A systematic comparison of statistical and hydrological methods for design flood estimation
Kenechukwu Okoli, Maurizio Mazzoleni, Korbinian Breinl and Giuliano Di Baldassarre

ABSTRACT
We compare statistical and hydrological methods to estimate design floods by proposing a framework that is based on assuming a synthetic scenario considered as ‘truth’ and use it as a benchmark for analysing results. To illustrate the framework, we used probability model selection and model averaging as statistical methods, while continuous simulations made with a simple and relatively complex rainfall–runoff model are used as hydrological methods. The results of our numerical exercise show that design floods estimated by using a simple rainfall–runoff model have small parameter uncertainty and limited errors, even for high return periods. Statistical methods perform better than the linear reservoir model in terms of median errors for high return periods, but their uncertainty (i.e., variance of the error) is larger. Moreover, selecting the best fitting probability distribution is associated with numerous outliers. On the contrary, using multiple probability distributions, regardless of their capability in fitting the data, leads to significantly fewer outliers, while keeping a similar accuracy. Thus, we find that, among the statistical methods, model averaging is a better option than model selection. Our results also show the relevance of the precautionary principle in design flood estimation, and thus help develop general recommendations for practitioners and experts involved in flood risk reduction.

Key words | design floods, hydrological modelling, model averaging, model selection, probability distribution, uncertainty

INTRODUCTION
At the heart of scientific and applied hydrology lies the well-known problem of how to obtain a reliable estimate of extreme flows that might occur at a given location (Blöschl et al. 2013). The need for reliable estimates – especially for extreme conditions – has become increasingly important, partly because of the need to plan and develop strategies that will mitigate against frequent floods that might occur due to climate change (Castellarin et al. 2004). For instance, flood risk assessment and the design of protection measures often require a reliable estimate of extreme flows with a given chance of occurrence (e.g., the 1% annual chance flood) from limited discharge records (Okoli et al. 2018). The engineering community commonly refers to these estimated extreme flows as ‘design floods’ because they have influence on some key parameters (i.e., size, dimensions, cost and safety) that are optimized for any given water related infrastructure (Rasekh et al. 2010; Brandimarte & Di Baldassarre 2012). Two main methods are presently used for the estimation of design floods:

1. statistical methods – generally referred to as ‘flood frequency analysis’ – which often consist of fitting a
probability distribution function, such as the generalized extreme value distribution (GEV), to a record of annual maximum flows (AMF) obtained for a gauged location. The fitted probability model is extrapolated to a flow magnitude corresponding to a selected probability of exceedance or return period (Moran 1957; Klemš 2000a). When the catchment is ungauged, regional frequency analysis of the type developed by Hosking & Wallis (1997) is one out of many techniques that are used for design flood estimations. Old concepts, such as probable maximum precipitation (PMP) developed by Hershfield (1961), which means the statistical estimation of the maximum precipitation to derive a probable maximum flood (PMF), are still in use for the design of hydraulic structures. However, the concept of PMP (including PMF) has been criticized due to the lack of physical justification for upper boundaries in meteorological factors used for storm maximization (Yevjevich 1968).

2. Hydrological methods, which often consist of estimating design floods generally based on the use of a mathematical model that describes the processes accounted for in the transformation of precipitation to runoff at a given catchment. Meteorological data (rainfall, snow and temperature) are common examples of inputs into a rainfall–runoff model with simulated river discharges as the output on completion of a model run (Beven 2012). Rainfall–runoff models can be used in two different simulation modes for the estimation of a design flood. The first one is the ‘event-based simulation’ (EBS) such as the type developed by Mulvaney (1851), otherwise known as the rational method. The event-based approach requires the selection of a design rainfall from an intensity duration frequency (IDF) curve of rainfall with a given duration and assumed profile and uses it as input into the rainfall–runoff model to derive the flood hydrograph (Rogger et al. 2012). Eagleson (1972) proposed a different kind of event-based modelling that requires the coupling of a stochastic weather generator with a catchment response function, which in his study was a kinematic wave model. Event-based approaches are limited in the fact that they do not have a realistic account of the role of antecedent moisture content in runoff generation. The second simulation mode is known as ‘continuous simulation’ (CS) and requires – just like event-based simulation – the coupling of a stochastic weather model with a runoff model. This approach treats the discharge as a single term without prior separation into overland flow and baseflow. The problem of antecedent moisture content of the catchment is addressed implicitly as part of the modelling procedure (Calver & Lamb 1995). CS can be viewed as a consolidated technique for flood estimation hydrology (Blazkova & Beven 1997; Cameron et al. 1999; Pathiraja et al. 2012). It is worth noting that the two hydrological methods of estimation (EBS and CS) are still influenced by the same subjective choices, such as choice of RR model and method of parameter estimation, etc.

Linsley (1986) recommends in his seminal paper on flood estimates testing and discussing the accuracy of hydrologic methods used for flood estimation. The need to test methods used for flood estimation is important for both scientific enterprise and also economic studies that are usually part of flood mitigation plans. One area that has led the charge in testing methods for flood estimation is statistical hydrology, with the literature awash with studies looking at a range of topics, for example, the influence of type of flow data, i.e., block maxima versus peaks over a threshold (Cunnane 1973; Madsen et al. 1997), errors due to rating curves (Di Baldassarre et al. 2012; Steinbakk et al. 2016), choice of plotting positions (Cunnane 1978), the use of model selection techniques (Di Baldassarre et al. 2009; Lallo et al. 2009), or the influence of fitting methods and choice of distributions (Slack et al. 1975; Landwehr et al. 1980), just to mention a few. The statistical approach is considered the standard method for design flood estimation but it has also faced a great deal of criticism. The main argument against its use is that the structure of the probability model used in extrapolating the flow record is derived based on axioms rooted in probability theory – a branch of mathematics whose main mission does not include the physical representations of flood generation processes (Klemš 1986, 2000a, 2000b). Therefore, extrapolating the flow record with a rainfall–runoff model can provide a more solid physical ground compared to statistical methods, since the former represents a formal statement about flood generation processes and allows including the influence of...
threshold effects (e.g., catchment wetness) that are known to affect the shape of the flood frequency curve. Thus, rainfall–rainfall modelling can support both understanding the underlying flood generation processes and interpreting them.

There are only a handful of studies dealing with the comparison of statistical approaches, event-based and continuous modelling. Grimaldi et al. (2012, 2013) compared continuous modelling with event-based approaches in small ungauged watersheds in Italy using spatially uniform rainfall at sub-hourly resolution. They concluded that the event-based approaches tend to underestimate design hydrographs in terms of flood volume and duration. Rogger et al. (2012) compared statistical, event-based and continuous approaches in ten small Austrian catchments based on observation data at sub-hourly resolution. Their main finding was that in catchments with a high storage capacity, there can be a step change in the flood frequency curve when exceeding a certain storage threshold. This step change cannot be represented by statistical flood frequency analysis, which thus tends to underestimate floods in such catchments. Breinl (2016) examined different types of weather generators coupled to a lumped hydrological model in two small Alpine catchments, concluding that continuous modelling and statistical approaches can lead to comparable results at daily resolution. Oliver et al. (2019) conducted a probabilistic flood risk assessment at daily time scale, starting from a set of stochastic weather models (Wilks 1998; King et al. 2012; Breinl et al. 2017) coupled to rainfall–runoff modelling. They conducted a full flood risk assessment across four major river catchments in India (up to an area of 21,100 km²), concluding that there is a need to generally better understand extreme hydrological events at the regional scale and, more specifically, examine the role of weather models, joint probability approaches and efficient simulations in continuous modelling. Winter et al. (2019) compared continuous modelling, event-based and statistical approaches in small Alpine catchments in Austria using sub-daily time series. While statistical and continuous modelling led to comparable results for design floods, event-based approaches led to an underestimation of flood volumes. Winter et al. (2019) however state that ‘it is hardly possible to identify the ‘correct’ estimation, as all methods are based on the extrapolation of observed patterns in one way or another’. All these studies mentioned above have in common that they compare different methods for deriving design floods with real-world observations.

However, design floods by their nature represent flows whose magnitudes are beyond what has been observed in the flow records. Since they are not known a priori in any practical application, it becomes difficult to assess the performance of different methods that are typically used for their estimation. Thus, in this study, we propose a framework that considers a synthetic scenario as ‘truth’ and use it as a benchmark to evaluate results derived from the two methods of estimation. A synthetic scenario as used in this paper refers to a hydrological model that suggests a representation of our understanding about the real world, i.e., in our case, the flood generation processes. Any rainfall–runoff model and a stochastic weather generator (ranging from simple to complex models) can be selected and its parameters calibrated with available observations. The calibrated models (a coupled weather and rainfall runoff model) are considered to be reality and their outputs – synthetic rainfall and discharges – assumed to be true realizations of the modelled process. This framework also allows the true design flood to be known in advance. Bashford et al. (2002) used a similar concept based on assumed truths to investigate the role of different kinds of data sets in model structure and parameter identification. However, numerical experiments that require the use of a certain model as the benchmark are affected by the fact that the selected model does not represent all processes, and the processes represented may not be the same as reality. Hence, they are affected by different sources of uncertainties (Beven et al. 2012).

**METHODS**

**Simulation framework**

Figure 1 illustrates the structure of the framework developed to test the approaches to design flood estimation. The framework is structured into four experiments; experiments A and B refer to the statistical methods, while C and D refer to the hydrological methods (CS).
The four experiments are connected to the benchmark scenario which is set up as follows:

1. A catchment is selected where meteorological inputs and information about discharge are available to calibrate a weather generator and rainfall–runoff model, respectively. The calibrated models are treated as the true representation of the process understanding.
2. The calibrated weather generator is used to compute a long synthetic sample of rainfall that is used as input into the rainfall–runoff model to derive a long synthetic sample of discharges.
3. AMF are extracted from the derived discharge sample and design floods for different return periods are computed using the Weibull plotting positions (Cunnane 1978). These design floods are used as the benchmark to compare the estimates based on statistical and CS approaches.

Two separate experiments (A and B) are used to test the performance when design flood is estimated based on statistical methods. Experiment A refers to when the estimation is based on a model selection criterion. Here, the Akaike information criterion is used to select the best probability distribution to use for estimation (Akaike 1973). Experiment B refers to a model-averaged estimate, i.e., the estimates from candidate probability distributions are weighted equally to yield a single-valued estimate (see Okoli et al. 2018). Table 1 shows the four candidate probability distributions considered in this study in terms of the PDF (probability density function) or CDF (cumulative distribution function). Experiment A is implemented as follows:

4. select a sub-sample of discharge data from the long sample of synthetic river discharge derived in Step 2;
5. conduct model selection using AIC and select the best fitting probability distribution;
6. estimate design floods for a given return period using the best probability distribution. Compute the percentage relative error as a way to compare between the estimated design flood and the true design flood derived in Step 3.

While experiment B is implemented as follows:

7. select a sub-sample of discharge data from the long sample of synthetic river discharge derived in Step 2;
8. conduct arithmetic model averaging by assigning equal weights to all candidate distributions;
9. estimate design floods for a selected return period and for each candidate probability distribution. The estimates are averaged by taking the mean of all design flood estimates to yield a single-valued estimate. Compute the percentage relative error as a way to compare between the estimated design flood and the true design flood derived in Step 3.

Experiments C and D are similar with the only difference being the hydrological model selected for use in CS. The two experiments are implemented as follows:

10. select a sub-sample of synthetic weather data and use it to calibrate a weather generator. Generate long time series of synthetic weather data and use that as input into a simple hydrological model (experiment C) or a complex hydrological model (experiment D);
11. simulate long time series of discharges and estimate the design flood for a selected return period using Weibull plotting positions. After that, compute the percentage relative error to compare the estimated design flood with the true design flood derived in Step 3.

In order to investigate the influence due to sampling uncertainty, experiments A, B, C and D are repeated \(m\) times, where \(m\) equals the number of sub-samples that can be created from the long samples derived in Step 3 and used in Step 10, and for different return periods.

**Weather generator**

To simulate long time series of catchment rainfall data we used a univariate Markovian rainfall generation algorithm, as proposed by Richardson (1981). Markovian rainfall modelling has been proven robust over decades for various applications (e.g., Gabriel & Neumann 1962; Stern & Coe 1982; Stowasser 2011; Mhanna & Bauwens 2012). The simplest model setup for univariate rainfall modelling of rainfall occurrence is a two-state (i.e., dry or wet day) first-order Markov process (i.e., the rainfall state on a particular day only depends on the previous day). According to the AIC (Katz 1981; Breinl et al. 2013), a two-state second-order Markov chain turned out to be most appropriate for the present rainfall time series in each calendar month. The model was fitted to the 12 calendar months to reproduce the seasonality of rainfall occurrences. To simulate rainfall amounts, we randomly sampled from a mixed exponential distribution on rainy days, which was estimated using maximum likelihood (Wilks 1998). The mixed exponential distribution was likewise fitted separately to each calendar month to reproduce the seasonality of rainfall amounts. Although the Gamma distribution has been widely applied in the simulation of daily rainfall, its validity is less than assumed (Vlcek & Huth 2009) and the mixed exponential distribution is a good alternative (Foufoula-Georgiou & Lettenmaier 1987; Wilks 1998; Li et al. 2015). The corresponding evapotranspiration and temperature time series were simulated by a

---

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

<table>
<thead>
<tr>
<th>Probability model</th>
<th>Parameters</th>
<th>PDF or CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gumble or EV1</td>
<td>((\theta_1, \theta_2))</td>
<td>(F(x, \theta) = \exp[-\exp(-(x - \theta_1)/\theta_2)])</td>
</tr>
<tr>
<td>Generalized Extreme Value (GEV)</td>
<td>((\theta_1, \theta_2, \theta_3))</td>
<td>(F(x, \theta) = \exp \left[-\left(1 - \frac{(x - \theta_2)}{\theta_3}\right)^\theta\right])</td>
</tr>
<tr>
<td>Gamma or Pearson Type III (P3)</td>
<td>((\theta_1, \theta_2, \theta_3))</td>
<td>(f(x, \theta) = \frac{1}{\theta_2} \frac{1}{\Gamma(\theta_3)} \left(\frac{x}{\theta_3}\right)^\theta \exp\left(-\frac{x}{\theta_3}\right))</td>
</tr>
<tr>
<td>Log-Normal (LN)</td>
<td>((\theta_1, \theta_2))</td>
<td>(f(x, \theta) = \frac{1}{\sqrt{2\pi}\theta_2} \exp\left[-\frac{1}{2} \left(\frac{x - \theta_1}{\theta_2}\right)^2\right])</td>
</tr>
</tbody>
</table>
random resampling procedure on an annual basis, i.e., we randomly resampled entire years of temperature and evapotranspiration to build new time series. Both variables were resampled from the same observed year to maintain the inter-variable correlation. The rationale behind this simplified approach for the simulation of evapotranspiration and temperature derives from the short time series available for our study area. More sophisticated simulation algorithms for evapotranspiration or temperature such as autoregressive models (e.g. Breinl et al. 2015), which can likewise better reproduce dry and wet temperatures, require a more solid data base meaning comparatively long observation time series.

Complex hydrological modelling

We selected the HBV model (Bergström 1992; Seibert & Viz 2012) as the complex (in a relative sense) hydrological model to represent the rainfall–runoff processes (Merz & Blöschl 2004; Bardossy 2007; Jin et al. 2009; Demirel et al. 2015; Vetter et al. 2015) to produce the benchmark scenario. The main model inputs are daily time series of precipitation rainfall, evapotranspiration and temperature, as well as estimates of potential evaporation. The HBV model is characterized by four model modules, which describe precipitation, soil moisture, upper zone and lower zone routing. In the precipitation routine, a degree-day approach is used to compute the sum of snowmelt and rainfall (P, mm/day). The soil moisture routine simulates the unsaturated-zone process, in which the soil wetness is estimated as:

\[
\frac{R}{P} = \left( \frac{SM}{FC} \right)^{\beta} \tag{1}
\]

where \(R\) is the recharge to the upper zone (mm/day), \(SM\) is the soil moisture storage (mm), \(FC\) is the maximum field capacity (mm), and \(\beta\) is the parameter that describes the non-linearity of the process. Besides the recharge \(R\) to the upper zone \(UZ\), the loss of water in the \(SM\) zone is due to evapotranspiration processes. The capillary flux (\(CF\)) entering into the \(SM\) zone by capillary action from the upper zone is estimated as:

\[
CF = CFLUX \left( 1 - \frac{SM}{FC} \right) \tag{2}
\]

where \(CFLUX\) is the parameter indicating the maximum value of capillary flux (mm/day). The main input for the lower zone is the percolation from the upper zone. Finally, the river discharge is calculated as the sum of the quickflow \(Q_0\) (from the upper zone) and baseflow \(Q_B\) (from the lower zone) as (Seibert & Vis 2012):

\[
Q_0 = K_0 \cdot \max(UZ - P_{UZ}) + K_1 \cdot UZ \tag{3}
\]

\[
Q_B = K_2 \cdot LZ
\]

where \(UZ\) and \(LZ\) are the model states representing the water content in the upper zone and lower zone, respectively. \(K_0, K_1 \) and \(K_2\) are recession coefficients, while \(P_{UZ}\) is a threshold value for the upper zone.

In our experiment, 18 parameters of the HBV model were calibrated with the least squares minimization technique using the Broyden–Fletcher–Goldfarb–Shanno variant of the Davidon–Fletcher–Powell minimization (DFPMIN) algorithm (Press et al. 1992). In particular, we aimed at minimizing the root mean squared error (RMSE) between the synthetic and the simulated annual maximum flow.

Probability models: model selection and averaging of estimates

Different kinds of probability models have been proposed for the estimation of design floods in hydrology (El Adlouni et al. 2008). Typically, a range of probability models are specified as potential candidates by the hydrologist and the model to be selected is usually based on a visual assessment of the fit to plotted discharge values, on the use of some metric based on goodness-of-fit or any model selection criterion. The ‘model selection’ criterion assumes that for the data available there is a single correct, or at least ‘best’ model that could be used for statistical inference. Therefore, the problem is formulated as a data-based search, over the candidate models, for that single best model (but with estimated parameters) that will be used for design flood estimation. The two well-known model selection criteria based on information-theoretic selection are the AIC and the Bayesian information criterion (BIC). For their mathematical and philosophical backgrounds, the reader is referred to Burnham & Anderson (2002).
Since all models (which includes both statistical and physical) are conceptually wrong, and the capability of different model structures and parameter sets to give reasonable fits to available data has been shown to be the normal state of affairs (Beven 2006), it is clear that any modelling exercise is faced with model structure uncertainty. One way to deal with model structure uncertainty is to use all candidate probability models for design flood estimation where the final estimate is a weighted average of all individual estimates. This process is commonly referred to as model averaging (Höge et al. 2019). The weights can be assigned in two ways. First, equal weights can be assigned to all probability models and this process is known as arithmetic model averaging (see Graefe et al. 2015; Okoli et al. 2018); it is the same as taking the mean of all design flood estimates. Second, the weights can also be assigned depending on how best the probability model fits the data, i.e., a weighted average where the probability model that fits the data best gets a higher weight and vice versa. Bayesian inference (Hoeting et al. 1999) is one way to estimate the weights assigned to a range of models considered for flood estimation purposes and Wood & Rodríguez-Iturbe (1975) showed an example application in flood frequency analysis. Höge et al. (2019) provide a review of the theoretical backgrounds of Bayesian model averaging (and selection) with hydrologists as the target audience.

In this study we have used AIC as a model selection criterion, while model averaging was applied by taking the mean of design flood estimates from the candidate models.

**Simple hydrological model**

The linear reservoir (LR) model is used as a simple hydrological model to estimate the annual maximum flow. The LR model is represented as a tank having rainfall as input, the reservoir storage as model state, and discharge as model output. The LR is based on the continuity equation and on the assumption that the outflow of a given catchment is linearly related to its storage. Combining these two concepts, the LR model can be represented as:

\[ Q = P\left(1 - e^{-\frac{t}{\theta}}\right) \]  \hspace{1cm} (4)

where \( K \) is the only parameter of the LR model and represents the storage capacity of the catchment. The main assumption in the LR model is that there are no losses from evapotranspiration and percolation. LR models have been used for different kinds of investigation in hydrology, e.g., the influence of land use change in mountainous catchments (Buytaert et al. 2004) and seasonal changes in surface meltwater recharge and glacier storage (Hannah & Gurnell 2001).

The calibration of the LR model is performed using a grid-search approach with 1,000 possible values of the parameter \( K \). The optimal value of \( K \) refers to the one which minimizes the RMSE between the observed and the simulated annual maximum flow.

**EXAMPLE APPLICATION**

**Test site**

The application of the proposed framework is based on the observed rainfall data available in the Brue catchment, located in Somerset, South West England. The Brue catchment has a predominantly rural use and modest slope (Moore et al. 2000). The drainage area of the catchment is about 135 km\(^2\). The average annual rainfall is about 867 mm, while the average flow values (measured at the outlet section of Lovington) are about 1.92 m\(^3\)/s.

The daily precipitation used as input in the univariate Markovian rainfall generation model is supplied by the British Atmospheric Data Centre from the NERC Hydrological Radar Experiment Dataset (HYREX) project (Moore et al. 2000; Wood et al. 2000) from 1965 to 2015. Unfortunately, temperature and evapotranspiration values were available only between 1994 and 1999, so that it was not possible to use sophisticated simulation algorithms to stochastically generate long time series. For this reason, we resampled on an annual basis the information in those few years and calculate the corresponding evapotranspiration and temperature time series for the entire period of the rainfall generation (see the section ‘Weather generator’).

The parameters of the complex hydrological model (HBV) applied on the Brue catchment were calculated by Shrestha et al. (2009).
Numerical experiments

We conducted four different experiments that are linked to the benchmark scenario that assumes a selected weather and hydrological model as true representation of our understanding of the processes. In this benchmark scenario, we first used the weather generator to compute 10,000 years of synthetic weather data that are used as input to drive a complex hydrological model (HBV). The resulting 10,000 years of synthetic discharge data are then assumed to be the ‘truth’, and the design floods of interest can be derived based on the sample by computing the return periods based on the Weibull plotting positions (Cunnane 1978). These design floods are treated as truth and used as benchmarks for all experiments. Experiment A describes the common statistical frequency analysis, i.e., based on a sub-sample of the generated discharge time series (in our case 50 years) and used for estimating the parameters of four different probability distributions by the method of maximum likelihood. The best probability distribution is selected using AIC and then used for design flood estimation AIC. Four different probability distributions (as shown in Table 1) were used as candidate models for estimating the design floods. The fitting procedure is conducted for 200 consecutive samples of 50 years (i.e., the entire 10,000 synthetic years), to allow for uncertainty analyses. Experiment B is similar to Experiment A, but fits a range of probability models to the 50 years’ samples (and like Experiment A, 200 times for uncertainty analysis). Based on the range of models, a weighted average of design flood estimates is computed by assigning equal weights to the candidate probability distributions. Based on these assigned weights, a weighted average of design floods is computed. Experiments C and D describe the continuous model experiments. In Experiment C, a sub-sample of 50 years of the synthetic weather time series is used to calibrate the weather generator and to generate 10,000 years of estimated weather time series. Accordingly, the same sub-sample of 50 years of synthetic rainfall is used to calibrate a simple hydrological model (LR) by minimizing the error between model results and the sub-sample of 50 years of synthetic discharges from the benchmark scenario. The estimated weather time series are then fed into the calibrated simple hydrological model to generate 10,000 years of synthetic discharge. The design floods are derived by Weibull plotting positions. Both the calibration of the weather generator and the hydrological model are conducted 200 times to cover the entire 10,000 years of the benchmark scenario. Experiment D is similar to Experiment C with the difference that a relatively complex hydrological model (HBV) is applied instead of the simple one (LR). Finally, the resulting design floods from all four experiments are compared to the ‘true’ design floods from the benchmark scenario.

Results

The accuracy and precision of design flood estimates based on both statistical and hydrological methods were assessed by conducting four different numerical experiments. Design flood estimates are made for return periods 2, 5, 10, 20, 50, 100, 200, 500 and 1,000 years. In this study, 2 years to 50 years are considered ‘low to moderate’ return periods, which are typically the range of return periods considered for the design of urban drainage systems. Return periods ranging from 100 to 1,000 years are considered as high return periods and are representative of the range selected for river structures ranging from diversion weirs to dams.

Figure 2(a) and 2(b) show details of performance when design flood is estimated based on statistical approach. Figure 2(a) refers to the distribution of errors when design flood estimate is based on model selection for each sub-sample of synthetic discharge (50 years) and repeated for 200 samples. Figure 2(b) refers to distribution of errors when design flood estimate is based on model averaging. There is a general tendency to underestimate for high return periods, i.e., from 50 years and above.

Also, the boxplots show that model averaging is more robust since it has no big outliers compared to model selection. For instance, Figure 2 shows that for return periods ranging from 200 to 1,000 years, model selection leads to an overestimation in the range of 52% to 150%, while for model averaging it is 20% to 25%. Table 2 summarizes the bias for statistical methods (but also hydrological models used for CS) in terms of their median values. The two statistical methods led to similar performance in accuracy.

In further analysis conducted for Experiment A, we found that the four candidate probability distributions
(i.e., Log normal, GEV, Gamma and EV1) were selected only 19, 63, 57, and 61 times, respectively, for the 200 sub-samples. Their performance in terms of accuracy and precision is shown in Figure 3. There is a general tendency to underestimate the design flood for higher return periods for all probability distributions. It is shown that GEV estimates led to higher variance in errors and the large outliers in Experiment A might be due to its frequent selection as the best model.

Experiments C and D (in Figure 2) show the performance when estimation is based on CS. The LR reservoir used in Experiment C underestimates the design flood for high return periods but with less variance in the errors. For a model with only one model parameter to calibrate, its precision is still better compared to statistical methods. Four different objective functions were used to calibrate the LR to see if there would be any improvement of some sort. The results presented in Figure 4 suggest a similar performance of LR for the four objective functions. The lower right-hand panel shows the performance when the true model that generated the calibration data is selected for estimation purposes. There is a tendency to underestimate for all return periods and the variance in error is, in part, due to limited sampling. Figure 5 shows the

Table 2 | Accuracy of design flood estimates for different estimation methods and return periods

<table>
<thead>
<tr>
<th>Remarks</th>
<th>Return period (years)</th>
<th>Model selection (%)</th>
<th>Model averaging (%)</th>
<th>Linear reservoir (%)</th>
<th>HBV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low to moderate</td>
<td>2</td>
<td>0.7</td>
<td>2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.5</td>
<td>2.5</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.2</td>
<td>0.9</td>
<td>-5</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>-3</td>
<td>-4</td>
<td>-7</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>-12</td>
<td>-15</td>
<td>-19</td>
<td>9</td>
</tr>
<tr>
<td>High</td>
<td>100</td>
<td>-17</td>
<td>-19</td>
<td>-25</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-22</td>
<td>-24</td>
<td>-29</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>-26</td>
<td>-29</td>
<td>-35</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>-27</td>
<td>-29</td>
<td>-36</td>
<td>10</td>
</tr>
</tbody>
</table>
quantile-quantile plots of simulated AMF against synthetic AMF, and developed for both LR and HBV models. The quantiles of LR tend to diverge significantly from the theoretical line, which implies that the distribution of AMF derived using LR is different from those derived from the HBV model.

Figure 3 | Accuracy and precision of the four candidate probability distributions used for design flood estimation.

Figure 4 | Boxplots show the precision of estimated design floods when different goodness-of-fit measures are used to evaluate the calibration of linear reservoir model.
DISCUSSION AND CONCLUSIONS

We proposed a simulation framework to compare statistical and hydrological methods for the estimation of design floods. The use of synthetic scenarios (i.e., assuming a given model as the truth) is exploited in order to assess the performance and errors of alternative approaches. To illustrate the framework, we showed an example application where we used model selection (Experiment A) and model averaging (Experiment B) as statistical methods, and continuous simulations made with a simple (Experiment C) or a perfect (Experiment D) rainfall–runoff model as hydrological methods.

The results of our numerical exercise showed that, as expected, using a perfect rainfall–runoff model (Experiment D) provides design flood estimates with the least errors. Such an experiment, however, does not reflect any real applications, because a perfect hydrological model does not exist in the real world. It was only meant to get a baseline for the discussion of the results obtained with other methods. Among them, despite its simplicity, the use of a LR model to estimate design floods (Experiment C) was associated with relatively small parameter uncertainty and limited errors for high return periods, i.e., the range from 100 to 1,000 years. Statistical methods (Experiments A and B) performed better than the latter (Experiment C) in terms of median errors for high return periods (Table 2), but the variance of their errors is larger (Figure 2). While errors in Table 2 are similar for the two statistical methods, selecting the best fitting probability distribution (Experiment A) is associated with numerous outliers, as depicted by the boxplots in Figure 2. This is because a limited sample can lead to choosing the wrong probability distribution, which can (by chance) fit well the limited data set, but can then generate large (often unacceptable) errors for high return periods. On the contrary, we found that using multiple probability distributions (Experiment B), regardless of their capability of fitting the data, leads to significantly fewer outliers (Figure 2), while keeping similar average errors (Table 2). Thus, reflecting the Keynesian saying that ‘it is better to be approximately right than precisely wrong’, from our point of view, model averaging is a better option than model selection.

While these outcomes are unavoidably associated with the test site considered here, as well as the choice of the perfect model (HBV) and the specific statistical and hydrological methods compared, our framework can be easily applied elsewhere to test alternative approaches for design flood estimation. Moreover, our results help develop...
general recommendations. They show that, even in this (simple) benchmark scenario, statistical and hydrological methods out- or under-perform each other because of many other sources of uncertainty that come into play, e.g., limited sample size. This suggest that, in the real world, relying on a single method (be it either statistical or hydrological) can lead to large errors in design flood estimation. As such, we argue that practitioners should always use both statistical and hydrological methods. Both methods are based on consolidated theories, and have complementary advantages and limitations. As such, by following the precautionary principle (Foster et al. 2000), which calls for erring on the side of least consequences, one should get two (or more) design flood estimates based on alternative methods and then pick the maximum value among them. This approach will minimize the likelihood of underestimating the design flood, and therefore help support the development and planning of measures for flood risk reduction.

ACKNOWLEDGEMENTS

The authors would like to thank the Swedish Research Council FORMAS for research funding, in the framework of the collaborative international consortium STEEP-STREAMS financed under the ERA-NET Cofund WaterWOrks2014 Call. This ERA-NET is an integral part of 2015 Joint Activities developed by the Water Challenges for a Changing World Joint Programme Initiative (Water JPI). The authors thank Alberto Montanari for discussion on previous versions of our work.

REFERENCES


Kleméš, V. 1986 Dilettantism in hydrology: transition or destiny? *Water Resources Research* 22 (9S), 177S–188S. https://doi.org/10.1029/Wr022i09Sp0177S.


Madsen, H., Pearson, C. P. & Rosbjerg, D. 1997 Comparison of annual maximum series and partial duration series methods...


