Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: https://www.tandfonline.com/loi/thsj20

Model averaging *versus* model selection: estimating design floods with uncertain river flow data

Kenechukwu Okoli, Korbinian Breinl, Luigia Brandimarte, Anna Botto, Elena Volpi & Giuliano Di Baldassarre

To cite this article: Kenechukwu Okoli, Korbinian Breinl, Luigia Brandimarte, Anna Botto, Elena Volpi & Giuliano Di Baldassarre (2018) Model averaging *versus* model selection: estimating design floods with uncertain river flow data, Hydrological Sciences Journal, 63:13-14, 1913-1926, DOI: 10.1080/02626667.2018.1546389

To link to this article: https://doi.org/10.1080/02626667.2018.1546389

HYDROLOGICAL SCIENCES JOURNAL
2018, VOL. 63, NOS. 13–14, 1913–1926
https://doi.org/10.1080/02626667.2018.1546389

Model averaging *versus* model selection: estimating design floods with uncertain river flow data

Kenechukwu Okoli [a,b], Korbinian Breinl[a,b], Luigia Brandimarte[c], Anna Botto[d], Elena Volpi[e] and Giuliano Di Baldassarre[a,b]

[a]Department of Earth Sciences, Uppsala University, Uppsala, Sweden; [b]Centre of Natural Hazards and Disaster Science (CNDS), Uppsala, Sweden; [c]Department of Sustainable Development, Environmental Science and Engineering, Royal Institute of Technology, Stockholm, Sweden; [d]Department of Civil, Environmental and Architectural Engineering, University di Padova, Padova, Italy; [e]Department of "Scienze dell' Ingegneria Civile", University of "Roma Tre", Rome, Italy

ABSTRACT

This study compares model averaging and model selection methods to estimate design floods, while accounting for the observation error that is typically associated with annual maximum flow data. Model selection refers to methods where a single distribution function is chosen based on prior knowledge or by means of selection criteria. Model averaging refers to methods where the results of multiple distribution functions are combined. Numerical experiments were carried out by generating synthetic data using the Wakeby distribution function as the parent distribution. For this study, comparisons were made in terms of relative error and root mean square error (RMSE) referring to the 1-in-100 year flood. The experiments show that model averaging and model selection methods lead to similar results, especially when short samples are drawn from a highly asymmetric parent. Also, taking an arithmetic average of all design flood estimates gives estimated variances similar to those obtained with more complex weighted model averaging.

ARTICLE HISTORY
Received 9 March 2018
Accepted 13 September 2018

EDITOR
A. Castellarin

ASSOCIATE EDITOR
S. Vorogushyn

KEYWORDS
model averaging; model selection; design flood; Akaike information criterion

1 Introduction

A common task in applied hydrology is the estimation of the design flood, i.e. a value of river discharge corresponding to a given exceedence probability that is often expressed as a return period in years. Flood risk assessment, floodplain mapping and the design of hydraulic structures are a few examples of applications where estimates of design floods are required. Two common approaches for estimating a design flood are either rainfall–runoff modelling (e.g. Moretti and Montanari 2008, Beven 2012, Breinl 2016) or the fitting of a probability distribution function to a record of annual maximum or peak-over-threshold flows (Viglione *et al.* 2013, Yan and Moradkhani 2016). The latter approach, which is the focus of this paper, has been referred to in the literature as the "standard approach" to the frequency analysis of floods (Klemeš 1993). The standard approach is affected by various sources of uncertainty, including: the choice of the sample technique (peak-over-threshold or annual maximum flows), a limited sample size, the selection of a suitable probability distribution function, the method of parameter estimation for the chosen distribution function, and errors in the observed annual peak flows derived from a rating curve (Sonuga 1972; Laio *et al.* 2009; Di Baldassarre *et al.* 2012)

It is common practice in any form of modelling or statistical analysis (including flood frequency analysis) to consider a range of models as possible representations of the observed reality. A single model is usually selected based on different criteria, such as (a) goodness-of-fit statistics, e.g. by using the chi-squared (χ^2) test; (b) prior selection of a distribution function as a result of what Chamberlain (1965) referred to as "parental affection" towards a given model; or (c) standardization, such as the log-Pearson Type III distribution used for flood frequency analysis in the USA (US Water Resources Council 1982). In the field of flood frequency analysis, the selection of a single best distribution function represents an implicit assumption that the selected model can adequately describe the frequency of observed and future floods, including the extreme ones. This

CONTACT Kenechukwu Okoli kenechukwu.okoli@geo.uu.se
Supplementary data for this article can be accessed here

assumption departs from the understanding that low and ordinary floods (which usually make up the annual peak flow record) are dominated by different processes compared to extreme floods, which are often the main focus in flood risk management. Therefore, the selection of a single distribution model, which is valid for the whole range of flows, may lead to uncertainty in the design flood estimates. Also, smaller floods are known to influence the smoothing and extrapolation of the largest discharges in the record and in turn may lead to uncertain estimates of the design flood (Klemeš 1986).

Experience from evaluating probability plots of discharge records shows that different distribution functions commonly used in flood frequency analysis give similar fits to data. The reason for this is that a majority of the parametric models used in flood frequency analysis have two or three parameters and are built to preserve the mean and variance of the calibration data (Koutsoyiannis 2004). Hence, there is always a model choice uncertainty when a particular distribution function is selected for estimation purposes. Within the hydrological modelling community, the phenomenon where different models give a similar fit to data has been referred to as "equifinality" (Beven 1993, 2006). In that context, Beven and Binley (1992) developed the generalized likelihood uncertainty estimation (GLUE) to support ensemble predictions of a model output variable. Just like GLUE, other techniques on how to combine estimates from different model structures (and parameter sets) using weights were developed and are generally referred to as model averaging (Hoeting *et al.* 1999, Burnham and Anderson 2002). Bayesian model averaging (BMA), for example, is used extensively in hydrogeology (Tsai and Li 2008, Ye *et al.* 2010, Foglia *et al.* 2013) to quantify predictive uncertainty when diverse conceptual models are used for recharge and/or hydraulic conductivity estimates. The reader is referred to Schöniger *et al.* (2014) and Volpi *et al.* (2017) for a detailed discussion on Bayesian model evidence (BME) for hydrological applications, especially when the problem of model selection is addressed using BMA.

Uncertainties present in the record of annual maximum flows are often neglected. For example, flood discharges, which are considerably larger than the directly measured discharges, and are therefore derived by extrapolating the rating curve, are subject to major errors, which may in turn impact the estimate of sample statistics such as the skewness (Potter and Walker 1985). Kuczera (1996, 1992) showed that significant uncertainty in the design flood estimate is often caused by errors in discharge data derived from a rating curve. Other studies made use of numerical approaches based on hydraulic modelling or Monte Carlo sampling to quantify the uncertainty in flow data due to rating curve errors (Di Baldassarre and Montanari 2009, Westerberg and McMillan 2015). According to their findings, the uncertainty present in derived discharges may add up to 30% or more.

Given this background, in this study we account for two sources of uncertainty that can significantly affect the design flood estimate: errors in the river flow data, i.e. annual maximum flows derived from a rating curve, and the choice of distribution function.

We compare model selection (denoted here as MS) with two different types of model averaging: arithmetic model averaging (denoted as MM) and weighted model averaging (MA). Model selection refers to a case where a single best distribution function is selected based on a selection criterion; MM describes the averaging by applying the arithmetic average of all estimated design floods; and MA refers to the application of a weighted average of design flood estimates from different probability functions (with weights based on a selection criterion). We used the Akaike information criterion (AIC) as a selection criterion for both MS and MA. The study was conducted in a simulation framework using the Wakeby distribution as the parent model for generating synthetic annual maximum flows of different sample sizes.

The aims of our study are as follows: (a) to simulate the systemic uncertainty in the real-world scenario; that is, in the real world, the parent distribution is unknown and likely more complex than the simpler distribution functions used for fitting and estimation purposes; (b) to make a systematic assessment and comparison of the performance of alternative methods for estimating design floods (MS *vs* MA *vs* MM) and; (c) to analyse the effect of flood data errors.

The comparison is based on the relative errors across the three techniques and the respective candidate distribution functions. The 1-in-100 year flood, i.e. the discharge value corresponding to a return period of 100 years (hereafter 100-year flood), is selected as the design flood of interest due to its wide use as a design standard in flood risk management (Brandimarte and Di Baldassarre 2012). For example, the current policy in the USA for flood defence design refers to the 100-year flood (Commission on Geosciences Environment and Resources). The analyses presented in this study are built on the assumption of stationarity, which has been widely discussed in hydrology (e.g.

Milly *et al.* 2008, Montanari and Koutsoyiannis 2014, Serinaldi and Kilsby 2015, Luke *et al.* 2017) and is not further discussed here.

2 Methods

The problem of the MS and MA methods is formulated as follows: a record of a random variable X is available and sampled from an unknown parent distribution $g(x)$. The samples are arranged in ascending order $x_1 \leq x_2 \ldots \leq x_N$. A set of probability distribution functions, whose general mathematical form can be written as $f(x_i|\theta)$ with θ as model parameter, are specified as potential candidates for design flood estimation. To implement the MS and MA techniques, we used the Akaike selection criterion, which is a commonly used method for model comparison in hydrology (e.g. Mutua 1994, Strupczewski *et al.* 2001). MS techniques based on information theory require the estimation of a measure of discrepancy, or amount of information loss, when a model is used to approximate the full reality (Linhart and Zucchini 1986). Akaike (1973) formulated the AIC as an estimator of information loss or gain when a model is fitted to data. The AIC index (I) is expressed as:

$$I = -2L\left(\hat{\theta}\right) + 2K \tag{1}$$

where K is the number of parameters, $L\left(\hat{\theta}\right)$ is the numerical value for the log-likelihood at its maximum point for the selected model and $\hat{\theta}$ is the maximum likelihood estimator of model parameters. For a detailed mathematical description, the reader is referred to Linhart and Zucchini (1986) and Burnham and Anderson (2002). A heuristic interpretation of Equation (1) suggests that the first term decreases with an increase in the second term. This shows a distinct property of the AIC in finding a trade-off between bias and variance of an estimator. The AIC is relative and – since the "truth" is not known – the relationship between AIC values of respective models indicates the model of choice, not AIC values *per se* (Burnham and Anderson 2002).

An extension of the AIC, denoted AIC_c, was proposed by Sugiura (1978) to correct for bias due to a short sample size n, the AIC_c index (I_c) is expressed as:

$$I_c = -2L\left(\hat{\theta}\right) + 2K + \frac{2K(K+1)}{n-K-1} \tag{2}$$

Burnham and Anderson (2002) suggested using AIC_c when the ratio n/K is small (e.g. <40), and the original formulation when the ratio is sufficiently large. We considered both AIC and AIC_c in this study. The AIC_c was used for the short samples, which in this application is a sample size of 30 years, and AIC was used for large sample sizes generated in our numerical experiments, as detailed in the following sections. In principle, the model with a minimum AIC (or AIC_c) value was considered the most suitable model.

2.1 Model selection

The aim of MS is to identify an optimal model from a set of possible candidates using a selection criterion (such as the aforementioned AIC). The MS technique can also be seen as a special case of model averaging (see Section 2.2 for details), where a weight of 1 is given to one distribution function and a weight of 0 is assigned to all other models considered. The efficiency of selecting the right parent model using various model selection techniques and their effect on design flood estimation has been discussed in detail in the hydrological literature (e.g. Turkman 1985; Di Baldassarre *et al.* 2009; Laio *et al.* 2009).

2.2 Model averaging (MA and MM)

Both model averaging methods (MA and MM) address the issue of uncertainty in the choice of probability distribution functions, by combining all model estimates of the design flood. Several studies have demonstrated the use of MA in dealing with model structure uncertainty (Bodo and Unny 1976, Tung and Mays 1981a, 1981b, Laio *et al.* 2011, Najafi *et al.* 2011, Najafi and Moradkhani 2015, Yan and Moradkhani 2016). Model averaging is similar to the concept of multiple working hypotheses (Chamberlain 1965), which is thought to cope better with the unavoidable bias of using a single model.

The weighted MA technique assigns different weights to the distribution functions considered for estimation. In order to compute these weights, models are first ranked based on their estimated AIC values, followed by the computation of weights for all the distribution functions. The distribution with the minimum AIC is assigned the highest weight. These weights are referred to as Akaike weights (w_i) (Burnham and Anderson 2002):

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta_r\right)}, \quad i = 1, 2, \ldots, R \tag{3}$$

where R is the number of models considered and Δ_i is called the Akaike difference, which represents the discrepancy between the best model with the minimum AIC and the ith model, and is expressed as:

$$\Delta_i = \mathrm{AIC}_i - \min_{i=1,..R}, \quad i = 1,\ 2,\ \ldots, R \qquad (4)$$

A zero value for the Akaike difference (i.e.$\Delta_i = 0$) points to the best distribution function to be used to fit the data. The arbitrariness in the use of Akaike weights is recognized in this work since, in practice, the "true" design flood is not known and the weighting only gives information about the adequacy of a model to fit the observations, not about the accuracy of the estimated discharge.

Let us consider R competing probability distribution functions that are denoted f_i. A posterior predictive distribution of a quantity of interest φ (e.g. a design flood) given the vector of observed data $\boldsymbol{X}$ can be expressed as:

$$p(\varphi|X) = \sum_{i}^{R} p(\varphi|X,\ f_i)\ p(f_i|X) \qquad (5)$$

where $p(.|X)$ represents the conditional probability distribution function and $p(f_i|X)$ is the posterior probability for a given model. Equation (5) was adapted from Hoeting *et al.* (1999) and provides a way of averaging the posterior distributions of the design flood under each of the models considered, weighted by the posterior model probability $p(f_i|X)$. The posterior model probability represents the degree of fit between a particular distribution function and the data, and can be assigned by expert judgement (Merz and Thieken 2005), estimated using Bayesian or Akaike techniques, with the latter already described earlier as Akaike weights w_i.

Uncertainty in the parameters of individual pdfs, and their effect on the accuracy of the estimated design flood is not considered in this study. However, the focus was on evaluating point estimates and not the posterior probability distribution of the design flood; a simplification of Equation (5) is required to implement MA and is expressed as:

$$\hat{Q}_T = \sum_{i=1}^{R} w_i \hat{Q}_{T,i} \qquad (6)$$

where $\hat{Q}_T$ is the estimated design flood for a given return period T. The estimated model weights w_i are assigned to candidate models, with the model that fits the data best having the highest weight.

As for MM, a simple arithmetic average is applied over the design flood estimates of all models, i.e. all models have equal weights. Similar to Graefe *et al.* (2015), we use it as a benchmark to assess the skill of MS.

3 Numerical experiments

3.1 Choice of parent distribution

The Wakeby distribution function was used as the parent distribution to generate synthetic annual maximum flows of different sample sizes. The synthetic samples were used for the systematic assessment of the MS, MA and MM techniques. Various distribution functions (see Table 1) were then used to fit these synthetic time series and estimate the 100-year flood. The Wakeby distribution function is a five-parameter distribution and was defined by Houghton (1977, 1978). The use of the Wakeby distribution first came about as a result of findings by Matalas *et al.* (1975), who showed that many commonly used distribution functions are not capable of reproducing the instability observed in sample estimates of skewness derived from flow records. In other words, the standard deviation of sample estimates of skewness derived from real-world flow data is higher than that derived from synthetic flow data. Matalas *et al.* (1975) called this behaviour the "separation-effect", a contradiction similar to the Hurst effect. The Wakeby quantile function is described as follows:

$$x = a\left[1 - (1-F)^{b}\right] - c\left[1 - (1-F)^{-d}\right] + m \qquad (7)$$

where $F \equiv F\ (x) = P\ (X \leq x)$ and x . The density function $f \equiv f(x)$ is defined as:

$$f = {}^{dF}/_{dx}\left[ab(1-F)^{b-1} + cd(1-F)^{-d-1}\right]^{-1} \qquad (8)$$

The distribution can be thought of in two parts: a left-hand tail $a\left[1-(1-F)^{b}\right]$ (small flows) and a right-hand tail $c\left[1-(1-F)^{-d}\right]+m$ (large flows). The letters *a, b, c, d* and *m* represent the distribution parameters, x is the flood quantile (or design flood) for a given return

Table 1. Probability distribution functions used in this study as operative models.

Probability model	Parameters	pdf or cdf
Gumbel or EV1	(θ_1, θ_2)	$F(x,\theta) = \exp[-\exp(-(x-\theta_1)/\theta_2)]$
Generalized extreme value (GEV)	$(\theta_1, \theta_2, \theta_3)$	$F(x,\theta) = \exp\left[-\left(1-(\theta_3(x-\theta_1)/\theta_2)^{\ 1}/_{\theta_3}\right)\right]$
Pearson Type III (P3)	$(\theta_1, \theta_2, \theta_3)$	$f(x,\theta) = [1/(\lvert\theta\rvert\Gamma(\theta_3+1))]((x-\theta_1)/\theta_2)^{\theta_3}\exp(-[(x-\theta_1)]/\theta_2)$
Lognormal (LN)	(θ_1, θ_2)	$f(x,\theta) = \frac{1}{x\sqrt{2\pi\theta_2}}\exp\left[-\frac{1}{2}\left(\frac{\log x-\theta_1}{\theta_1}\right)^2\right]$

period T, and F is the non-exceedence probability, i.e. $F = 1 - 1/T$. If $F = 0$, then $x = m$ and $f = 1/(ab + cd)$. Note that since $f \geq 0 \forall x, (ab + cd) \geq 0$, for $F = 1$, the values of x and f depend upon the values of the parameters of the distribution, the upper bound on x being $+\infty$ or $(m + a - c)$. Not all parameterizations of the Wakeby distribution are capable of accounting for the conditions of separation mentioned above. However, in an extensive Monte Carlo experiment, Landwehr and Wallis (1978) found that when $b > 1$ and $d > 0$ (i.e. long stretched upper tails) the Wakeby distribution accounts for conditions of separation. The parameter combinations used in this study (i.e. fixed values) in defining a Wakeby parent are listed in Table 2 and were taken from Landwehr and Matalas (1979). A detailed presentation about parameter limits and valid parameter combinations for the Wakeby distribution is provided by Landwehr and Wallis (1978).

We chose the Wakeby distribution for the following reasons: first, we want to simulate the epistemic uncertainty that affects any design flood estimation exercise, i.e. the understanding that the flood generation processes are complex (and not completely known), while simpler models are commonly used for fitting and estimation purposes. The Wakeby distribution has a higher level of complexity in the form of more parameters than the other distribution functions commonly used for estimation purposes. Second, it mimics the upper tail structures typical of flood distributions, which are essential to capture in any synthetic data, i.e. the occasional presence of an outlier (in this case an extreme flood peak), which is not expected, but probable. Third, its quantile function is expressed explicitly in terms of the unknown variable, making the generation of synthetic data straightforward (Hosking and Wallis 1997).

It should be noted that previous numerical studies on flood frequency analysis used more common distribution functions, e.g. lognormal (Matalas *et al.* 1975, Slack *et al.* 1975, Matalas and Wallis 1978) as the parent model. However, our choice was based on the need to simulate the fact that, in the real world, the parent distribution is unknown and likely more complex than the simpler distribution functions.

Table 2. Wakeby distribution functions; μ, σ, Cv, γ and λ denote: mean, standard deviation, coefficient of variation, skewness and kurtosis, respectively.

	Parameters					Statistical characteristics				
Distribution	m	a	b	c	d	μ	σ	Cv	γ	λ
Wakeby-1	0	1	16.0	4	0.20	1.94	1.34	0.69	4.14	63.74
Wakeby-2	0	1	7.5	5	0.12	1.56	0.90	0.58	2.01	14.08
Wakeby-3	0	1	1.0	5	0.12	1.18	1.03	0.87	1.91	10.73
Wakeby-4	0	1	16.0	10	0.04	1.36	0.51	0.38	1.10	7.69
Wakeby-5	0	1	1.0	10	0.04	0.92	0.70	0.76	1.11	4.73

Source: Landwehr and Matalas (1979)

3.2 Choice of probability distribution functions

Four commonly used distribution functions were selected as operational models (i.e. $R = 4$) to fit the synthetic flows and to estimate the 100-year event. The distribution functions considered are (i) the EV1 (Gumbel) distribution, (ii) the generalized extreme value (GEV) distribution, (iii) the generalized gamma or Pearson Type III (P3) distribution, and (iv) the lognormal (LN) distribution. Table 1 provides their cumulative distribution functions (cdf), $F(x, \theta)$, and the probability density functions (pdf), $f(x, \theta)$; the latter are shown for those distribution functions whose cdf is not invertible.

3.3 Simulation framework

We set up a Monte Carlo simulation framework consisting of the following steps, in which the procedure is repeated for each of the Wakeby parent distributions fully determined by the five sets of parameters reported in Table 2. We also let the sample size n vary by assuming values of 30, 50, 100 and 200 years.

(1) One of the Wakeby pdfs, with a fixed set of parameters (Table 2) is selected as parent distribution $g(x)$. As parameters are fixed, the "true" design flood value Q_{100} is the quantile corresponding to a return period of 100 years, which is computed using Equation (7), with $F = 1 - 1/100$.
(2) The Wakeby cdf described in Equation (7) is used to generate a sample of synthetic annual maximum flows Q of fixed length; these values are considered true discharges. Introducing observation error, corrupted discharges Q^* are generated using the error model for uncorrelated observation error (Kuczera 1992) as follows:

$$Q^* = Q + \beta Q \varepsilon \tag{9}$$

where ε denotes a standard Gaussian random variable (i.e. zero mean and standard deviation of 1), Q is the true discharge, and β is a positive valued coefficient denoting the magnitude of observation error. Values for β of 0.00, 0.15 and 0.30 (i.e. 0%, 15% and 30%) are magnitudes of observation error considered in this study and taken from Di Baldassarre *et al.* (2012). A β value of 0% represents the scenario in which observed discharge equals the true discharge; thus there is no observation error.

(3) Using the corrupted discharges Q^*, the parameters of the four pdfs (Table 1) are estimated using the method of maximum likelihood. For the P3 and GEV distributions, maximum likelihood estimators are either not available or asymptotically efficient in a few non-regular cases. Due to this, Smith's estimators (Smith 1985) were used instead of maximum likelihood estimators.
(4) The four pdfs are used to estimate the design flood as the quantile corresponding to a return period of 100 years.
(5) AIC is applied for both MS and MA:
   (i) MS: AIC or AIC_c (depending on the sample size generated, see Section 2) is applied by using Equation (1) or (2), respectively, for the four distribution functions and the optimal distribution is used to estimate the design flood as the flood quantile corresponding to a return period of T years.
   (ii) MA: Using Equation (4), the Akaike differences Δ_i ($i = 1, 2, 3, \ldots, R$) are evaluated and used for the computation of model weights using Equation (3); the estimated design floods for each of the candidate distribution functions are combined by applying Equation (6).
(6) The arithmetic average (MM) of design floods estimated using the candidate distribution functions (Step 4) is implemented.
(7) A percentage relative error is computed in order to compare the true design flood (derived in Step 1) with the design floods estimated by: each of the four candidate models (as in Step 4), model selection (MS, Step 5(a)), weighted model averaging (MA, Step 5(b)), and arithmetic model averaging (MM, Step 6). Thus, we obtained seven relative error estimates (four candidate distribution functions, MS, MA, and MM).

Steps 2–7 are repeated 1000 times, generating 1000 synthetic flow samples from a given parent Wakeby distribution and of a fixed sample size. A generated sample size of 30 and 50 years reflects the typical length of historical observations, while samples of length 100 and 200 years represent an optimistic case in hydrology.

4 Results

Box plots are used to support the comparison between the different techniques (four candidate distribution functions, MS, MA and MM). Figures 1 to Figures 5 show the results of the numerical experiments and summarize the performance of MA, MM and MS, and also the candidate distributions, in estimating the 100-year flood, for different statistical characteristics of the underlying parent distribution, different record lengths and levels of observation uncertainty.

Figure 1 shows box plots of percentage relative estimation errors when Wakeby-1 is used as the parent distribution. Observation errors for a given sample size increase from the left to the right panels, while the sample size for a given observation error increases from the top to the bottom panels. In general, a tendency towards underestimation is observed for all techniques, namely MS, MA and MM, and the individual distribution functions when the parent is highly skewed, as shown in Figures 1 and 2, respectively. For instance, considering Wakeby-1 as the parent model, an error magnitude of 15% and a sample size of 50 years, on average, MA underestimates the true design flood by 19.6%, while MS and MM give an equal underestimation of 22.3%. Major deviations across all techniques and distribution functions appear to be reasonable, as the underlying population was based on a complex parent distribution with five parameters, while the fitting is conducted using distribution functions with only two or three parameters.

Figures 3–5 show the boxplots obtained by using Wakeby-3 to Wakeby-5 as parent distributions, and in that order refer to the reduction in skewness of the parent distributions (see Table 2 for details about the value of skewness for each Wakeby parent). These diagrams show that, in general, all three techniques (MS, MA and MM) tend towards overestimation. For instance, considering Wakeby-3 as the parent model, with an error magnitude of 15% and sample size of 50 years, on average, MS, MA and MM overestimate the true value by 1.3, 3.6 and 6.88%, respectively. Looking at the panels of Figures 3–5 from left to right, this overestimation is influenced by increasing observation errors. This is due to the fact that these errors tend to increase the variance of the sample (see Equation (7)), which in turn leads to increased variance of the design flood estimates (Di Baldassarre *et al.* 2012).

Another set of box plots was produced to help understand the influence of Akaike weights used in MA on the overall accuracy and variance of design flood estimates. For example, if one considers the centre panel of the first row in Figure 6 (i.e. the case of β = 15% and sample size 30), the interpretation is as follows: on average, the best model

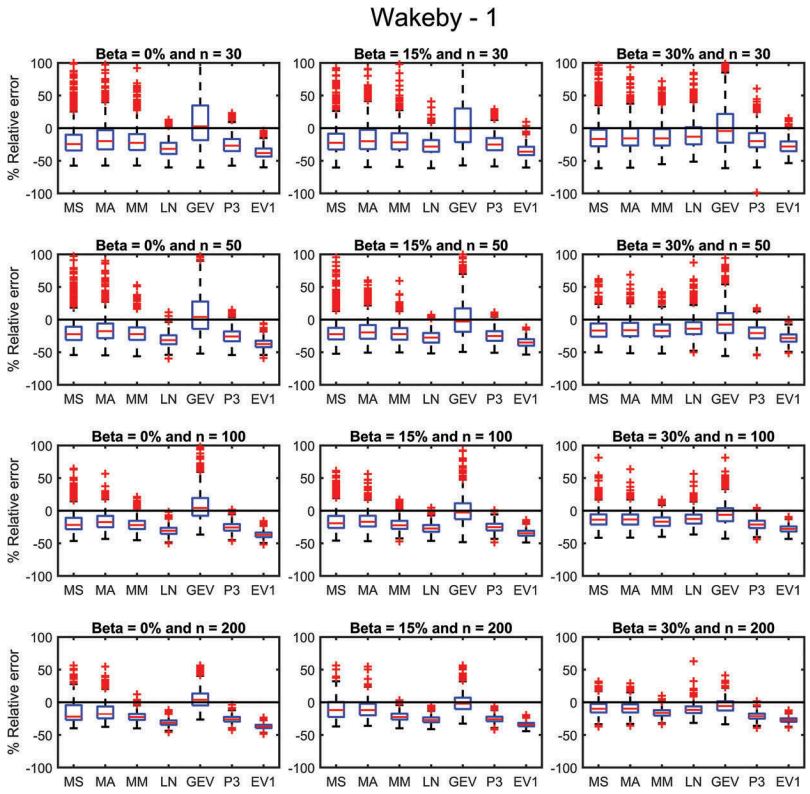

Figure 1. Box plots of percentage relative error for MS, MA, MM and all candidate models, with Wakeby-1 as parent model. The red line represents the median (50th percentile) and the lower and upper ends of the blue box represent the 25th and 75th percentiles, respectively. Outliers are represented by red crosses.

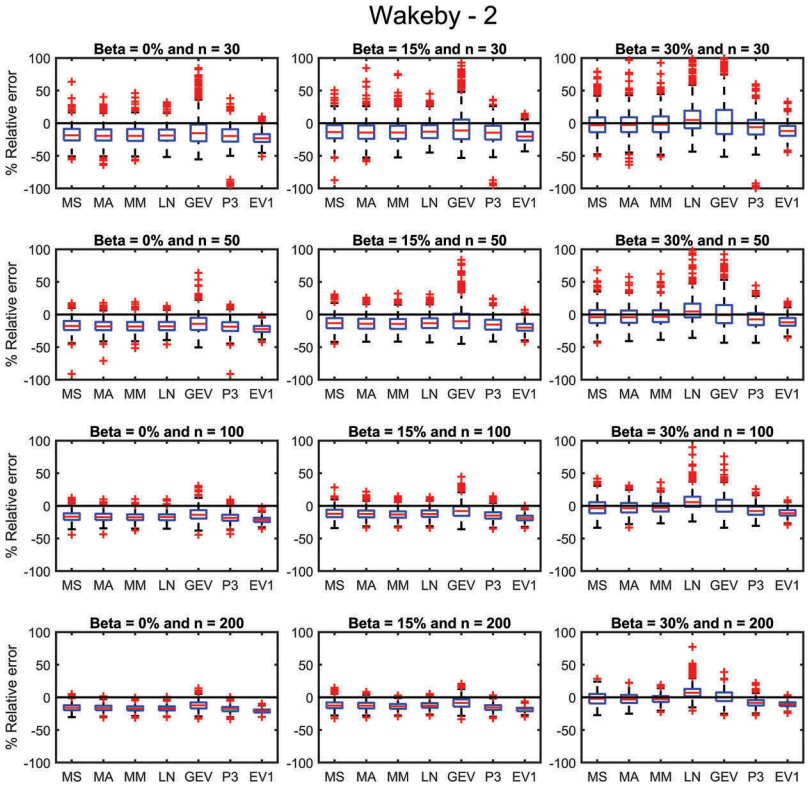

Figure 2. Box plots of percentage relative error for MS, MA, MM and all candidate models, with Wakeby-2 as parent model. Symbols as in Figure 1.

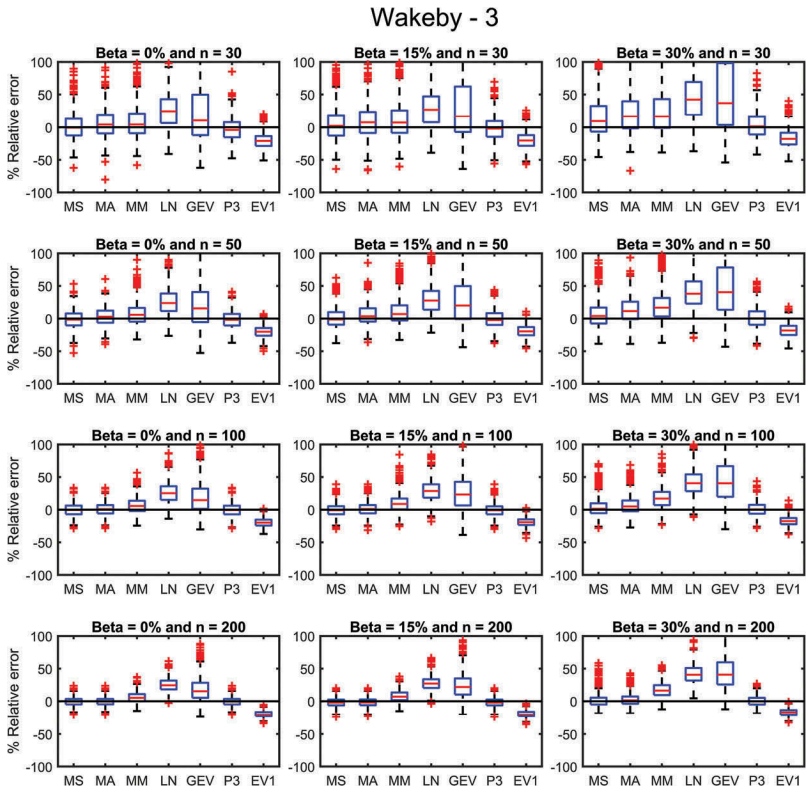

Figure 3. Box plots of percentage relative error for MS, MA, MM and all candidate models, with Wakeby-3 as parent model. Symbols as in Figure 1.

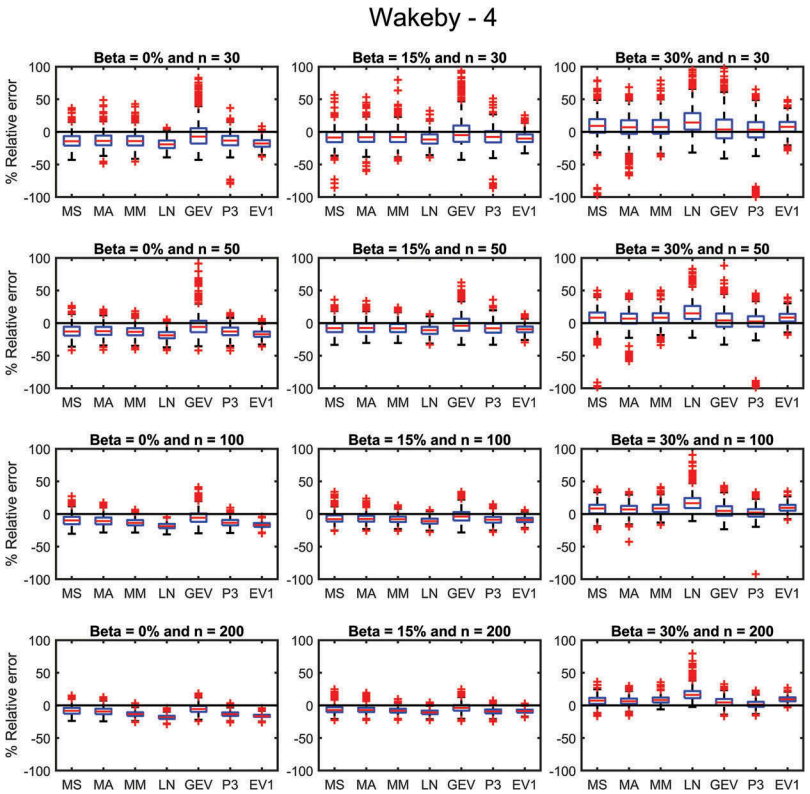

Figure 4. Box plots of percentage relative error for MS, MA, MM and all candidate models, with Wakeby-4 as parent model. Symbols as in Figure 1.

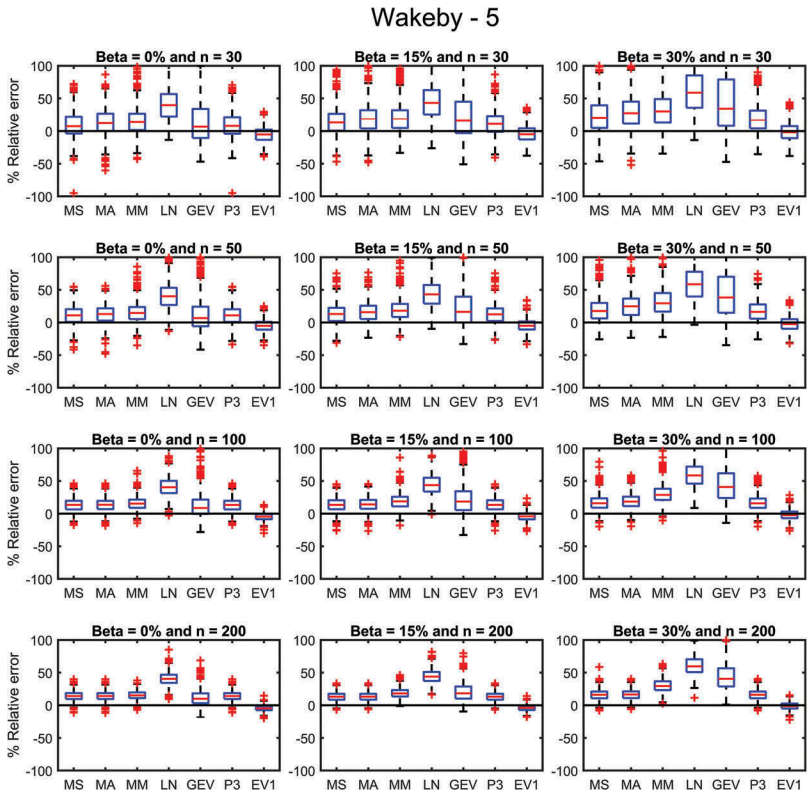

Figure 5. Box plots of percentage relative error for MS, MA, MM and all candidate models, with Wakeby-5 as parent model. Symbols as in Figure 1.

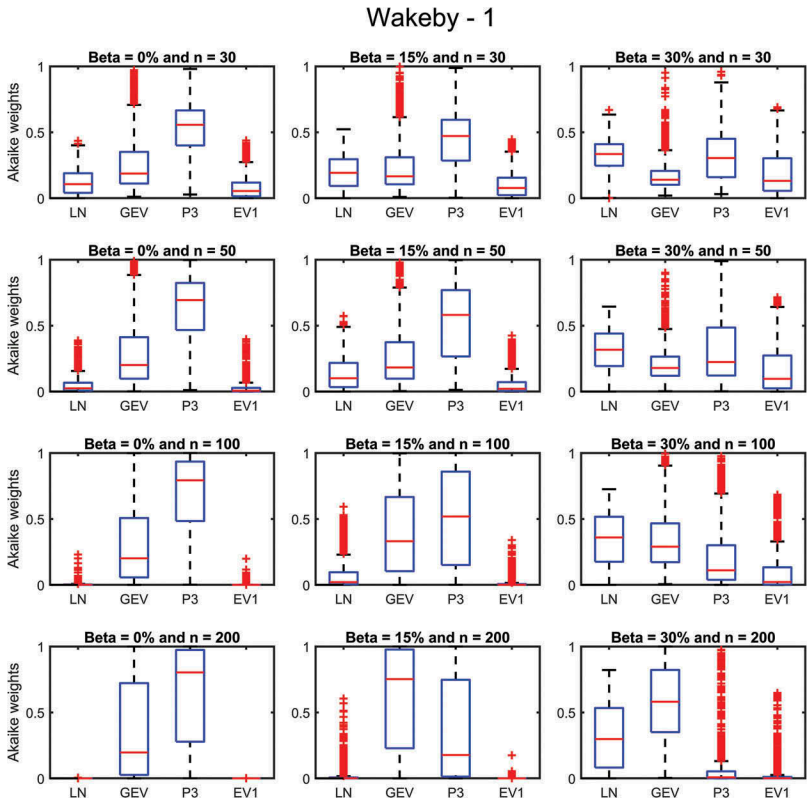

Figure 6. Box plots of Akaike weights for all candidate distribution functions with Wakeby-1 as parent model. Symbols as in Figure 1.

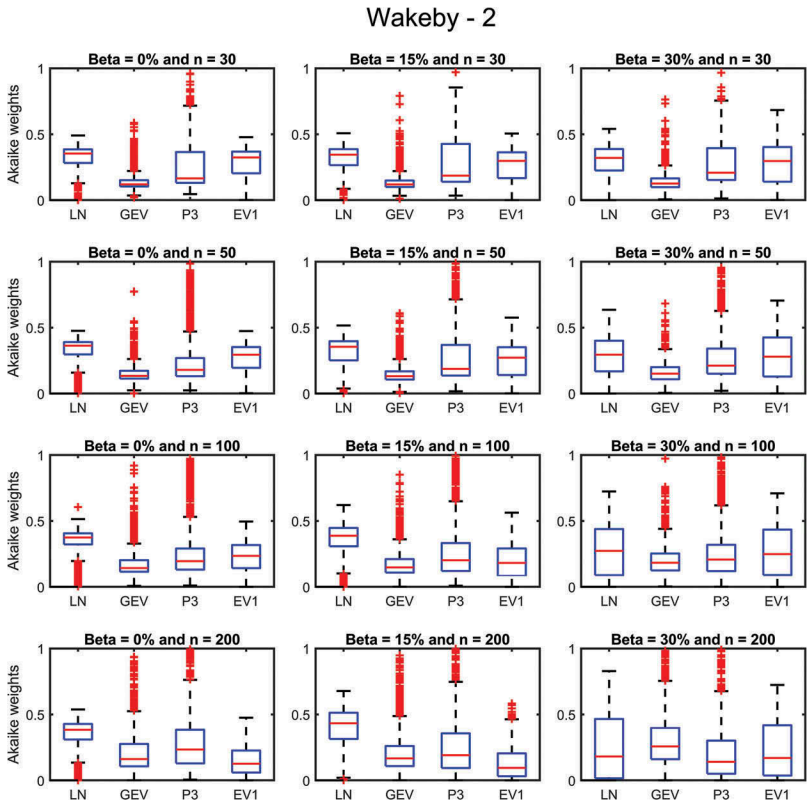

Figure 7. Box plots of Akaike weights for all candidate distribution functions with Wakeby-2 as parent model. Symbols as in Figure 1.

among the candidates is P3, and it accounts for approximately 50% of the weighted average; LN and GEV account for 20 and 18% respectively; while EV1 accounts for 12% of the weighted average. Thus, P3 is clearly the best distribution function in terms of Akaike weights when the parent is Wakeby-1. Figure 1 shows that – for the same 15% error – P3 has a highly biased estimate (with less variance) compared to the GEV, which has less bias but increased variance. The selection of P3 as the best distribution that fits the data in Figure 6 is fairly consistent, as the sample size increases from 30 to 100 for an error of 15%. Note that this behaviour changes when the sample size increases to 200, with GEV as the best distribution function followed by P3. If Wakeby-2 is selected as the parent, Figure 7 suggests that LN is the distribution function that fits the data best for almost all sample sizes and error magnitudes considered. Comparing Figures 2 and 7, it is observed that all models have almost the same accuracy and variance, except for EV1, which is slightly more biased. In summary, it is observed that, on the one hand, the performance between P3 and GEV (comparing Figs. 1 and 6) presents a scenario where a distribution function with the highest weight has less accuracy but small variance, and, on the other, comparing LN with all other distribution functions (Figs. 2 and 7) presents a scenario where distribution functions with a smaller weight have almost equal variance, with the distribution function having the highest weight. That is, having a higher (or lower) Akaike weight for a given distribution function does not necessarily translate to better (or worse) estimates of the design flood. The reason is that the Akaike weight, or any other model selection criterion, refers only to how well the model fits the data and not to how good the estimation is. Box plots of Akaike weights for Wakeby-3 to Wakeby-5 can be found in the Supplementary material.

Tables 3 and 4 show the root mean square error (RMSE) and average percentage relative error (RE %), respectively, for the three methods and the four candidate distribution functions. Table 3 shows that for Wakeby-1, which has a true design flood of 7.05 m^3/s, MA has a slightly better accuracy (±1.71 m^3/s) when compared to MS (±1.87 m^3/s) and MM (±1.77 m^3/s). The same pattern was observed (Table 4) for the three techniques in terms of RE%. As skewness reduces, i.e. from Wakeby-2 to Wakeby-5, we see that the three techniques have similar performance in terms of RMSE

Table 3. Root mean square error (RMSE) for all techniques and distribution functions for a sample size of 50 and a magnitude error of 15%.

Distribution	True 1 in 100 (m^3/s)	RMSE						
		MS	MA	MM	LN	GEV	P3	EV1
Wakeby-1	7.05	1.87	1.71	1.77	2.05	2.19	1.88	2.50
Wakeby-2	4.69	0.79	0.81	0.82	0.78	0.90	0.87	0.98
Wakeby-3	4.68	0.68	0.77	1.04	1.69	2.78	0.62	0.99
Wakeby-4	3.02	0.37	0.36	0.36	0.41	0.46	0.38	0.35
Wakeby-5	3.01	0.59	0.67	0.83	1.50	1.73	0.59	0.31

Table 4. Average percentage relative error (RE%) for all techniques and distribution functions for a sample size of 50 and a magnitude error of 15%.

Distribution	True 1 in-100 (m^3/s)	Average RE (%)						
		MS	MA	MM	LN	GEV	P3	EV1
Wakeby-1	7.05	5.74	5.83	5.91	−26.95	2.72	−24.42	−34.63
Wakeby-2	4.69	4.09	4.06	4.06	−12.72	−8.19	−14.96	−19.39
Wakeby-3	4.68	4.70	4.92	5.16	29.36	32.09	−0.75	−19.09
Wakeby-4	3.02	2.82	2.82	2.82	−10.80	−1.88	−7.52	−9.42
Wakeby-5	3.01	3.39	3.48	3.58	44.45	24.67	12.47	−4.31

and average RE%. For instance, for Wakeby-3, with a true design flood of 4.68 m^3/s, MS has an accuracy of ±0.68 m^3/s when compared to MA and MM, with an RMSE of ±0.77 and ±1.04 m^3/s, respectively. However, all three models (MS, MA and MM) have an average RE% of 2.8%. Table 4 shows that, overall, AIC techniques always have a smaller average RE%, except for the case of Wakeby-1, where GEV has the lowest value. This may be seen as a positive outcome of model selection methods in selecting distribution functions for estimation purposes.

5 Discussion and conclusions

When it comes to flood frequency analysis, the true distribution of floods (which includes the true design flood corresponding to a given return period) is not known *a priori*. Therefore, the task for model selection – leading to a single best distribution function – and model averaging methods is driven towards better estimation, rather than the search for the true distribution that generated the data.

In this study, the MM approach assigns equal weights to all candidate distribution functions without taking into account how well these distribution functions fit the data. The MA approach is different from MM in the sense that the former takes into account individual performance of all distribution functions in fitting the data. The MA approach assigns higher weights to distribution functions that give better fits to the data.

There are certainly situations in which the distribution functions are all similar, i.e. having almost the same AIC values and Akaike weights, which will lead to similar estimates as MM. It seems this behaviour, where candidate distribution functions have similar AIC values, may be the norm rather than the exception, as seen in studies by Mutua (1994) and Strupczewski *et al.* (2001). This might be the reason why, in our study, MA has about the same level of performance as the MM approach. However, MA can only surpass MM in terms of accuracy of estimates if one or more distributions have sufficient weights, and their estimates are close to the true value of the design flood.

The MM approach is usually neglected as a sort of outcast because of its obvious simplicity when compared to Bayesian and Akaike approaches for model averaging. However, studies in social sciences have demonstrated that MM can perform well compared to ensemble Bayesian model averaging (e.g. Graefe *et al.* 2015), and similar conclusions were drawn from studies focusing on operational and financial forecasts (Clark and McCracken 2010, Graefe *et al.* 2014). By assigning equal weights to all candidate distribution functions when implementing the MM method, one ignores the relative adequacy of fit of individual distributions, thereby deliberately introducing bias by taking into account distribution functions with inadequate fit. The MA approach, however, tends to assign higher weights to models with large degrees of freedom, even though the AIC is formulated to take into account overfitting. The effect of overfitting due to the MA approach may lead to improved accuracy, but at the expense of increased variance of the estimated design floods. However, introducing bias by implementing the MM approach may lead to less overfitting but reduced variance.

The trade-off between accuracy and variance observed for some candidate distribution functions might be the reason for the similar performance between MA and MM. To illustrate that trade-off, let us consider the top right corner of Figure 1: the GEV model provides good accuracy, but high variance when compared to the two averaging approaches MA and MM. The same figure shows that LN has less variance but also less accuracy when compared to GEV. However, a comparison of the distribution of Akaike weights (see Fig. 6,

top right corner) shows that GEV has less weight compared to LN. This shows that there is no clear relationship between the calculated weights and the accuracy and variability of the estimates, and therefore demands that one must give some thought to the estimation problem before making up a list of distribution functions suitable for reliable design flood estimates. Also, one can speculate that the similar performance provided by the MA and MM approaches relates to the fact that none of the candidate distributions deviate too much from the parent distribution.

For water management, selecting a distribution function with a high variance of the estimated design flood will complicate the design of an infrastructure, i.e. there is potential for substantial overdesign if the upper limit of the confidence interval is considered. An unbiased estimate in flood estimation is desirable, but, given the numerous sources of uncertainty, engineers and planners usually do not mind sacrificing accuracy in exchange for reduced variability in estimations (Slack *et al*. 1975). However, this ethos of preferring a distribution function with reduced variability can be problematic, since the true design flood is not known in advance; it may lead to an increased risk of over- or under-design of water-related infrastructure if the true design flood is outside the calculated confidence intervals. Furthermore, the decision on the design flood for a given infrastructure does not only depend on estimates based on distribution functions, but also on risk perception and economic feasibility.

Our study likewise shows that, when facing short sample sizes (30–50 years), which are common in hydrology and water resources engineering applications, model averaging (MA and MM) and model selection (MS) lead to better results than arbitrarily selecting a single distribution function. Moreover, for very large sample sizes (100–200 years), which are rare in real-world applications, our study shows that MS, MA and MM have similar variance also when observation uncertainty is introduced. This is related to the fact that the sample sizes are large enough for a better estimation of parameters (even for highly parameterized distribution functions such as GEV), but may not lead to reduced variance due to over-fitting.

It is important to note that our work is focused purely on the estimation of design floods using statistical techniques. Several limitations, such as the distribution functions considered, have unavoidably influenced our results. Future studies on design flood estimation could be extended to consider the physical processes behind flood generation.

Acknowledgements

This research was carried out within the CNDS (Centre of Natural Hazards and Disaster Science) research school, www.cnds.se. We thank Francesco Laio, two anonymous reviewers and the editor for providing critical comments that helped to improve an earlier version of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Kenechukwu Okoli http://orcid.org/0000-0002-5880-607X

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. *In*: B.N. Petrov and F. Csáki, eds. *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 2–8, 1971, Budapest: Akadémiai Kiadó, 267–281.

Beven, K., 1993. Reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, 16, 41–51.

Beven, K., 2006. A manifesto for the equifinality thesis. *Journal of Hydrology*, 320 (1–2), 18–36. doi:10.1016/j.jhydrol.2005.07.007

Beven, K., 2012. *Rainfall - runoff modelling the primer*. 2nd ed. West Sussex: John Wiley & Sons.

Beven, K. and Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, 6, 279–298.

Bodo, B. and Unny, T.E., 1976. Model uncertainty in flood frequency analysis and frequency-based design. *Water Resources Research*, 12 (6), 1109–1117.

Brandimarte, L. and Di Baldassarre, G., 2012. Uncertainty in design flood profiles derived by hydraulic modelling. *Hydrology Research*, 43 (6), 753. doi:10.2166/nh.2011.086

Breinl, K., 2016. Driving a lumped hydrological model with precipitation output from weather generators of different complexity. *Hydrological Sciences Journal*, 61 (8), 1395–1414.

Burnham, K.P. and Anderson, D.R., 2002. *Model selection and multimodel inference*. 2nd ed. New York: Springer.

Chamberlain, T.C., 1965. The method of multiple working hypotheses [reprint of 1890 science article]. *Science*, 148, 754–759.

Clark, T.E., and McCracken, M.W., 2010. Averaging forcasts from VARs with uncertain instabilities. *Journal of Applied Econometrics*, 25 (1), 5-29. doi:10.1002/jae.1127

Di Baldassarre, G., Laio, F., and Montanari, A., 2009. Design flood estimation using model selection criteria. *Physics and Chemistry of the Earth*, 34 (10–12), 606–611, Parts A/B/C. doi:10.1016/j.pce.2008.10.066

Di Baldassarre, G., Laio, F., and Montanari, A., 2012. Effect of observation errors on the uncertainty of design floods. *Physics and Chemistry of the Earth*, 42–44, 85–90.

Di Baldassarre, G. and Montanari, A., 2009. Uncertainty in river discharge observations: a quantitative analysis. *Hydrology and Earth System Sciences Discussions*, 6 (1), 39–61. doi:10.5194/hessd-6-39-2009

Foglia, L., *et al.*, 2013. Evaluating model structure adequacy: the case of the Maggia Valley groundwater system, southern Switzerland. *Water Resources Research*, 49 (1), 260–282. doi:10.1029/2011WR011779

Graefe, A., *et al.*, 2014. Combining forecasts: an application to elections. *International Journal of Forecasting*, 30 (1), 43–54. doi:10.1016/j.ijforecast.2013.02.005

Graefe, A., *et al.*, 2015. Limitations of ensemble Bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31 (3), 943–951. doi:10.1016/j.ijforecast.2014.12.001

Hoeting, J.A., *et al.*, 1999. Bayesian model averaging: a tutorial. *Statistical Science*, 14 (4), 382–417. doi:10.2307/2676803

Hosking, J.R.M. and Wallis, J.R., 1997. *Regional frequency analysis: an approach based on L-moments*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511529443

Houghton, J.C. 1977. *Robust estimation of the frequency of extreme events in a flood frequency context*. Ph.D dissertation. Harvard University, Cambridge, MA.

Houghton, J.C., 1978. Birth of a parent: the Wakeby distribution for modeling flood flows. *Water Resources Research*, 14, 6. doi:10.1029/WR015i005p01288

Klemeš, V., 1986. Dilettantism in hydrology: transition or destiny? *Water Resources Research*, 22 (9S), 177S–188S. doi:10.1029/WR022i09Sp0177S

Klemeš, V., 1993. Probability of extreme hydrometeorological events –a different approach. *In*: *Proceedings of the Yokohama Symposium, Extreme Hydrological Events: Precipitation, Floods and Droughts*, Vol. 213, Yokohama, Japan, IAHS Publ. Wallingford, UK: IAHS Press, Centre for Ecology and Hydrology, 167–176.

Koutsoyiannis, D., 2004. Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. *Hydrological Sciences Journal*, 49, 4.

Kuczera, G., 1992. Uncorrelated measurement error in flood frequency inference. *Water Resources Research*, 28 (1), 183–188.

Kuczera, G., 1996. Correlated rating curve error in flood frequency inference. *Water Resources Research*, 32 (7), 2119–2127.

Laio, F., *et al.*, 2011. Spatially smooth regional estimation of the flood frequency curve (with uncertainty). *Journal of Hydrology*, 408 (1–2), 67–77. doi:10.1016/j.jhydrol.2011.07.022

Laio, F., Di Baldassarre, G., and Montanari, A., 2009. Model selection techniques for the frequency analysis of hydrological extremes. *Water Resources Research*, 45 (7), W07416. doi:10.1029/2007WR006666

Landwehr, J.M. and Matalas, N.C., 1979. Estimation of parameters and quantiles of Wakeby distributions 1. Known lower bounds. *Water Resources Research*, 15 (6), 1361–1372.

Landwehr, J.M. and Wallis, J.R., 1978. Some comparisons of flood statistics in real and log space. *Water Resources Research*, 14 (5), 902–920.

Linhart, H. and Zucchini, W., 1986. *Model selection*. Hoboken, NJ: John Wiley.

Luke, A., *et al.*, 2017. Predicting nonstationary flood frequencies: evidence supports an updated stationarity thesis in the United States. *Water Resource Research*, 53 (7), 5469–5494.

Matalas, N.C., Slack, J.R., and Wallis, J.R., 1975. Regional skew in search of a parent. *Water Resources Research*, 11 (6), 815–826.

Matalas, N.C. and Wallis, J.R., 1978. Some comparisons of flood statistics in real and log space. *Water Resources Research*, 14 (5), 902–920.

Merz, B. and Thieken, A.H., 2005. Separating natural and epistemic uncertainty in flood frequency analysis. *Journal of Hydrology*, 309 (1–4), 114–132. doi:10.1016/j.jhydrol.2004.11.015

Milly, P.C.D., *et al.*, 2008. Climate change - stationarity is dead: whither water management? *Science*, 319 (5863), 573–574.

Montanari, A. and Koutsoyiannis, D., 2014. Modeling and mitigating natural hazards: stationarity is immortal! *Water Resources Research*, 50 (12), 9748–9756.

Moretti, G. and Montanari, A., 2008. Inferring the flood frequency distribution for an ungauged basin using a spatially distributed rainfall–runoff model. *Hydrology and Earth System Sciences Discussions*, 5 (1), 1–26. doi:10.5194/hessd-5-1-2008

Mutua, F.M., 1994. The use of the Akaike Information Criterion in the identification of an optimum flood frequency model. *Hydrological Sciences Journal*, 39 (3), 235–244. doi:10.1080/02626669409492740

Najafi, M.R. and Moradkhani, H., 2015. Multi-model ensemble analysis of runoff extremes for climate change impact assesment. *Journal of Hydrology*, 525, 352–361.

Najafi, M.R., Moradkhani, H., and Jung, I.W., 2011. Assessing the uncertainties of hydrologic model selection in climate change impact studies. *Hydrological Processes*, 25. doi:10.10002/hyp.8043

Potter, K.W. and Walker, J.F., 1985. An empirical study of flood measurement error. *Water Resources Research*, 21 (3), 403–406. doi:10.1029/WR021i003p00403

Schöniger, A., *et al.*, 2014. Model selection on solid ground: rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, 50, 5342–5350. doi:10.1002/2012WR013085

Serinaldi, F. and Kilsby, C.G., 2015. Stationarity is undead: uncertainty dominates the distribution of extremes. *Advances in Water Resources*, 77, 17–36.

Slack, J.R., Wallis, J.R., and Matalas, N.C., 1975. On the value of information to flood frequency analysis. *Water Resources Research*, 11 (5), 629–647. doi:10.1029/WR011i005p00629

Smith, R.L., 1985. Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72 (1), 67–90.

Sonuga, J.O., 1972. Principal of maximum entropy in hydrologic frequency analysis. *Journal of Hydrology*, 17, 177–191.

Strupczewski, W.G., Singh, V.P., and Feluch, W., 2001. Non-stationary approach to at-site flood frequency modelling I. Maximum likelihood estimation. *Journal of Hydrology*, 248, 123–142.

Sugiura, N., 1978. Further analysts of the data by akaike's information criterion and the finite corrections: further

analysts of the data by akaike's. *Communications in Statistics - Theory and Methods*, 7 (1), 13–26.

Tsai, F.T.-C. and Li, X., 2008. Water resources research. *Inverse Groundwater Modeling for Hydraulic Conductivity Estimation Using Bayesian Model Averaging and Variance Window*, 44 (9), n/a-n/a. doi:10.1029/2007WR006576

Tung, Y. and Mays, L.W., 1981a. Optimal risk-based design of flood levee systems. *Water Resources Research*, 17 (4), 843–852.

Tung, Y.K. and Mays, L.W., 1981b. Risk models for flood levee design. *Water Resources Research*, 17 (4), 833–841. doi:10.1029/WR017i004p00833

Turkman, R.F., 1985. The choice of extremal models by Akaike's information criterion. *Journal of Hydrology*, 82, 307–315.

US Water Resources Council, 1982. *Guidelines for determining flood flow frequency: bulletin 17B, hydrology subcommittee, office of water data coordination, US geological survey, Reston Virginia*. Washington, DC: U.S. Government Printing Office.

Viglione, A., *et al.*, 2013. Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research*, 49 (2), 675–692. doi:10.1029/2011WR010782

Volpi, E., *et al.*, 2017. Sworn testimony of the model evidence: Gaussian Mixture Importance (GAME) sampling. *Water Resources Research*, 53 (7), 6133–6158. doi:10.1002/2016WR020167

Westerberg, I.K. and McMillan, H.K., 2015. Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences*, 3951–3968. doi:10.5194/hess-19-3951-2015

Yan, H. and Moradkhani, H., 2016. Towards more robust extreme flood prediction by Bayesian hierarchical and multimodeling. *Natural Hazards*, 81, 203–225. doi:10.1007/s11069-015-2070-6

Ye, M., *et al.*, 2010. A model-averaging method for assessing groundwater conceptual model uncertainty. *Ground Water*, 48 (5), 716–728. doi:10.1111/j.1745-6584.2009.00633.x