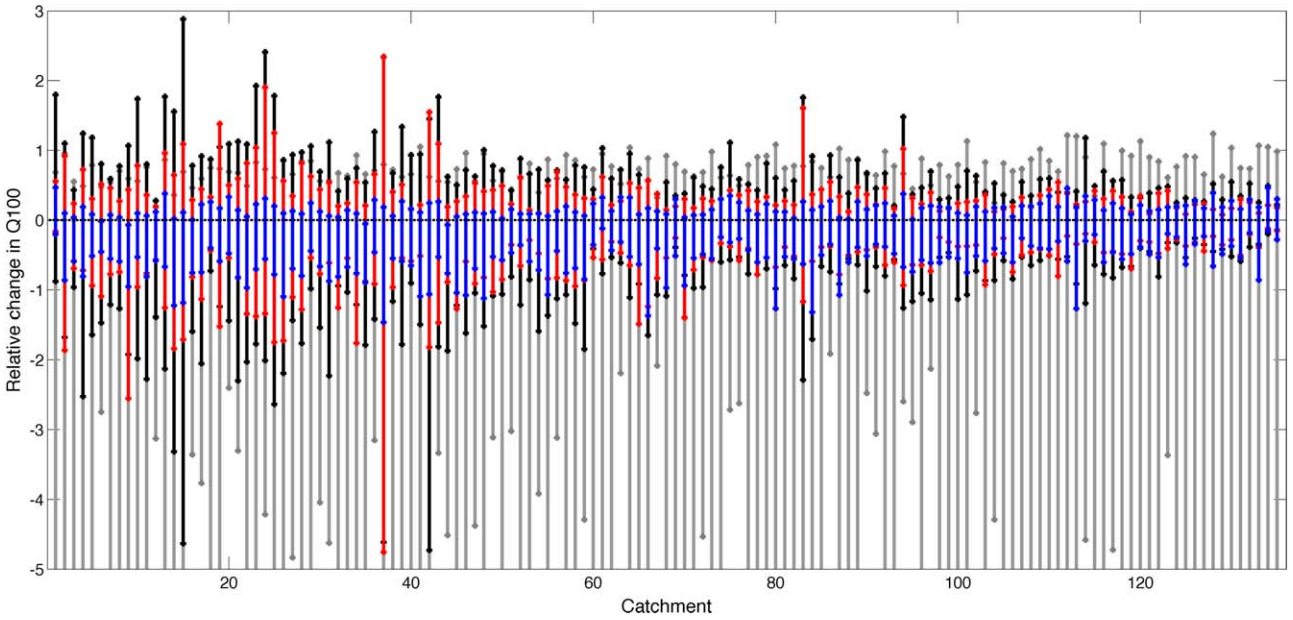


Figure 5. Relative changes in *Q100* with respect to the observation-based estimate of *Q100*, arranged in the observation-based *Q100* increasing order. Blue indicates 95% credible intervals based on combined information for K = 0, red indicates 95% credible intervals based on median predictions, black indicates 95% credible intervals based on all 11 (or 22) models, and grey indicates 95% credible intervals based on combined information for K = 1.

credible intervals for the median change in *Q*100 include zero), but the combined information does not support change in *Q*25, as the corresponding posterior 95% credible interval for the change in *Q25* includes zero (Figure 6c). To assess change in *QT*, the currently adopted protocol selects a representative ensemble member based on the median of the return periods of the past *QT* estimates on the corresponding future flood frequency scales [e.g., *Hirabayashi et al*., 2013; p. 816]. This may result (as shown in Figure 6c) in selecting an ensemble member that provides the highest future projection for *Q*25, which is not intuitively representative of the change in the ensemble flood magnitudes.

Table 2 shows the descriptor estimates for each data source for the catchment in Figure 6b (the catchment with a large shared model discrepancy). MME strongly underestimates the scale parameter $\boldsymbol{\alpha}^*$ (log-transformed in the table), underestimates the location parameter ξ, and overestimates the shape parameter $\boldsymbol{\kappa}$ that collectively results in underestimation of T-year return events by the MME members (see (A3) and Figure 6b). Table 2 gives the expected value for the posterior distribution of the descriptor $\boldsymbol{\theta_0}$ (which is unaffected by the choice of K), as well its standard deviations when the shared model discrepancy spread does not change between the past and the future periods ($K = 0$), and when the spread largely increases ($K = 1$). The posterior 95% credible region for change in the descriptor θ_0 between the two periods excludes 0 for both $K = 0$ and $K = 1$, suggesting a change in flood frequency. However, change in T-year return flow is not supported by the posterior 95% credible intervals for $T = 100$ (as shown in Figure 6b), but is supported for smaller return periods (e.g., $T = 10$). This illustrates an interaction of the GEV parameters in the calculation of the T-year flow magnitudes (A3), so that different GEV parameter sets can results in similar estimates of extreme flows, and vice versa.

Figure 7 compares uncertainty in historic *Q*100 estimated from combined information with uncertainty in historic *Q*100 estimated from the observations and MME, and shows a relative uncertainty in the combined estimates with respect to the largest of uncertainties in observational and MME estimates. The uncertainty here is evaluated as the width of the 95% credible prediction intervals. Figure 7 demonstrates that combining information reduces uncertainty over the most uncertain information source, i.e., models or observations (shown by black squares); and that accounting appropriately for the model discrepancies (among the other factors) does not result in a prohibitively large uncertainty, and thus may be valuable for practical applications. Improvement relative to the observation source can be anticipated from equation (8), indicating that the combined estimate precision matrix $\boldsymbol{S}^{-1}$ is no less than the observation estimate precision matrix $\mathbf{J}_0{}^{-1}$ (a larger precision matrix corresponds to a better defined, or less uncertain, estimate). This

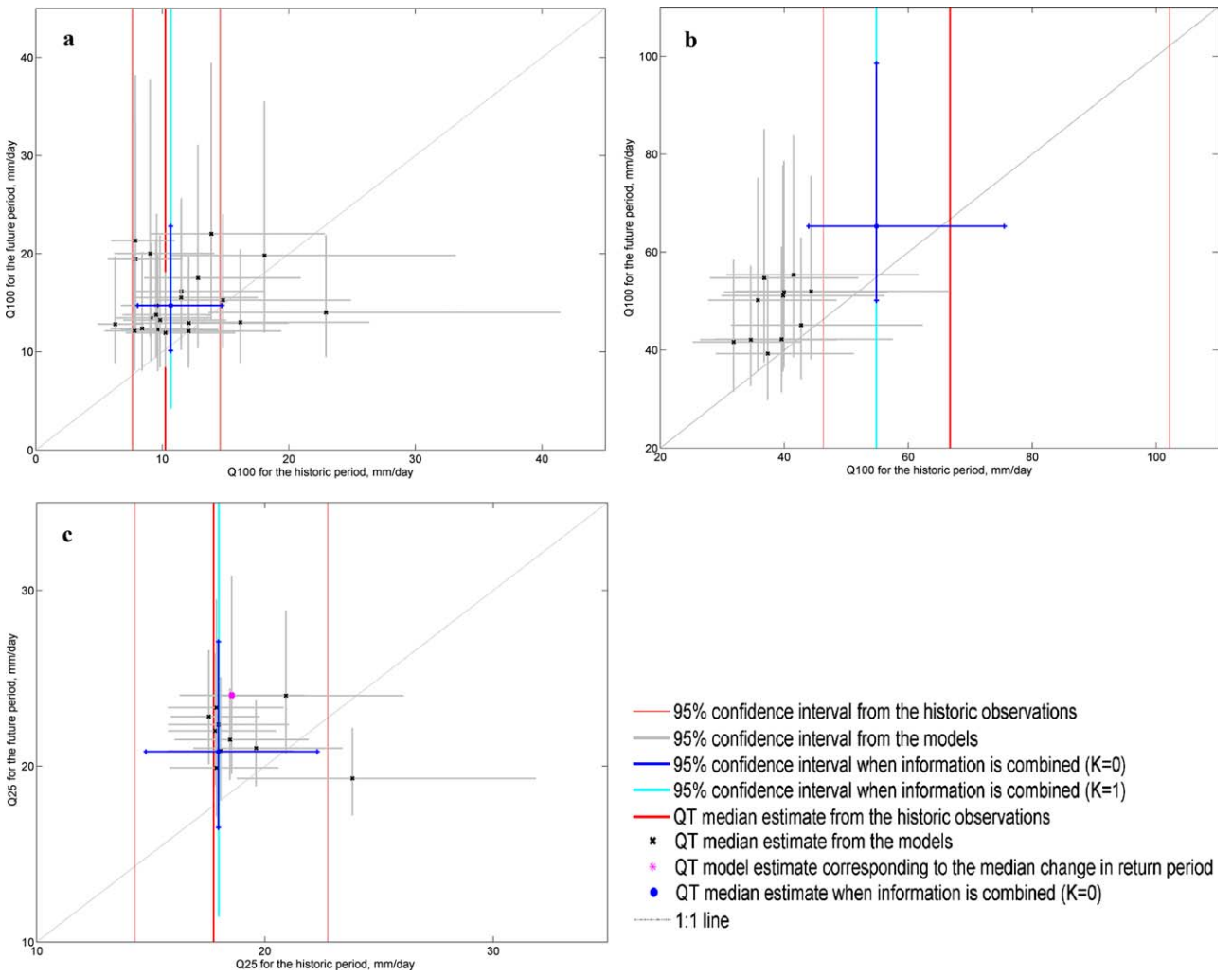


Figure 6. *QT* estimates given by different methods for (a) "*Great Stour at Horton*" station (South England), (b) "*Girvan at Robstone*" station (West Scotland), and (c) "*Fal at Tregony*" station (South West England). The thick vertical red lines denote *QT* median estimate and the thin red lines denote the corresponding 95% credible intervals derived from the historic observations. Black crosses indicate *QT* median estimate and grey bars indicate the corresponding 95% credible intervals derived from the model predictions. Blue dot is the *QT* median estimate and blue (cyan) lines are the corresponding 95% credible intervals derived from combined information when $K=0$ ($K=1$); blue and cyan lines overlap for the historic period. Magenta star is *Q*25 model estimate corresponding to the median change in return period.

relationship, however, does not guarantee that the combined estimate is better defined than an individual model estimate, due to additional shared multimodel discrepancy (characterized by $\mathbf{\Lambda}$). The combined estimate is superior to both observational and model estimates only when model discrepancy is small. This is illustrated in Figure 7 that connects the relative uncertainty to relative bias in (median) ensemble estimate of *Q*100 with respect to observations (shown by grey stars) that serves as a proxy for shared multimodel discrepancy. When *Q*100 is considerably underestimated by the ensemble (large negative values of relative bias), the MME provides overconfident estimates of *Q*100 due to low variability in simulated flow; and the combined uncertainty of *Q*100 is largely defined by the observational uncertainty of *Q*100 alone.

It is well established that the magnitude of the 100 year flood is highly uncertain when estimated using only a 30 year period of data [*Hosking and Wallis*, 1987]. It is, however, problematic to use longer periods of data due to the heightened risk of inadvertently capturing time trends in flood frequency. Therefore, the lower-flow but more-frequent 10 year flood event (*Q*10) is considered here to illustrate the impact of the reduced uncertainty on the flood frequency analysis. As in the previous analysis, the Bayesian analysis results in estimates of *Q*10 that are consistent with *Q*10 estimated from the observational evidence (for the historic period). Figure 8a shows projected changes in flood frequency based on combined information, for a 10 year flood. Mainly due to the reduced uncertainty in individual MME and observation-based 10 year

Table 2. Descriptor Estimates for Observations (Denoted as $\hat{\theta}_0$), MME (Denoted as $\hat{\theta}_i$, $i=1.m$), and Combined Information for "*Girvan at Robstone*" Station (West Scotland)[a]

	Historical			Future		
	$ln\ \xi$	$ln\ \alpha^*$	κ	$ln\ \xi$	$ln\ \alpha^*$	κ
Descriptors						
$\hat{\theta}_0$	2.70	1.20	−0.45			
$\hat{\theta}_1$	2.49	0.87	−0.30	2.59	0.86	−0.45
$\hat{\theta}_2$	2.50	0.93	−0.30	2.56	0.79	−0.37
$\hat{\theta}_3$	2.51	0.73	−0.36	2.53	0.88	−0.49
$\hat{\theta}_4$	2.50	0.71	−0.43	2.59	0.95	−0.47
$\hat{\theta}_5$	2.45	0.65	−0.37	2.62	1.04	−0.30
$\hat{\theta}_6$	2.48	0.66	−0.31	2.59	0.85	−0.37
$\hat{\theta}_7$	2.47	0.85	−0.37	2.51	0.90	−0.46
$\hat{\theta}_8$	2.51	0.78	−0.42	2.65	0.98	−0.36
$\hat{\theta}_9$	2.51	0.85	−0.36	2.56	0.91	−0.46
$\hat{\theta}_{10}$	2.38	0.77	−0.40	2.55	0.81	−0.40
$\hat{\theta}_{11}$	2.50	0.75	−0.44	2.64	1.08	−0.39
Consensus						
$\bar{\hat{\theta}}$	2.48	0.78	−0.37	2.58	0.91	−0.41
$\hat{\omega}$	−0.22	−0.42	0.09			
Posterior						
$E[\theta_0]$	2.70	1.16	−0.40	2.80	1.28	−0.43
std. dev.($K=0$)	0.02	0.08	0.06	0.03	0.11	0.08
std. dev. ($K=1$)	0.02	0.08	0.06	0.78	0.87	0.22

[a] $\bar{\hat{\theta}}$ is an average of descriptor estimates from MME, $\hat{\omega}$ is an estimate of the shared model discrepancy derived as $\bar{\hat{\theta}} - \hat{\theta}_0$ and available only for the historic period. Posterior for θ_0 is estimated using Gibbs sampling with two different values of *K*.

flood estimates, there are 85 out of 135 sites with zero change in *Q10* outside of the posterior 95% credible intervals, when model discrepancy spread is held constant between the past and the future (K = 0); and one site only, when model discrepancy spread is assumed to increase in the future (K = 1). Maximum sizes of the posterior credible intervals to exclude zero in *Q10* change (*K*=0) are shown on Figure 8b; there are further 14 out of 135 basins with zero change in *Q10* outside of the posterior 90% credible intervals.

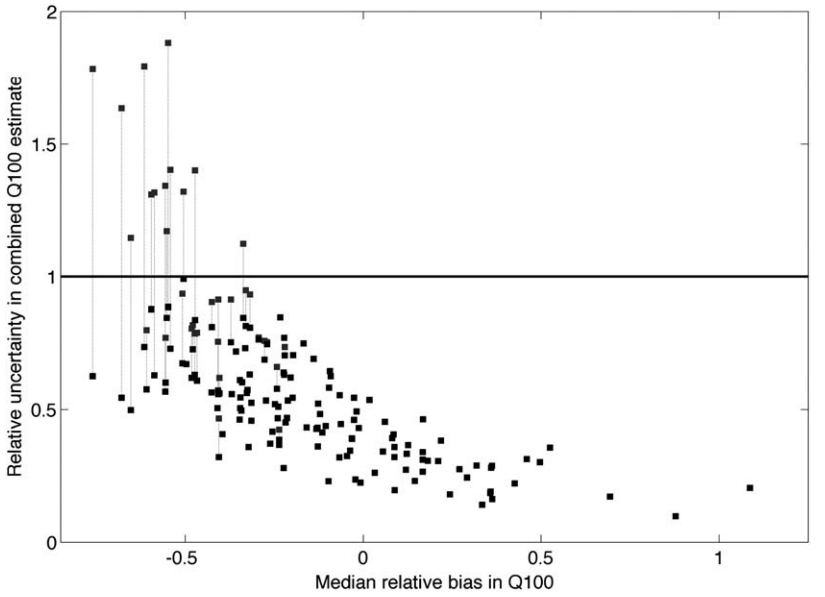


Figure 7. Relative uncertainty in historic *Q*100 estimated from combined information. Black squares denote relative uncertainty with respect to the largest uncertainty estimated from both observations and MME; and grey squares indicate relative uncertainty with respect to the largest uncertainty estimated from MME only. Grey-dotted vertical lines connect points representing the same basin, when black squares and grey stars do not overlap. Black horizontal line indicates the relative uncertainty of one.

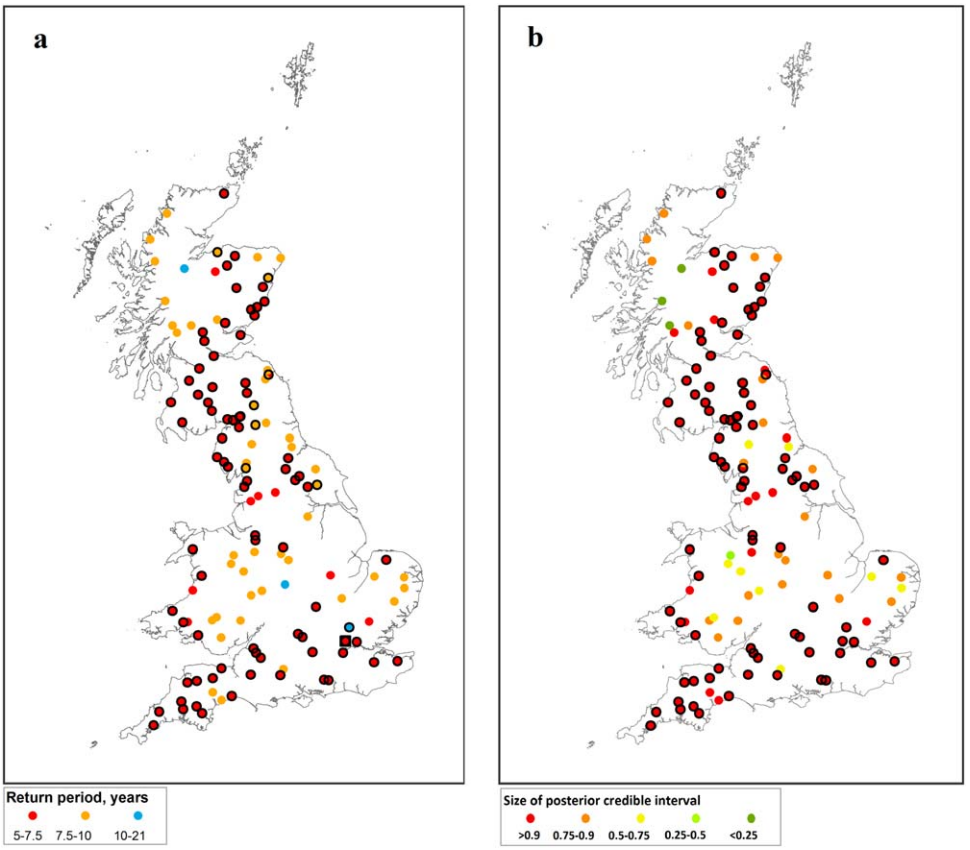


Figure 8. Projected change in *Q*10 based on MME combined information. (a) Return period (years) at the end of 21st century for discharge corresponding to the historic 10 year flood; (b) maximum size of posterior credible intervals (defined via probability mass contained in the interval) that exclude zero in *Q10* change (K = 0)—the larger the size the stronger the support for change in *Q10*. Basins with the posterior 95% credible intervals that exclude zero change in *Q10* are denoted by black circles (K = 0) and black squares (K = 1).

6. Conclusions

Current protocol to handle multimodel flood projections focuses on the median change as well as model consistency [*Arnell and Gosling*, 2014; *Hirabayashi et al*., 2013; *Ward et al*., 2014]. In this paper, it is demonstrated that the standard practice of combining outputs from a multimodel ensemble provides widely spread, uncertain and possibly biased estimates of extreme flow events. This, first, raises questions regarding the reliability of "baseline" conditions (to assess future changes against) that are provided by a multimodel ensemble. And second, it prevents the extraction of a meaningful signal from the widely distributed estimates.

The study examines the potential of a hierarchical Bayesian approach to combine information from multiple models for the purposes of flood frequency change analysis. The approach explicitly considers shared multimodel discrepancy alongside the probabilistic nature of flood estimates, and treats the selected models as a sample from the complete (but unobservable) set of models. The AMS/GEV descriptors are proposed for selection as statistical summaries of the observed and MME flow time series, thus allowing efficient and consistent combination of information regarding the entire flood frequency curve.

This information-combining method hinges on a number of assumptions, and is a subject to various limitations, as described in section 3. The proposed method is general, however, allowing the assumptions to be readily and transparently altered to reflect other beliefs and to evaluate the sensitivity of outcomes. Representation of future model discrepancy is one of the most impactful, but (based on the data available) not a priori knowable, assumptions (see discussion in section 3.2.3). Although its explicit representation and its differentiation between the past and the future may increase uncertainty, the technique represents an advance in MME processing where it is typically (and erroneously) assumed that model performance is

similar between the past and the future. The latter is comparable to the assumption that characteristics of the model discrepancy remain the same between the two periods. When applied to a flood frequency change problem, the information-combining method exhibits the following attractive properties:

1. It provides "baseline" conditions that are consistent with the observed evidence, and therefore are appropriate for comparison with future changes.
2. The method reduces uncertainty in flood estimates, as compared to the estimates of uncertainty derived from both MME and observations.
3. For low values of shared model discrepancy, the approach allows extraction of information from multiple small and weak effects by bringing together predictions that individually lack statistical power, but collectively provide meaningful information.
4. For high values of shared model discrepancy, the method shows that the traditionally used model consistency cannot be regarded as indicative of change in flood frequency.
5. The method allows an explicit assessment of the influence of the (model performance) stationarity assumption on the resulting predictions.

The approach indicates that the current reliance on the median change in flood estimates may be misleading. Moreover, the method suggests that reducing shared model discrepancy is necessary to further reduce uncertainty in flood frequency analysis. A large value of model discrepancy inflates uncertainties associated with the MME flood estimates, and prohibits distinguishing a useful signal out of the noise around the combined estimates. Efficient reduction of uncertainty is especially relevant while evaluating changes in extreme event magnitudes (e.g., 100 year event) that are estimated from relatively short periods (e.g., 30 years). Finally, an explicit representation of model discrepancy and uncertainty in flood magnitude may have a significant effect on impact/attribution assessment studies. Failure to credibly capture the uncertainty in future projections (i.e., lacking to account for model biases, inter-model dependences, etc.), and a subsequent propagation of this through models for vulnerability and loss may arrive at wrong risk assessments.

Appendix A: Annual Maximum Series/Generalized Extreme Value Model

In the AMS/GEV approach, annual peak flow maxima are assumed to follow a generalized extreme value distribution [*Madsen et al.*, 1997] with the following cumulative distribution

$$F(q)=\begin{cases}\exp\left(-\exp\left(-\dfrac{q-\xi}{\alpha^*}\right)\right), & \kappa=0\\ \exp\left(-\left(1-\kappa\dfrac{q-\xi}{\alpha^*}\right)^{1/\kappa}\right), & \kappa\neq 0\end{cases} \tag{A1}$$

The three parameters in (A1) (ξ, α^*, and κ) are related to PDS/GP model parameters λ, q_0, α, and κ, such that the shape parameter κ is identical between the models, and

$$\xi=\begin{cases}q_0+\alpha\ \ln(\lambda), & \kappa=0\\ q_0+\dfrac{\alpha}{\kappa}\ (1-\lambda^{-\kappa}), & \kappa\neq 0\end{cases} \tag{A2}$$

$$\alpha^*=\alpha\ \lambda^{-\kappa}$$

The mean annual number of threshold exceedances λ defines the flow threshold q_0 in PDS/GP. In the UK, five events, on average, per year are included; with standard rules employed to ensure that extracted flood peaks are independent events (by imposing a minimum separation time period of three times time-to-peak, and specifying that the flow between two peaks must drop to at least two thirds of the higher peak [*Robson*, 1999, p.276]).

While no explicit solution exists for the maximum likelihood $\hat{\xi}$, $\hat{\alpha}^*$, and $\hat{\kappa}$ estimator of ξ, α^*, and κ, a numerical procedure is typically applied based on the Newton-Raphson iteration [*Hosking and Wallis*, 1987; *Madsen et al.*, 1997]. The asymptotic covariance matrix for the ML estimator is obtained by inverting the Fisher information matrix given by *Prescott and Walden* [1983].

The T-year return period event magnitude for AMS is defined as the $1-1/T$ quantile of the GEV distribution (A1), so that

$$q_T = \begin{cases} \xi + \alpha^* y_T, & \kappa = 0 \\ \xi + \frac{\alpha^*}{\kappa}(1 - \exp(-\kappa y_T)), & \kappa \neq 0 \end{cases} \quad \text{(A3)}$$

where $y_T = -ln\left(-ln\left(1-\frac{1}{T}\right)\right)$ is the Gumbel reduced variate.

The quantile estimator $\hat{q}_T$ is defined by substituting estimators $\hat{\xi}$, $\hat{\alpha^*}$, and $\hat{\kappa}$ in the above equation (A3). The credible intervals for q_T can be estimated using Monte-Carlo sampling of parameters of ξ, α^*, and κ from their joint distribution and calculating the corresponding flow magnitudes from (A3). The same procedure can be employed to test a null-hypothesis of "no change" between T-year flow estimates for different time periods, including estimation of a corresponding p-value (a two-sided test is employed in the work).

When both the location parameter ξ and scale parameter α^* are log-transformed (shape parameter κ is kept unchanged), the maximum likelihood parameter estimate equals ln $\hat{\xi}$, ln $\hat{\alpha^*}, \hat{\kappa}$. Further, the Fisher information matrix for the transformed parameters can be calculated as $\mathbf{J'*I_F*J}$, where $\mathbf{I_F}$ is the Fisher information matrix for the original parameters ξ, α^*, and κ [*Prescott and Walden*, 1983] and $\mathbf{J}$ is the Jacobian matrix with zero off-diagonal elements and ξ, α^*, and 1 on the main diagonal. The inverse of the Fisher information matrix provides the covariance matrix for the asymptotically normal distribution of the three parameters.

Appendix B: Selection of Initial Values of $\boldsymbol{C}, \boldsymbol{\Lambda}$

The Gibbs sampling described in section 2.2.2 requires starting values of $\boldsymbol{C}$, $\boldsymbol{\Lambda}$ to initiate the sampling. The estimates $\hat{\boldsymbol{C}}$, $\hat{\boldsymbol{\Lambda}}$ based on $\left\{\hat{\boldsymbol{\theta}}_i\right\}_{i=0}^{m}$ provided by *Chandler* [2013] are used for the purposes and are described below.

The propensity to deviate around the model consensus is described by the covariance matrix $\boldsymbol{C}$ and can be estimated as

$$\hat{\boldsymbol{C}} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\hat{\boldsymbol{\theta}}_i - \bar{\hat{\boldsymbol{\theta}}}\right)\left(\hat{\boldsymbol{\theta}}_i - \bar{\hat{\boldsymbol{\theta}}}\right)' \quad \text{(B1)}$$

m represents the number of models considered to be equally credible and $\bar{\hat{\boldsymbol{\theta}}} = \frac{1}{m}\sum_{i=1}^{m} \hat{\boldsymbol{\theta}}_i$ represents the mean of the estimated model descriptors.

$\boldsymbol{\Lambda_h} = E\left(\boldsymbol{\omega_h}\boldsymbol{\omega_h'}\right)$ is the part of the covariance matrix $\boldsymbol{\Lambda}$ corresponding to the historic period, and can be estimated as $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{h}} = \hat{\boldsymbol{\omega}}_{\boldsymbol{h}}\hat{\boldsymbol{\omega}}_{\boldsymbol{h}}'$, where $\hat{\boldsymbol{\omega}}_{\boldsymbol{h}} = \bar{\hat{\boldsymbol{\theta}}}_{\boldsymbol{h}} - \hat{\boldsymbol{\theta}}_{0\boldsymbol{h}}$. As no data are available to estimate the remainder of the matrix $\boldsymbol{\Lambda}$, the assumption about the MME shared discrepancy change in the future characterized by a constant $K > 0$ is used to provide an estimate of $\boldsymbol{\Lambda}$ as

$$\hat{\boldsymbol{\Lambda}} = \begin{pmatrix} \hat{\boldsymbol{\Lambda}}_h & \hat{\boldsymbol{\Lambda}}_h \\ \hat{\boldsymbol{\Lambda}}_h & (1+K)\hat{\boldsymbol{\Lambda}}_h \end{pmatrix} \quad \text{(B2)}$$

The assumption is that the shared discrepancy would change in the future as $\boldsymbol{\omega}_f = \boldsymbol{\omega}_h + \boldsymbol{\eta}$, where random variable $\boldsymbol{\eta}$ is independent of $\boldsymbol{\omega}_h$, has a mean of zero and a covariance matrix $K\boldsymbol{\Lambda}_h$ for some constant $K > 0$.

Appendix C: Asymptotic Normality Approximation

The maximum likelihood parameter estimates $\left(ln\ \hat{\xi},\ ln\ \hat{\alpha}^*,\ \hat{\kappa}\right)$ for GEV distribution are asymptotically normal and unbiased (a sample consists of independent flow peaks) [*Millar*, 2011, chap. 12.2]. An accuracy of this assumption when the sample size is finite and equals 150 (representing 30 years of flow data, with five flow threshold exceedances per year, on average) is examined here. A computer experiment is run to calculate bias and sample variance of the descriptor estimates, as well as to perform a number of statistical tests on normality. Simulations are performed with the shape parameter of the GEV distribution taking values $\kappa = -1,\ -0.7,\ -0.4,\ -0.1$. The considered values of the shape parameter are representative of the parameter ranges obtained for observations and simulations of flow used in the study that vary between -1.15 and -0.05. The maximum likelihood method of estimation is invariant under linear transformation of the data, so that without loss of generality $\xi = 1$ and $\alpha^* = 1$ ($ln\ \xi = 0$, $ln\ \alpha^* = 0$). For each value of κ, 1000 random sets of 150 samples are generated from the GEV distribution; and parameters $ln\ \xi$, $ln\ \alpha^*$, and κ are estimated for each set of samples by the maximum likelihood method. Table C1 shows that the values of

Table C1. Values of the Asymptotic (ASD) and Sample (SSD) Standard Deviations, and Bias in the Parameter Estimates for Each Value of the Shape Parameter κ

	$ln\ \xi$				$ln\ \alpha^*$				κ			
κ	−1	−0.7	−0.4	−0.1	−1	−0.7	−0.4	−0.1	−1	−0.7	−0.4	−0.1
ASD	0.09	0.09	0.09	0.09	0.12	0.10	0.08	0.07	0.10	0.09	0.07	0.06
SSD	0.09	0.09	0.09	0.09	0.12	0.10	0.08	0.07	0.11	0.09	0.08	0.06
Bias	−0.01	0.00	0.00	0.00	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	0.00

the asymptotic standard deviation are approximated by the corresponding sample standard deviations, and bias in the parameter estimates is small for each value of κ. Further the Lliefors test [*Lilliefors*, 1967] on univariate normality for each of the variables and for each value of κ cannot be rejected at 0.95 confidence level. However, the Mardia tests for multivariate normality based on multivariate skewness and kurtosis measures [*Mardia*, 1970] indicated deviations from normality for low values of κ (0.95 confidence level). In particular, the normality was rejected for $\kappa = -1$ and $\kappa = -0.7$ based on the skewness measure, and for $\kappa = -1$ based on the kurtosis measure.

Appendix D: Necessary Conditions for Independence of Model Discrepancies

As represented by equation (2), model discrepancies are assumed to be normally distributed for each catchment:

$$\boldsymbol{\delta_i^k} \sim N\left(\boldsymbol{\omega^k}, \boldsymbol{C^k}\right),\ i=1,m; k=1,n \tag{D1}$$

index k indicates each catchment's identifier and, for simplicity, the covariance matrix is considered constant for each catchment, i.e., $\boldsymbol{C_1^k} = \ldots = \boldsymbol{C_m^k} = \boldsymbol{C^k}$ (it can be shown that the further derivation can be generalized for a case with distinct matrices $\boldsymbol{C}_i$). The model discrepancies are assumed to be independent; and cross-covariances are considered below to establish a necessary condition for the assumption.

When model discrepancies come from the same distribution for all catchments (i.e., $\boldsymbol{\omega^k}$ and $\boldsymbol{C^k}$ are constant for all catchments), the sample cross-covariance based on n draws $\left\{\boldsymbol{\delta_i^k}, \boldsymbol{\delta_j^k}\right\}_{k=1}^{n}$, $i \neq j$ can be computed and the significance hypotheses can be tested. In a general case, the n draws cannot be used directly to estimate the cross-covariances, as each draw $\left(\boldsymbol{\delta_i^k}, \boldsymbol{\delta_j^k}\right)$ comes from a different (catchment-specific) distribution (D1). Instead, the normalized values of model discrepancies $\boldsymbol{\gamma_i^k} = \boldsymbol{G^k} * \left(\boldsymbol{\delta_i^k} - \boldsymbol{\omega^k}\right)$ are considered, where $\boldsymbol{C^k} = \boldsymbol{G^{k'}} * \boldsymbol{G^k}$. so that all standardized discrepancies come from the same standard normal distribution $N(0, \boldsymbol{I})$. Then, the cross-covariance can be expressed as $cov\left(\boldsymbol{\gamma_i^k}, \boldsymbol{\gamma_j^k}\right) = \boldsymbol{G^k} * cov\left(\boldsymbol{\delta_i^k}, \boldsymbol{\delta_j^k}\right) * \boldsymbol{G^{k'}}$, $i \neq j$. For a positive-definite covariance matrix $\boldsymbol{C^k}$, zero cross-covariance between the normalized model discrepancies is equivalent to zero cross-covariance between the original model discrepancies.

The cross-covariance $cov\left(\boldsymbol{\gamma_i^k}, \boldsymbol{\gamma_j^k}\right)$, $i \neq j$ can be estimated based on n draws $\left\{\boldsymbol{\gamma_i^k}, \boldsymbol{\gamma_j^k}\right\}_{k=1}^{n}$, $i \neq j$, as the draws now come from the same standard normal distribution, and a family of hypotheses of pair-wise cross-covariance significance can be tested. Family-wise error rate (type I error) can be defined using either Bonferroni correction [*Dunn*, 1961], or Holm's procedure [*Holm*, 1979].

Acknowledgments

The research was supported by the *UK Natural Environment Research Council's programs: Changing Water Cycle* (grant *NE/I022558/1*) and *Probability, Uncertainty and Risk in the Environment* (grant *NE/J016578/1*). The Future Flows *Hydrology* data and flow observations were accessed via the National River Flow Archive (Centre for Ecology and Hydrology): http://www.ceh.ac.uk/data/nrfa/data/search.html. Future Flows Hydrology is also available via: http://dx.doi.org/10.5285/f3723162-4fed-4d9d-92c6-dd17412fa37b following user registration. The software is available on request from the author.

References

Arnell, N. W., and S. N. Gosling (2014), The impacts of climate change on river flood risk at the global scale, *Clim. Change*, *134*(3), 1–15.

Arnell, N. W., et al. (2013), A global assessment of the effects of climate policy on the impacts of climate change, *Nat. Clim. Change*, *3*(5), 512–519.

Beven, K. J. (2011), *Rainfall-Runoff Modelling: The Primer*, John Wiley, Hoboken, N. J.

Carlin, B., and T. Louis (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., 419 pp., Chapman and Hall, London, U. K.

Casella, G., and R. Berger (1990), *Statistical Inference*, Wandsworth Inc., Belmont, Calif.

Chandler, R. (2013), Exploiting strength, discounting weakness: Combining information from multiple climate simulators, *Philos. Trans. R. Soc. A*, *371*(1991), 20120388.

Choulakian, V., and M. Stephens (2001), Goodness-of-fit tests for the generalized Pareto distribution, *Technometrics*, *43*(4), 478–484.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer, London, U. K.
Crooks, S., and P. Naden (2007), CLASSIC: A semi-distributed rainfall-runoff modelling system, *Hydrol. Earth Syst. Sci.*, *11*(1), 516–531.
Cunnane, C. (1979), A note on the Poisson assumption in partial duration series models, *Water Resour. Res.*, *15*(2), 489–494.
Dunn, O. J. (1961), Multiple comparisons among means, *J. Am. Stat. Assoc.*, *56*(293), 52–64.
Furrer, R., S. Sain, D. Nychka, and G. Meehl (2007), Multivariate Bayesian analysis of atmosphere–ocean general circulation models, *Environ. Ecol. Stat.*, *14*(3), 249–266.
Geyer, C. J. (1992), Practical Markov chain Monte Carlo, *Stat. Sci.*, *7*(4), 473–483.
Giorgi, F., and L. Mearns (2002), Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging"(REA) method, *J. Clim.*, *15*(10), 1141–1158.
Greene, A., L. Goddard, and U. Lall (2006), Probabilistic multimodel regional temperature change projections, *J. Clim.*, *19*(17), 4326–4343.
Griffiths, J., A. Young, and V. Keller (2006), Continuous Estimation of River Flows (CERF), *Tech. Rep. Task 1.3*, Model scheme for representing rainfall interception and soil moisture, Centre for Ecology and Hydrology (CEH), Wallingford, U. K.
Hallegatte, S., C. Green, R. J. Nicholls, and J. Corfee-Morlot (2013), Future flood losses in major coastal cities, *Nat. Clim. Change*, *3*(9), 802–806.
Haxton, T., et al. (2012), *Future flows hydrology data*, NERC Environmental Information Data Centre, doi:10.5285/f3723162-4fed-4d9d-92c6-dd17412fa37b.
Hirabayashi, Y., R. Mahendran, S. Koirala, L. Konoshima, D. Yamazaki, S. Watanabe, H. Kim, and S. Kanae (2013), Global flood risk under climate change, *Nat. Clim. Change*, *3*(9), 816–821.
Hoeting, J. A., D. M. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial (with discussion), *Stat. Sci.*, *14*, 382–401.
Holm, S. (1979), A simple sequentially rejective multiple test procedure, *Scand. J. Stat.*, *6*(2), 65–70.
Hosking, J., and J. Wallis (1987), Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics*, *29*(3), 339–349.
Institute of Hydrology (1999), *Flood Estimation Handbook*, Wallingford, U. K.
Kay, A. L., and D. A. Jones (2012), Transient changes in flood frequency and timing in Britain under potential projections of climate change, *Int. J. Climatol.*, *32*(4), 489–502.
Kennedy, M., and A. O'Hagan (2001), Bayesian calibration of computer models, *J. R. Stat. Soc., Ser. B*, *63*(3), 425–464, doi:10.1111/1467-9868.00294.
Knutti, R. (2010), The end of model democracy?, *Clim. Change*, *102*(3-4), 395–404.
Kuczera, G. (1982), Combining site-specific and regional information: An empirical Bayes approach, *Water Resour. Res.*, *18*(2), 306–314.
Leith, N., and R. Chandler (2010), A framework for interpreting climate model outputs, *J. R. Stat. Soc. Ser. C*, *59*(2), 279–296.
Lilliefors, H. W. (1967), On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *J. Am. Stat. Assoc.*, *62*(318), 399–402.
Lindley, D. V., and A. F. Smith (1972), Bayes estimates for the linear model, *J. R. Stat. Soc. Ser. B*, *34*(1), 1–41.
Madsen, H., P. F. Rasmussen, and D. Rosbjerg (1997), Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events .1. At-site modeling, *Water Resour. Res.*, *33*(4), 747–757.
Mardia, K. V. (1970), Measures of multivariate skewness and kurtosis with applications, *Biometrika*, *57*(3), 519–530.
McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener (2005), Ensemble predictions of runoff in ungauged catchments, *Water Resour. Res.*, *41*, W12434, doi:10.1029/2005WR004289.
Millar, R. (2011), *Maximum Likelihood Estimation and Inference With Examples in R, SAS and ADMB*, John Wiley, Hoboken, N. J.
Moore, R. (2007), The PDM rainfall-runoff model, *Hydrol. Earth Syst. Sci. Discuss.*, *11*(1), 483–499.
Pawitan, Y. (2001), *In All Likelihood*, Clarendon, Oxford.
Penner, J., M. Andreae, H. Annegarn, L. Barrie, J. Feichter, D. Hegg, A. Jayaraman, R. Leaitch, D. Murphy, and J. Nganga (2001), *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge Univ. Press, Cambridge.
Prescott, P., and A. Walden (1983), Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples, *J. Stat. Comput. Simul.*, *16*(3-4), 241–250.
Prudhomme, C., S. Crooks, C. Jackson, J. Kelvin, and A. Young (2012a), *Future Flows and Groundwater Levels*, 118 pp., Centre for Ecology and Hydrology (CEH), Wallingford, U. K.
Prudhomme, C., S. Dadson, D. Morris, J. Williamson, G. Goodsell, S. Crooks, L. Boelee, H. Davies, G. Buys, and T. Lafon (2012b), Future Flows Climate: An ensemble of 1-km climate change projections for hydrological application in Great Britain, *Earth Syst. Sci. Data Discuss.*, *5*(1), 475–490.
Prudhomme, C., et al. (2013), Future flows hydrology: An ensemble of daily river flow and monthly groundwater levels for use for climate change impact assessment across Great Britain, *Earth Syst. Sci. Data*, *5*(1), 101–107.
Raftery A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*, 1155–1174.
Räisänen, J., and T. Palmer (2001), A probability and decision-model analysis of a multimodel ensemble of climate change simulations, *J. Clim.*, *14*(15), 3212–3226.
Renard B (2011), A Bayesian hierarchical approach to regional frequency analysis, *Water Resour Res.*, *47*, W11513, doi:10.1029/2010WR010089.
Robson, A. (1999), Statistical procedures for flood frequency estimation, in *Flood Estimation Handbook*, Inst. of Hydrol., Wallingford, U. K.
Smith, T., L. Marshall, and A. Sharma (2014), Predicting hydrologic response through a hierarchical catchment knowledgebase: A Bayes empirical Bayes approach, *Water Resour. Res.*, *50*, 1189–1204, doi:10.1002/2013WR015079.
Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic climate projections, *Philos. Trans. R. Soc. Ser. A*, *365*(1857), 2053–2075.
Tebaldi, C., R. Smith, D. Nychka, and L. Mearns (2005), Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles, *J. Clim.*, *18*(10), 1524–1540.
Ward, P., S. Eisner, M. Flörke, M. Dettinger, and M. Kummu (2014), Annual flood sensitivities to El Niño–Southern Oscillation at the global scale, *Hydrol. Earth Syst. Sci.*, *18*(1), 47–66.
Wu, W., J. S. Clark, and J. M. Vose (2010), Assimilating multi-source uncertainties of a parsimonious conceptual hydrological model using hierarchical Bayesian modeling, *J. Hydrol.*, *394*(3-4), 436–446, doi:10.1016/j.jhydrol.2010.09.017.