German tanks and historical records: the estimation of the time coverage of ungauged extreme events.

Ilaria Prosdocimi *

February 21, 2017

Abstract

The use of historical data can significantly reduce the uncertainty around estimates of the magnitude of rare events obtained with extreme value statistical models. For historical data to be included in the statistical analysis a number of their properties, e.g. their number and magnitude, need to be known with a reasonable level of confidence. Another key aspect of the historical data which needs to be known is the coverage period of the historical information, i.e. the period of time over which it is assumed that all large events above a certain threshold are known. It might be the case though, that it is not possible to easily retrieve with sufficient confidence information on the coverage period, which therefore needs to be estimated. In this paper methods to perform such estimation are introduced and evaluated. The statistical definition of the problem corresponds to estimating the size of a population for which only few data points are available. This problem is generally refereed to as the *German tanks problem*, which arose during the second world war, when statistical estimates of the number of tanks available to the German army were obtained. Different estimators can be derived using different statistical estimation approaches, with the maximum spacing estimator being the minimum-variance unbiased estimator. The properties of three estimators are investigated by means of a simulation study, both for the simple estimation of the historical coverage and for the estimation of the extreme value statistical model. The maximum spacing estimator

*Department of Mathematical Sciences, University of Bath, UK. Email: prosdocimi.ilaria@gmail.com

is confirmed to be a good approach to the estimation of the historical period coverage for practical use and its application for a case study in Britain is presented.

1 Introduction

Natural hazards like floods, sea surges or earthquakes are some of the most dangerous threats both to human lives and infrastructures. Throughout history, strategies to manage the risks connected to natural hazards have been devised, and still at present these risks cannot be eliminated, but must be managed and planned for. A key step in the management of risks is the estimation of the frequency of events of large magnitude, which is needed to assess the likelihood of severe damages happening in specific areas. However by definition, very large events happen rarely and there are consequently few records available to perform such estimation. This is particularly true when the estimation is based on systematic measures of the process of interest, which might cover a period of time much shorter than the time scale at which one would imagine to actually record very rare events, such as events happening less frequently than once every 100 years. The statistical models typically used to estimate the frequencies of rare events are based on extreme value theory, which provides some general asymptotic results on the behaviour of events of great magnitude. Moreover the methods generally used in the estimation procedure make an attempt to use as much data as is available. For example regional methods, which pool together the information of a large number of stations are used to estimate the frequency of large storm surges (Bernardara et al., 2011) and floods (Hosking and Wallis, 1997). Alternatively, it would be possible to augment the data available at the time of analysis by including not only the systemically measured data, but also additional data from past events on which some information is still available from historical records or evidence in the landscape. This approach has been shown to greatly reduce the uncertainty of estimates at sites of interest for different natural hazards like coastal water levels (e.g. Bulteau et al., 2015), volcanic eruptions (e.g. Coles and Sparks, 2006) and peak river flow (e.g. Macdonald et al., 2014). This study gives some results on the estimation of a specific quantity needed when fitting statistical models on data series of historical and systematic records, namely the time of coverage of the historical record. Although this quantity can often be retrieved in

the investigation which leads to the construction of the historical record, it is in some cases unknown, so that an estimated value needs to be used instead. In the remainder of this paper the focus will be the use of historical records in flood frequency estimation applications, but the results could be useful in any situation in which historical data would be used to improve an estimate of the frequency of rare events and no clear information can be retrieved on the actual historical period covered by the non-systematic data. Statistical models for inference are presented in Section 2, with extreme value modeling briefly discussed in Section 2.2. Section 3 introduces the model used to include historical data in extreme value frequency estimation, while the different estimators for the coverage of the historical record are presented in Section 4. The performance of these estimators is investigated by means of a simulation study in Section 5 while Section 6 shows an application for the assessment of the rarity of large floods for a gauging station in the UK which recently experienced a record breaking event. Finally, Section 7 gives a brief summary and discussion of the results.

2 Statistical Models

Broadly speaking, statistical inference aims at characterising the behaviour of a process of interest using some relevant sample of data. It is typically assumed that the available sample is representative of the process of interest (e.g. large floods) so that it can be used to infer properties of the stochastic distribution of the process. It is generally assumed that the process under study follows a certain known distribution f, parametrised by some parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ whose values are unknown. Finding estimates of the distribution parameters $\boldsymbol{\theta}$ gives a full description of the behaviour of the process under study.

In the simplest case it is assumed that each element x_i in the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a realisation of independent and identically distributed (i.i.d.) random variables X_i, whose probability distribution function is a certain $f(x_i, \boldsymbol{\theta})$. In the following subsection some methods commonly used to estimate the parameter vector $\boldsymbol{\theta}$ are discussed. These methods have long been established and are discussed in most introductory book to statistical inference (e.g. Rice, 2006); only some basic details are provided here as a reference.

2.1 Statistical parameter estimation approaches

Maximum Likelihood

A very common method used to estimate $\boldsymbol{\theta}$ is maximum likelihood, which, under some conditions, provide asymptotically unbiased and efficient estimators. Maximum likelihood estimates are obtained as the $\boldsymbol{\theta}$ values which maximise the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{x})$, defined as

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{i=1}^{n} f(\boldsymbol{\theta}, x_i).$$

The ML estimate $\hat{\boldsymbol{\theta}}_{ML}$ can be thought of as the value of $\boldsymbol{\theta}$ which make the data more likely to have happened under the assumed distribution. In some cases the $\boldsymbol{\theta}$ which maximise the likelihood function can be found analytically, but in many applications numerical methods are used to maximise $L(\boldsymbol{\theta}; \boldsymbol{x})$ and find the estimated values $\hat{\boldsymbol{\theta}}_{ML}$.

Method of moments

Another very intuitive and commonly used approach for the estimation of $\boldsymbol{\theta}$, is the method of moments, in which the parameters are first expressed as functions of the distribution moments (e.g. $\mu_1 = E[X]$, $\mu_2 = E[X^2]$, and so forth) and then directly estimated by plugging in the sample estimates of the moments (e.g. $\hat{\mu}_1 = \sum_{i=1}^{n} x_i$, $\hat{\mu}_2 = \sum_{i=1}^{n} x_i^2$, and so forth). For example, the mean and variance of a normal distribution $X \sim N(\mu, \sigma)$, can be expressed as $\mu = E[X]$ and $\sigma^2 = Var[X] = E[X^2] - (E[X])^2$. Method of moment estimates are then obtained as $\hat{\mu}_{Mom} = \hat{\mu}_1 = \sum_{i=1}^{n} x_i$ and $\hat{\sigma}_{Mom} = [\hat{\mu}_2 - (\hat{\mu}_1)^2]^{1/2} = [\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2]^{1/2}$. Method of moments estimates $\hat{\boldsymbol{\theta}}_{Mom}$ do not enjoy the optimal asymptotic properties of ML estimates, but can be shown to be consistent and are computationally easy to derive.

Maximum spacing method

A less widespread, but also useful, inference approach is the maximum spacing method introduced simultaneously with a different naming and a different reasoning by Cheng and Amin (1983) and Ranneby (1984). Defining the ordered sample $(x_{(1)}, \ldots, x_{(n)})$ such that $x_{(1)} < x_{(2)} < \ldots < x_{(n-1)} < x_{(n)}$, the spacing between the cumulative distribution functions

of successive points is taken to be

$$D_i(\boldsymbol{\theta}) = F(\boldsymbol{\theta}, x_{(i)}) - F(\boldsymbol{\theta}, x_{(i-1)}), \quad i = 1, \ldots, (n+1)$$

taking, for convenience, $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$. The maximum spacing estimator is defined as the value $\hat{\boldsymbol{\theta}}_{MSP}$ which maximises

$$S(\boldsymbol{\theta}) = \frac{1}{n+1} \sum_{i=1}^{n+1} \ln D_i(\boldsymbol{\theta}) = \frac{1}{n+1} \sum_{i=1}^{n+1} \ln(F(\boldsymbol{\theta}, x_i) - F(\boldsymbol{\theta}, x_{(i-1)})).$$

The estimate $\hat{\boldsymbol{\theta}}_{MSP}$ can be thought of as the value of $\boldsymbol{\theta}$ which makes the distribution of the estimated cumulative distribution function (cdf) as close as possible to the Uniform(0,1) distribution, which is how the cdf of a i.i.d. sample is expected to behave. The maximum spacing method can give valid results in cases for which the likelihood approach fails and is shown to be consistent.

All the above methods have been developed under the general framework in which it is assumed that the available sample is representative of an existing parent distribution parametrised by some true parameters $\boldsymbol{\theta}$ whose values need to be estimated. Another very popular approach to statistical inference is the Bayesian approach in which it is assumed that the distribution parameters are also random variables, and that the aim of the inference is to characterise the distribution of these random variables given the available sample. The method is not discussed further in the paper, but its use is widespread in statistical applications and should be mentioned, in particular given its wide use for the estimation of extreme value models in presence of historical data.

2.2 Statistical models for the frequency of extremes events

Most statistical applications aim at describing the behaviour of the central part of the distribution of the process under study. It is often the case though, that it is not the typical behaviour of the process that is of interest, but its tail behaviour, i.e. the rarely observed events. When the interest of the estimation lies in the frequency of extreme events it is common practice to use only a subset of the available data which is actually informative of the behaviour of the tail of the distribution rather than its central part. A frequently used approach is to only use the maximum value of the measured process in a block, for example a year or another fixed period

of time. The block maxima are assumed to follow some appropriate long-tailed distribution, with the Generalised Extreme Value (GEV) distribution being motivated by the asymptotic behaviour of maxima of stationary processes (see Coles, 2001). The GEV is often used in practice when investigating the frequency of rare events, although other distributions have been proposed in some cases as discussed in Salinas et al. (2014). The Generalised Logistic (GLO) distribution, for example, has been shown to provide a better goodness of fit for samples of British peak flow annual maxima (Kjeldsen and Prosdocimi, 2015) and the Pearson-Type-III distribution is frequently used when modelling peak flow values of basins in the USA (U.S. Interagency Advisory Committee on Water Data, 1982). Once a decision is made on the appropriate form of $f(x, \boldsymbol{\theta})$ to represent the distribution of the data, and the values of $\boldsymbol{\theta}$ are estimated, the magnitude of the events which are expected to be exceeded with a certain probability p can be derived via the quantile function $q(1-p, \hat{\boldsymbol{\theta}})$. Conversely, it is possible to obtain an estimate of the frequency at which an event of magnitude $\tilde{x}$ is expected to be exceeded via the cumulative distribution function $F(\tilde{x}, \hat{\boldsymbol{\theta}})$. In practice, since only a subset of a record is used in the estimation of the frequencies of extreme events, samples tend to be relatively small and long observations are needed to obtain large samples of annual maxima. For example, gauged flow records in the UK tend to be less than 40-year long (see Kjeldsen and Prosdocimi, 2016), which means that samples of less than 40 units would be used in the estimation of the frequency of rare events when annual maxima are analysed. The review carried out in Hall et al. (2015) indicate that records throughout Europe are of similar length. Given that typically the interest is in the estimation of events which are expected to be exceeded at most every 100-year, there is a large difference between the available information and the target of the estimation. Several strategies, aiming at augmenting the available information, have been developed. A popular approach is to somehow pool together information across different series: this is referred to as the regional approach and has been widely used in flood frequency applications following, for example, in the work of Hosking and Wallis (1997). The justification for the regional approach is that, given that series only cover a short period of time, one can trade space for time and augment the available information by combining different stations. The idea of augmenting the information used in the inference process pooling is also used in probabilistic regional envelop curves, which pool together information on extreme events and are used to estimate exceedance probabilities for homogeneous regions (see for example Lam

et al., 2016) Finally methods which are less reliant on the theoretical statistical properties of the peak flow process, but make use of the understanding of hydrological processes are often used. For example rainfall-runoff models use information on the catchment to provide estimates of the entire hydrograph, for rainfall events of given rarity. ReFEH (Kjeldsen, 2007) is the model used in the UK within the Flood Estimation Handbook, but several other models are proposed in the literature. In general, when estimating flood frequency curves, it would be ideal to use as much knowledge as possible about the site for which the estimation is carried out, combining both the hydrological knowledge of the analysis and using all available data in the best possible way. This is strongly advocated in a series of companion papers by Merz and Blöschl (2008a,b) and Viglione et al. (2013), which showcase the usefulness and importance of combining different sources of information to improve the accuracy of flood frequency estimation. A similar message is also found in Environment Agency (2017), which showcase how the use of catchment-specific information can improve the quality of the estimation of flood risk. The usage of information on past large event, for example, is often suggested as a way to improve inference about flood risk. Indeed, historical data can be used to extend the length of time covered by the available series, thus diminishing the discrepancy between the estimation horizon and the amount of data used in the estimation. These type of events would not have been gauged using the modern-day technology, but would nevertheless be informative of the size of very large events which happened in the past. The usefulness of including historical data in flood frequency analysis has long been recognised (e.g. Hosking and Wallis, 1986; Stedinger and Cohn, 1986). Different methods to combine historical and systematic data have been proposed (e.g. Cohn et al., 1997; Gaume et al., 2010), historical flow series have been reconstructed for several river basins (see among others Macdonald, 2014; Elleder, 2015; Machado et al., 2015, in a recent HESS special issue) and several countries in Europe at present recommend that evidence from past floods is included when estimating the magnitude of rare flood events (Kjeldsen et al., 2014). The case study in Section 6 gives some discussion of the possible difficulties and advantages of using historical data in flood frequency estimation for a specific location in the UK. The standard framework to include historical data builds on the construction of the likelihood outlined in Section 3.

3 The inclusion of historical data for frequency estimation

Assume that a series of gauged annual maxima $\boldsymbol{x} = (x_i, \ldots, x_n)$ is available and that additionally some information on the magnitude of k historical events $\boldsymbol{y} = (y_1, \ldots, y_k)$ pre-dating the systematically recorded observations is also available. It is assumed that all k events are bigger than a certain value X_0, which is referred to as perception threshold, since it corresponds to a magnitude above which events would have been large enough to leave visible marks in the basin or be worthy of being recorded for example in diaries, local newspapers or as epigraphic marks in urban developments. Further, it is assumed that the underlying process generating the extreme events in the past and in the present day can be modelled using the same distribution X with pdf $f_X(x, \boldsymbol{\theta})$ and cdf $F_X(x, \boldsymbol{\theta})$. One important assumption that is made is that all events above X_0 in the period of time covered by the historical information, denoted by h, are available. The different quantities involved in the inclusion of historical data are exemplified in Figure 1 which shows the systematic and selected historical data for the Sussex Ouse at Lewes case study described in Macdonald et al. (2014). The number of historical events k can then be thought of as a realisation of a Binomial distribution $K \sim Bin(h, p)$, with $p = P(X > X_0) = [1 - F_X(X_0)]$. Finally, by taking $f(y) = f_X(y|y \leq X_0)P(y \leq X_0) + f_X(y|y > X_0)P(y > X_0)$ and reworking some of the formulae (see Stedinger and Cohn, 1986) the likelihood for the combined sample of historical and gauged records $(y_1, \ldots, y_k, x_1, \ldots, x_n)$ can be written as

$$L(\boldsymbol{x}, \boldsymbol{y}, h, k; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i, \boldsymbol{\theta}) \binom{h}{k} F_X(X_0)^{(h-k)} \prod_{j=1}^{k} f_X(y_j, \boldsymbol{\theta}). \quad (1)$$

Numerical methods are generally used to maximise the above likelihood and the use of Bayesian methods has extensively been advocated for this type of applications (e.g. Parent and Bernier, 2003; Reis and Stedinger, 2005; Neppel et al., 2010). As discussed in Stedinger and Cohn (1986) the likelihood in Equation (1) can be modified when only the number of historical events and not their magnitude can be ascertained with sufficient confidence, but this case is not explored in the present work.

A number of features on the historical data are required in Equation (1), namely h, k and $\boldsymbol{y}$, and these are assumed to be correctly specified. In particular it is assumed that the period of

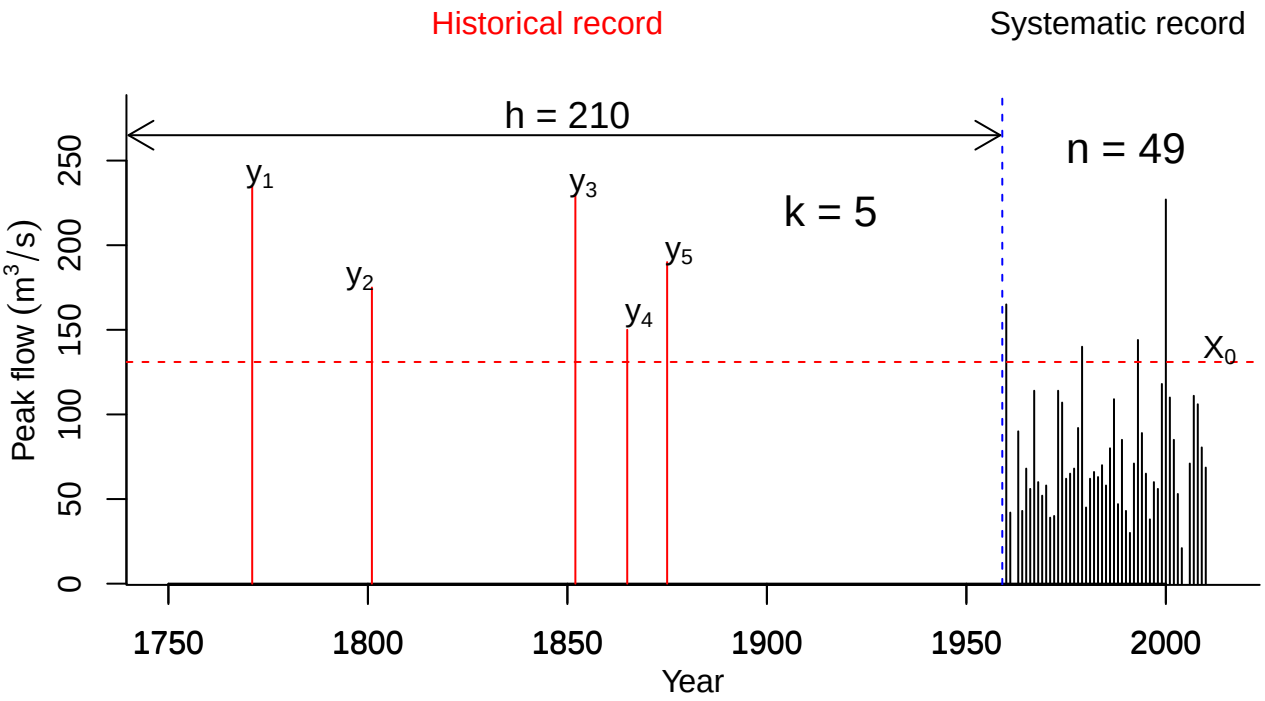

Figure 1: Display of the quantities involved when combining historical and systematic records using the Sussex Ouse at Lewes data from Macdonald et al. (2014).

time covered by the historical information h is correctly known: this paper discusses methods to estimate h when it can not be accurately quantified from the historical information. The impact of the value of h on the final estimation outcome can be seen in Figure 2 where the different estimated flood frequency curves obtained using a range of h values and the Sussex Ouse data shown in Figure 1 are shown. Using different values of h can have a noticeable effect of the estimated flood frequency curves, in particular the magnitude of rare events would be estimated very differently depending on which value of h is used. Note that some of the values of h in Figure 2 are of course not possible given the historical record for the station: results for values of h smaller than 190 year are given as reference and they correspond to the case in which the historical events would have all happened in the years just before the beginning of the systematic record. The importance of correctly assessing the value of h is discussed in Hirsch (1987) and Hirsch and Stedinger (1987), which indicate that biases can be introduced in the assessment of the of extreme events if the wrong value of h is used, and Bayliss and Reed (2001) state that no guidelines appear to be available on how to correctly asses a realistic period of record to historical information. This is an indication that the issue of the correct identification of h has been given little attention in the large literature on the use of historical

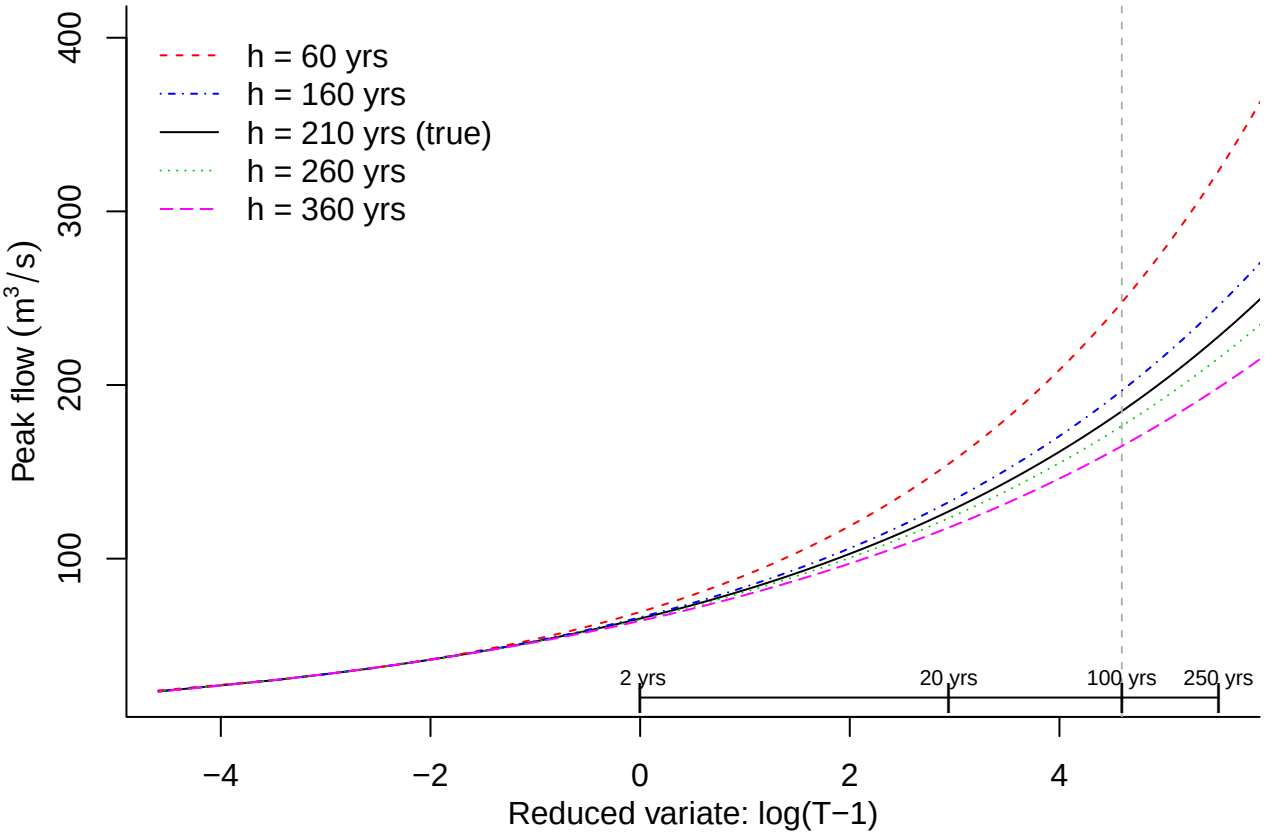

Figure 2: The impact of using different values of h on the estimated flood frequency curve.

data and in several studies which combine historical and systematic data it is unclear whether a realistic value of h could be determined in the retrieval of the historical information and which value of h is effectively used in the estimation. It is often the case that the value of h is taken to be the time between the first historical record available and the beginning of the systematic record. The drawbacks of this approach are discussed later in the paper, and have been already pointed out in Strupczewski et al. (2014), which is to the author's knowledge, the only effort to give guidance on how to obtain reliable values of h since the review by Bayliss and Reed (2001).

Finally some cautionary warnings on the routine inclusion of historical data in flood frequency estimation should be given. An important assumption that is made in the estimation procedure is that all the information in sample, i.e. both the historical and the systematic peaks, comes from the same underlying distribution. That is to say that the process from which the high flows are extracted is stationary throughout time. In simpler words, it is assumed that all peaks are somewhat representative of the flood risk at the present time, and that flood risk is unchanged throughout the record. Given the natural and anthropogenic

changes in climate and the potential impacts that changes in the catchment would have on flood risk, this is indeed a very strong requirement and assumption. The significance of potential non-stationarities driven from climatic variations and man-made changes to the river basin can be derived by means of statistical models, but to be reliable and representative of large scale changes these would need to be based on long records, whether from instrumental measurements (Mudersbach et al., 2015), or from a mixture of instrumental measurements and historical records as in Machado et al. (2015). Again, combining all available knowledge about the history and properties of the river basin under study allows for a more precise characterisation of risk in the area of interest and how this changes in relationship to anthropogenic changes and natural variability. See for example Silva et al. (2015, 2016) for an analysis which combines different sources of information to assess how flood risk change as a function of several explanatory variables. Further, methods to include uncertain values of for historical peak flows values have been widely employed (see for example Viglione et al., 2013; Gaume et al., 2010) and these can be used to acknowledge that the value of the past peaks corresponds to a range of possible values for the flow as it would have been recorded in the present time.

4 The estimation of h

To estimate the parameter h it is assumed that some reliable information on the timing of the historical events $\boldsymbol{y}$ is available. The differences between the time of occurrence of the historical events and the start of the systematic records are denoted as $\boldsymbol{t} = (t_1, \ldots, t_k)$: each value of t_i corresponds to the number of years between each historical event y_i and the onset of the systematic records. Since annual maxima are assumed to be independent the exceedance of the high perception threshold can happen in any year of the historical periods with equal probability. Each t_i can then be seen as a realisation of a uniform distribution with lower limit equal to 1 and an unknown upper limit h: $T \sim U(1, h)$. Alternatively, one could see the sample $\boldsymbol{t}$ as a random draw without replacement of k elements from the population of past annual maxima which happened in the years $(1, \ldots, h)$. The estimation of h would then correspond to the estimation of the size of the population of annual maxima from which the sample $\boldsymbol{t}$ is extracted. This problem corresponds to the so called German tanks problem, which arised during World War II when an estimate of the total number of German tanks and warfare

was obtained based on the serial number of the captured items. As discussed in Ruggles and Brodie (1947) the statistical estimates of the number of weapons and components available to the German army proved to be more accurate than the numbers deducted by intelligence. Johnson (1994) presents a series of possible estimators of h derived on the population size characterisation of the problem, listing their expected values and variances. The same estimators, derived using the Uniform distribution characterisation, are presented below with an indication of their advantages and issues.

Maximum likelihood

Assuming that $T \sim U(1, h)$, the likelihood function to be maximised to estimate h corresponds to:

$$L(\boldsymbol{t}, h) = \prod_{i=1}^{k} f(t_i, h) = \begin{cases} (h-1)^{-k} & \text{for } 1 < t_{(1)} < \ldots < t_{(k)} < h \\ 0 & \text{otherwise} \end{cases}$$

so that the maximum likelihood estimator of h, $\hat{h}_{ML}$, corresponds to the largest value of the sample for which the likelihood has a positive value: $\hat{h}_{ML} = t_{(k)} = \max(t_1, \ldots, t_k)$. In other words, the estimated time span for which historical information is available is estimated to be starting at the time at which the first historical event is recorded. The ML estimate h_{ML} can be shown to be biased, as $E[\hat{h}_{ML}] = hk/(k+1) + k/(k+1)$.

Method of moments

The Method of moments estimator of the upper bound h of a uniform distribution $T \sim U(1, h)$ can be derived knowing that $E[T] = (h+1)/2$, so that $h = 2E[T] - 1$. Taking $\bar{t} = \sum_{i=1}^{k} t_i$, the average time before the start of the systematic record at which historical events happened, the Method of Moments estimator can be written as $\hat{h}_{Mom} = 2\bar{t} - 1$. The estimator h_{Mom} is unbiased since $E[\hat{h}_{Mom}] = h$. Notice though that in practice the value of $\hat{h}$ might be a non-integer and might be smaller than the maximum value observed, $t_{(k)}$. The first issue is easily fixed by rounding $\hat{h}_{Mom}$ to the nearest integer, and one could take the estimate of h to be the maximum between $\hat{h}_{Mom}$ and $t_{(k)}$, but it is undesirable for an estimator to produce results that are not possible for a given sample.

Maximum spacing method

The maximum spacing estimator of the upper bound h of a uniform distribution $T \sim U(1, h)$

can be derived as the value which maximises the function

$$S(h, \boldsymbol{t}) = \{\ln(t_{(1)}-1)-\ln(h-1)+\sum_{i=2}^{n}\ln(t_{(i+1)}-t_{(i)})-(n-1)\ln(h-1)+\ln(h-t_{(n)})-\ln(h-1)\}.$$

The estimator $\hat{h}_{MSP}$ is then found to be $\hat{h}_{MSP} = t_{(k)}(k+1)/k - 1 = t_{(k)} + t_{(k)}/k - 1$. Note that this corresponds to taking the maximum value of $t_{(k)}$ and add the average gap between the observed timings. The expected value of $\hat{h}_{MSP}$ is $E[\hat{h}_{MSP}] = h$. For the case at hand, the MSP estimator can also be shown to have minimal variance (see. Johnson, 1994), and should therefore be the preferred estimator. In the case in which only one observation is available, $k = 1$, the h_{Mom} and h_{MSP} estimator are identical and their form corresponds to the one already presented in Strupczewski et al. (2014).

The different estimators of h correspond to different approaches that hydrologists could use when including historical data in flood frequency estimation: the three approaches are listed in Section 4.4.3. of Bayliss and Reed (2001) where they are presented using common sense reasoning rather than a statistical framework. Using the time of the first historical event as an indicator of the whole period of coverage of the historical record corresponds to using the maximum likelihood estimator of h: this is relatively easy to apply, but it has been shown to give less reliable results. Indeed if a large historical record was recorded at a distant time $t_{(k)}$, it would be unlikely that a similarly large event would have happened the year before, so it would be reasonable to shift the starting period of the historical record to a earlier date. Taking the starting point to be the point in time that precedes the first event by the average time between the historical events corresponds to the maximum spacing estimator. Finally it seems plausible to think that the amount of time passed between the start of the historical coverage and the first historical record should be the same as the amount of time between the first historical record and the starting of systematic record: when $k = 1$ this corresponds to the method of moments estimator.

For the Sussex Ouse data presented in Figure 1 the $\boldsymbol{t}$ sample is found to be equal to $\boldsymbol{t} = (85, 95, 108, 159, 189)$, and the known value of h is 210. If h was unknown the different estimates would correspond to: $\hat{h}_{ML} = 189$, $\hat{h}_{Mom} = 220.6$ and $\hat{h}_{MSP} = 225.8$. The impact of using the different estimated values can be guessed by comparing the flood frequency curves in Figure 2.

The performance of the different estimation approaches for the estimation of h in practice

is investigated by means of a simulation study in the next Section. The impact of the different estimation approaches within the wider scope of return curve estimation when combining historical and systematic data is further investigated within the simulation study.

5 Simulation study

A simulation study is performed to investigate the performance of the different methods in estimating the value of h and successively the impact of estimating h on the overall performance of flood frequency estimation when augmenting systematic data with historical information. The simulation study is designed to be representative of possible data availability situations in real applications and realistic distributional assumptions based on observed characteristics of British peak flow data are used. The parent distribution for the synthetic data generated in the simulation study is taken to be a GLO, which is the default distribution for British peak flow data, with location, scale and shape parameter taken to be, respectively, equal to 33, 6.5 and -0.3, approximately the median values of the at-site estimates of the 960 stations included in the National River Flow Archive (NRFA) peak flow dataset v3.3.4 (`http://nrfa.ceh.ac.uk`). Samples of systematic sample size n equal to 20, 40 and 80 are generated sampling from the known parent distribution. The true historical period covered by the historical information h is taken to vary among the values of 200, 400 and 800 years. In the data generation procedure for the simulation study, exactly h data points using the same parent distribution of the systematic data, are generated and the largest k points which are also larger than the defined perception threshold are taken to constitute the historical information used in the estimation procedure. The values of k are taken to be 1, 3, 5 and 10. For each combination of k and h values the perception threshold X_0 is taken to be the $(1 - k/h)^{th}$ quantile of the parent distribution, which is to the say the value above which one would expect to observe k values over h years. A total of 36 combinations of parameters are included in the study, to allow a full investigation of the impacts of different properties on the final performance of the estimation procedures. For each of the total 36 combinations, 10000 samples of historical and systematic data are generated and analysed: different estimation procedures are applied to estimate h and these estimates are then plugged in the methods to estimate the distribution parameters discussed in Section 3 which finally allowes to estimate the magnitude of rare events for each generated

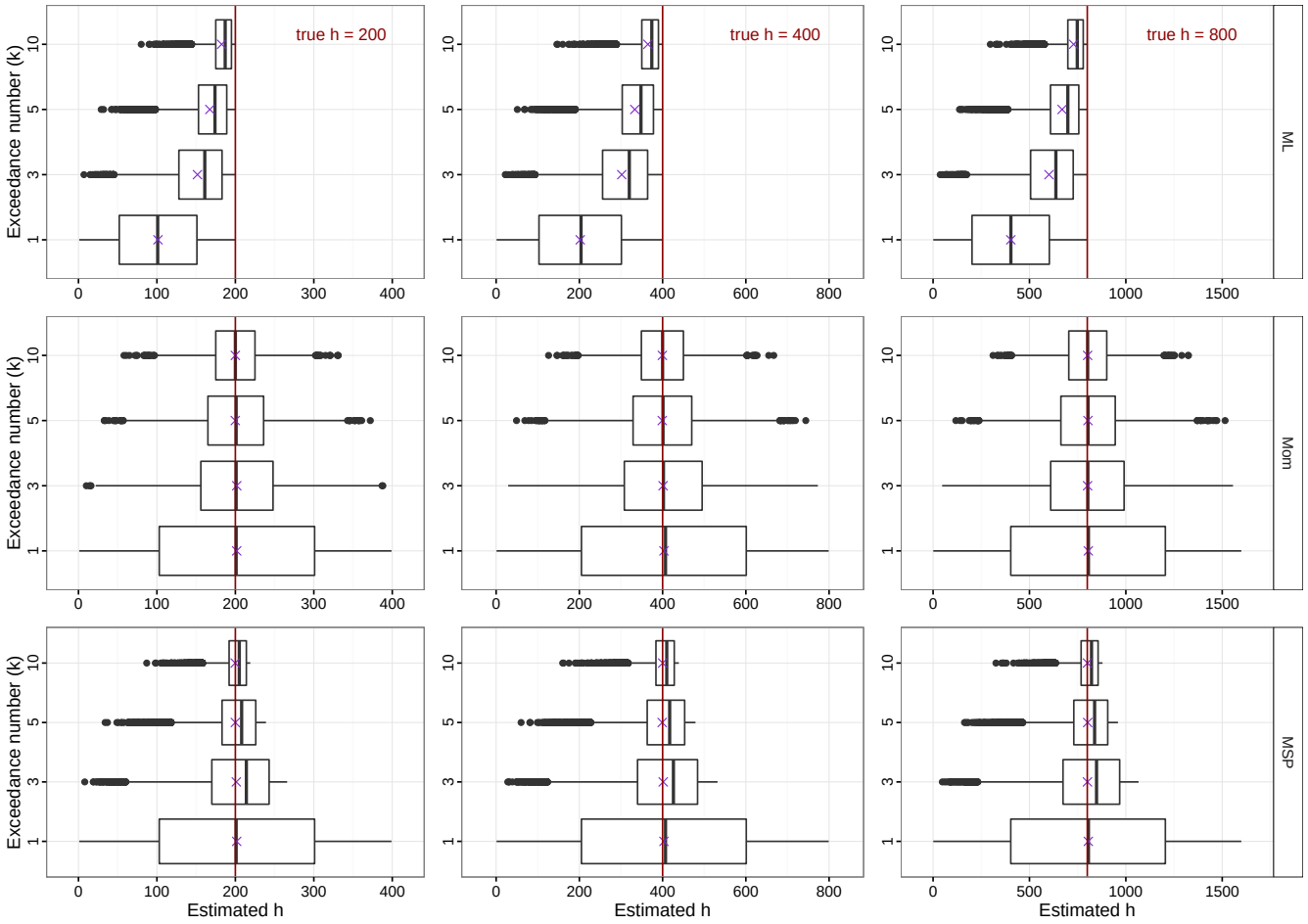

Figure 3: Boxplots showing estimates of h for different values of k: each row shows the estimates obtained using the three proposed methods, while each column shows the results for different true values of h. The true value of h is indicated as a vertical line in each plot and indicated in the text. The violet cross in each boxplot indicates the average value of the estimates

sample.

At first the ability of the different estimators presented in Section 4 to estimate h is assessed. Figure 3 shows boxplots of the estimated values of h using the maximum likelihood (ML) approach, the method of moments (Mom) and the minimum spacing (MSP) approach for different values of k (e.g. the number of historical events in the sample) and h (the historical period cover which corresponds to the quantity being estimated). Also shown in the Figure is the true value of h. It is clearly visible that for all estimation procedure the variability of the estimators decreases for increasing values of k: this is not surprising, as a larger sample (i.e. more information) is used in the estimation. Another remarkable feature is that the ML estimates are indeed biased, while both the MSP approach and the method of moments give unbiased estimates, with the MSP approach estimator being less variable, as expected

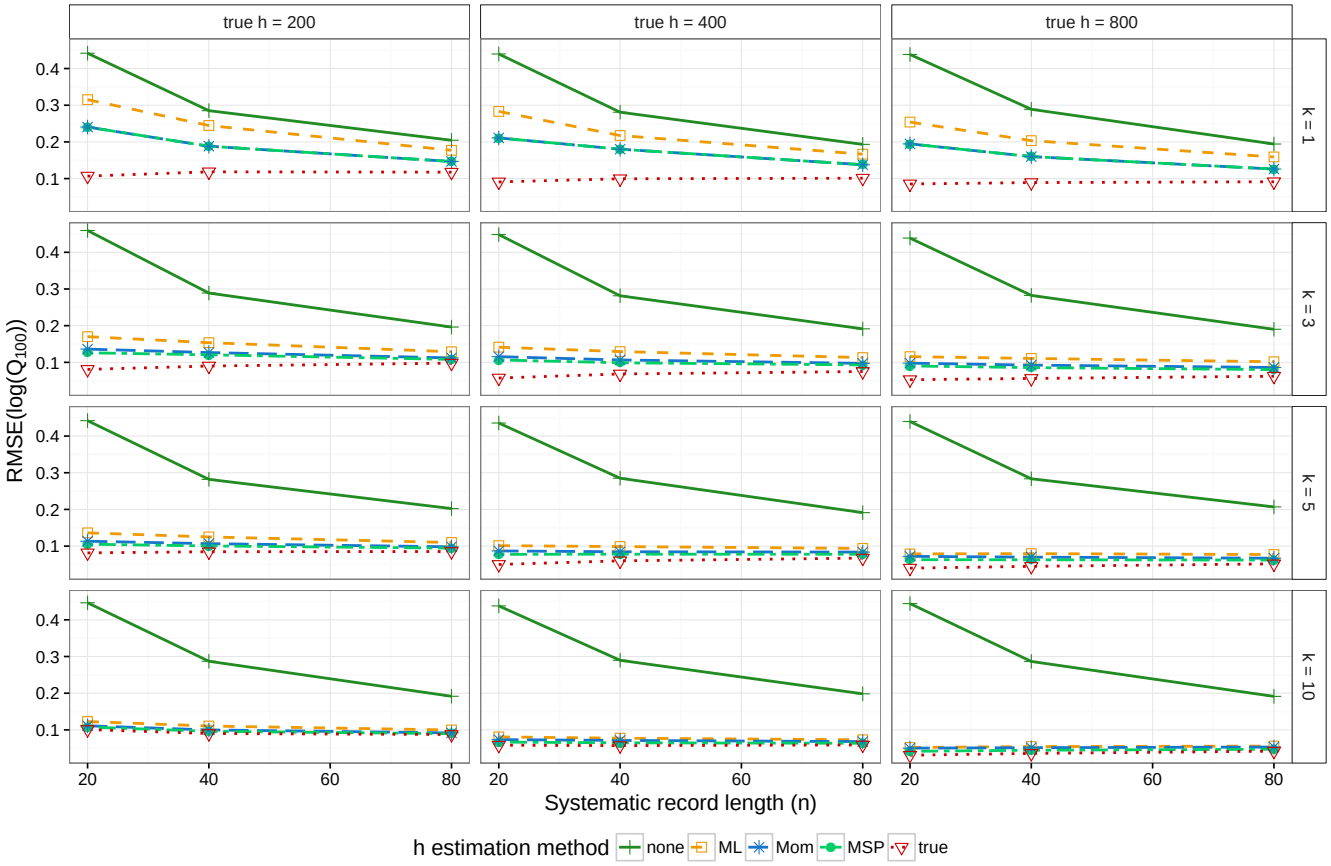

Figure 4: RMSE for $\log(Q_{100})$ for different values of k and h. Each line shows the RMSE for $\log(Q_{100})$ using different approaches to estimate the parameter h.

from the theoretical result. The asymmetric shape of the boxplots for the ML and MSP estimates is consistent with the behaviour of the distribution of the maximum shown in Johnson (1994), while the MOM estimates, which are based on the average value of samples from a uniform distribution, exhibit a more symmetric behaviour, which is not surprising under the Central Limit Theorem. The properties of the estimation procedures is not impacted by the actual value of h, and the MSP approach consistently gives unbiased estimated with smaller variability. When $h = 1$ the boxplots for the MSP approach the method of moment are identical, since the two estimators have the same form. It would then appear that in cases where no reliable information can be retrieved on the real value of h the MSP approach should be used to estimate the length of time covered by the historical data.

Nevertheless, when including historical data in flood frequency analysis, the aim is generally to estimate the parameters of the flood distribution and to then obtain estimates for its upper quantiles. The performance of this inference when using the different estimators for h are explored in Figure 4 and Figure 5. Figure 4 shows the RMSE values of the $\log(Q_{100})$ for each

combination of systematic record length (n), number of historical record (k) and historical period length (h). Each line shows the RMSE values obtained using a different approach to estimate h, including the solid line which corresponds to the case in which only systematic data are used and no estimation of h takes place and the case in which the true value of h is used (dotted line and triangle). The RMSE for any estimated quantity τ, either a parameter of a property of the distribution, is defined as the square root of the expected value of the squared difference between the estimated value $\hat{\tau}$ and its true value. In a simulation study with s synthetic data, the $RMSE(\hat{\tau})$ can be determined taking:

$$RMSE(\hat{\tau}) = \frac{1}{s}\sqrt{\sum_{i=1}^{s}(\hat{\tau}_i - \tau)^2}.$$

Low values of RMSE indicate that the estimated value do not vary much around the true value of τ, giving an indication of a good performance in estimation. Figure 4 shows that more precise estimates can be obtained for high quantiles when using historical data, compared to when using systematic data only, even when there is some uncertainty on the actual time covered by the historical data. Further it would appear that when including historical data, if the MSP approach or the method of moments are used to estimate h the RMSE values for the Q_{100} are similar to the one obtained when the true value of h is used even if only $k = 3$ historical events are available.

To investigate the impact of the different methods of estimating h on the flood frequency model estimation the RMSE for the shape parameter of the GLO distribution are shown in Figure 5. The shape parameter plays an important role in extreme value modeling and its estimate are generally quite variable due to the limited sample sizes normally available for estimation. Figure 5 shows that, when the parameter h is estimated, both the method of moments and the MSP approach lead to similar reductions in RMSE. The use of the $\hat{h}_{ML}$ estimator is heavily discouraged, although when a large number of historical events are available it appears to still give, on average, a good improvement on the quality of the estimation. Interestingly, from both Figure 4 and 5 it can be concluded that it is not necessary to have a very large amount of historical events to obtain large improvements in the precision of flood frequency analysis. Another interesting aspect to notice from Figure 4 is that, when knowing the real value of h, the decrease in RMSE is already very large when including just one historical event for an historical period of $h = 200$ year. There is also an improvement in the estimation

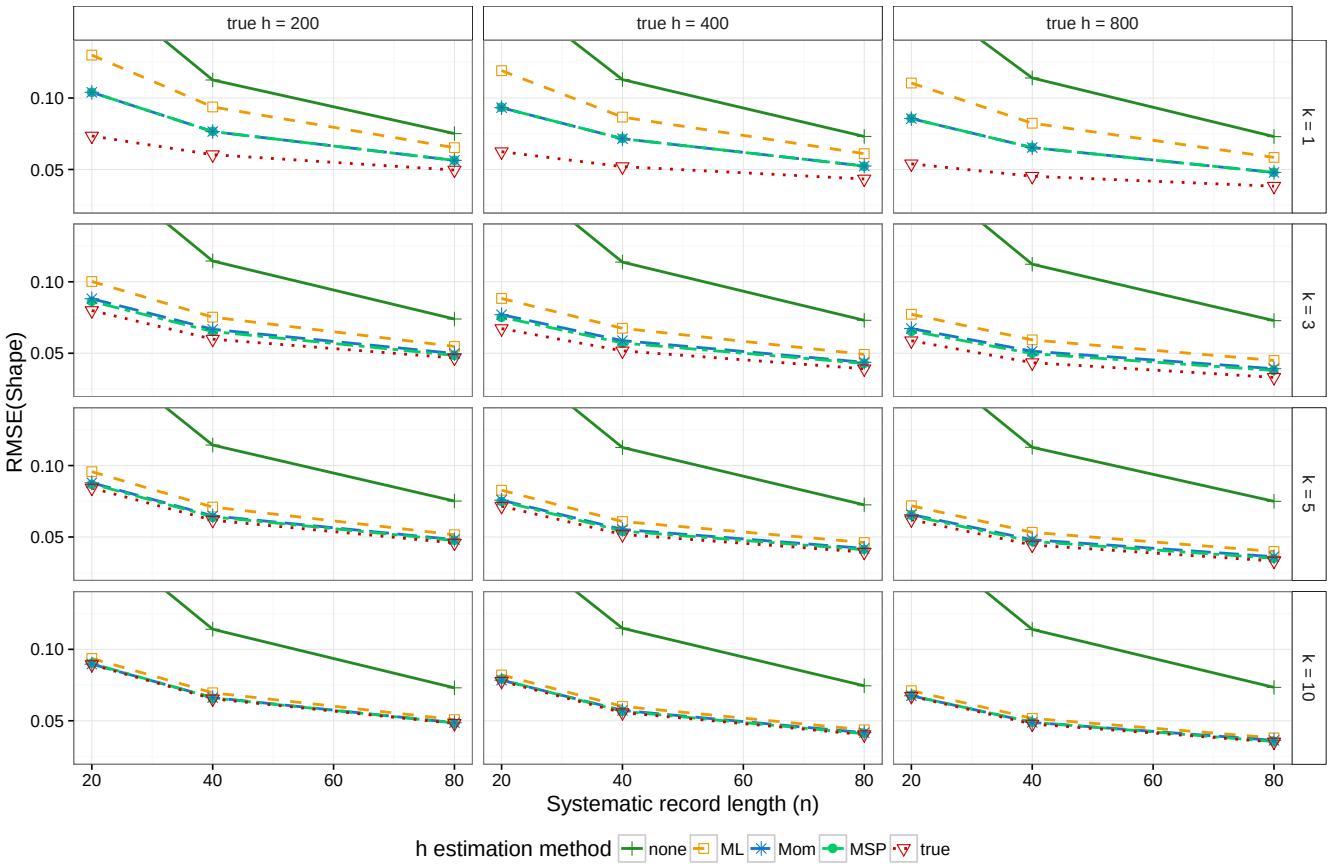

Figure 5: RMSE for the shape parameter for different values of k and h. Each line shows the RMSE for the shape parameter when different approaches to estimate the parameter h are used. Values for the RMSE obtained when using systematic data only (solid line) are truncated to make the Figure more readable.

when longer historical periods are considered and more historical events are included in the estimation, but the RMSE does not decrease very much. In Figure 5 one can notice that the RMSE of the shape parameter obtained when using systematic data with a 80-year long record is comparable to the one obtained using a 20-year systematic record combined with $k = 3$ historical events covering a time span of $h = 200$ year. Thus, for practical purposes it might be very useful to retrieve even sparse knowledge on the recent past to improve the overall estimation.

6 The Lune at Caton and the December 2015 floods: an historical perspective

In December 2015 several large flood events were recorded in northern Britain as a result of extremely large storms (Desmond, Frank and Eva) which occurred after a period of substantial rain. The extent of the flooding was reported to be unprecedented and several record-breaking events were recorded (Met Office, 2016; Parry et al., 2016). Information on historical records of large floods in the proximity of some of the gauging stations involved in the Winter 2015 floods is readily available and could be used to put the recent floods into the historical context. A full appraisal of the rarity of the events recorded in Winter 2015 is out of the scope of this investigation, which aims at discussing some practical aspects of the inclusion of historical data for a specific location. Volume IV of the Flood Studies Report (FSR, Natural Environment Research Council, 1975) contains a long list of gauging stations for which some form of historical data is available, with historical series of extreme flow values available for several stations. In particular, a series of peak flow annual maxima between 1968 and 2013 is available for the site of the present day station measuring the peak flow of the Lune at Caton (NRFA Station 72004).

The Lune at Caton peaked at about 1700 m^3/s on December 5^{th} 2015 (Parry et al., 2016): this peak corresponds to the highest peak ever recorded at the gauging station, exceeding the previous record of 1395.222 m^3/s of January 1995, the highest peak registered in the 46 years of available data since the station started operating in 1967. Annual maxima in the UK are extracted as the highest peak recorded in a Water Year, which runs from October to

Date	Water Year	Peak Flow
02 Sep 1892	1891	977.000
26 Jan 1903	1902	1104.000
10 Feb 1920	1919	878.000
13 Nov 1923	1923	1119.000
21 Sep 1927	1926	906.000
03 Nov 1927	1927	1048.000
14 Feb 1936	1935	991.000
02 Dec 1954	1954	1161.000

Table 1: Historical peak flow values for the Lune at Caton listed in Volume IV of the FSR

September, thus the event of January 1995 would be listed as the maximum of Water Year 1994. The gauged peak flow records for this station, and all other stations in the UK, can be easily retrieved via the National River Flow Archive (NRFA) website. Volume IV of the FSR lists eight major annual maxima events recorded at the gauging station location in the years before the station started its regular recording in 1967 (see Table 1).

Beside the information on the peak flows, the following note is given:

Chapman & Buchanan - Frequency of Floods of normal maximum intensity in upland areas of Great Britain. ICE Symposium, River Flood Hydrology, 1965. Listed and ranked are discharges estimate at Caton Low mill, 2.5 miles upstream of Halton gauging station. The figures 'might be said to give a complete record of the very highest floods for a period of some 80 years...' [...].

It is therefore likely then that the historical record give information for the period starting in a year between 1875 and 1885. Note that the list of historical records was compiled some years before the gauging station started its operation, but it is very likely that no major event happened in the catchment in the few years between the creation of the list and 1967, otherwise this would have been noted in the FSR. For this station the estimates of h using the different methods can be derived from $\boldsymbol{t} = (13, 32, 40, 41, 44, 48, 65, 76)$ as: $h_{ML} = 76$, $h_{Mom} = 91.75$ and $h_{MSP} = 85.625$, which would lead to the historical coverage to start, respectively, in 1892, 1876 and 1882.

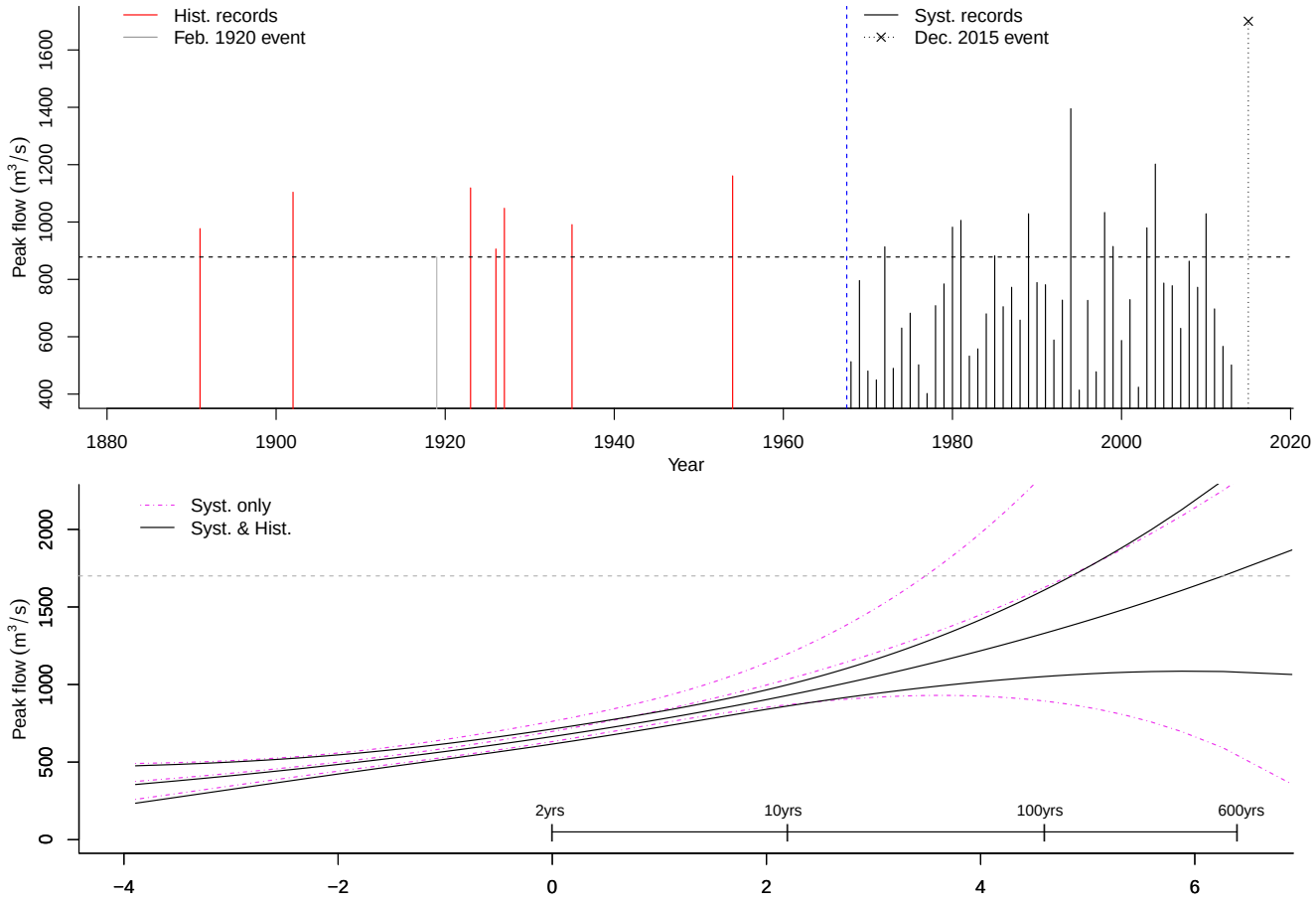

Figure 6: Upper panel: historical and systematic data combined. The perception threshold used in the analysis and the magnitude of the December 2015 flood are also shown. Lower panel: at-site estimates of flood frequency curve with 95% confidence intervals for the Lune at Caton using systematic data only and systematic data combined with historical records.

To use the historical information to estimate the flood frequency curve for the station a value for the perception threshold X_0 also needs to be identified. Nevertheless, from the text reported above it is not evident what perception threshold should be used in the estimation procedure. To eliminate possible subjective choices on which perception threshold to use, it was decided to take the lowest event in the historical record, the February 1920 event, as perception threshold and to only include the largest seven events in the record in the analysis. Indeed, if we are confident that the historical events capture accurately past large events we should be quite certain that all the highest seven events in the record are larger than the 1920 event. When only the highest seven historical events are used the estimated values of h correspond to $h_{ML} = 76$, $h_{Mom} = 90.86$ and $h_{MSP} = 87$: the difference compared to when using the full historical record is very small. The combined historical and systematic records

are displayed in the upper panel of Figure 6. The lower panel of Figure 6 shows the flood frequency curves obtained when using only the 46 years of systematic data available for the gauging station and the curve obtained when historical data are also included in the analysis using the estimated $h_{MSP} = 87$ value. Also shown in the Figure is a line which indicates the magnitude of the December 2015 flood (1700 m^3/s): this event was not included in the estimation procedures. For both curves the GLO distribution was assumed to be the underlying distribution for the high flow process and 95% confidence intervals are derived by means of the delta method as in Macdonald et al. (2014). Comparing the two curves, it is immediately noticeable that when including historical data in the analysis the probability of high flows exceeding the magnitude recorded in December 2015 is much smaller than when systematic data only are used. In terms of return period, these two probabilities correspond to 126 years (annual exceedance probability equal to 0.0080) and 526 years (annual exceedance probability equal to 0.0019) respectively. The at-site estimate gives results comparable to the estimates obtained when using the regional analysis approach described in Environment Agency (2008), which gives an estimated return period for the December 2015 event of 132 years (annual exceedance probability equal to 0.0076).

To assess the difference that the different estimates of h would have on the overall estimation, Figure 7 show the 95% confidence intervals for the 100-year (annual exceedance probability equal to 0.01) and 1000-year event (annual exceedance probability equal to 0.0001) using different values of h, either estimated using the different estimators presented in Section 4 and some hypothetical high values of h: these are included to showcase the potential benefit of including records covering very long periods of time. It is immediately obvious that using the historical information in the estimation procedure gives much tighter confidence intervals, but little difference can be seen in the estimate and the variability obtained when using the three estimators for h. This is not so surprising given the fact the the actual estimated values of the h parameter are not so different for the three estimation methods, but this might not always be true for other case studies, especially if information is available only on few historical events. On the other hand, if the seven historical events would have been recorded in a much longer period of time than the one available for the Lune, the reduction in the estimation variability would be even more significant. This indicates how including information on long records can be extremely beneficial in terms of uncertainty reduction. Viglione et al. (2013)

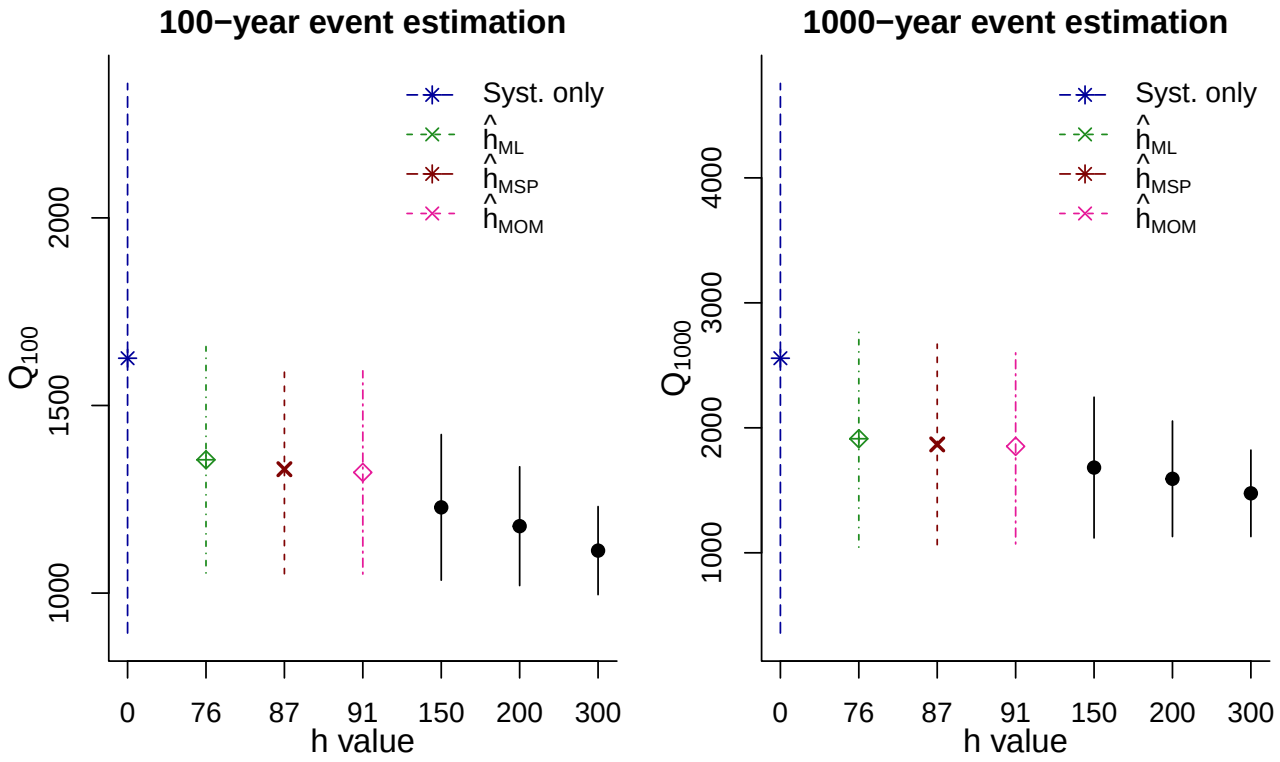

Figure 7: 95% confidence intervals for Q_{100} and Q_{1000} using systematic data only and the seven historical events higher than 878 m^3/s with varying parameter h values (i.e. the coverage of the historical period).

had already noted that very large increases in the historical coverage would be needed for the variability in the estimation to reduce significantly, and it is often the case that decisions about the values of the perception threshold X_0 and the number of historical events for which some information on the flow values can be determined also play a role in the assessment of the time coverage used in the analysis. To give a more complete assessment of the sensitivity of the estimation of flood risk for the Lune at Caton, the 95% confidence intervals for the 100-year and 1000-year event using increasing subsets of the historical records are shown in Figure 8. For each value of k, the number of historical events in the record, the perception threshold is set to be the peak flow of the largest historical event smaller than the largest k peaks. For example when $k = 2$, only the November 1923 and December 1954 events are included in the analysis, and the perception threshold is set to $X_0 = 1104$, the peak flow value of the January 1903 event, i.e. the third largest event in the historical record. The case of $k = 7$ corresponds to the curve shown in Figure 6. The historical record length is kept fixed at h=87. It is quite striking how little is the effect of including a smaller set of historical values and changing the

perception threshold for this case study. This might not be the case in all situations, and it is sometimes the case the very different flood frequency curves can be obtained depending on the decision of which subset of the available historical data is used in the estimation. Once again the most striking feature in Figure 8 is how including historical information results in much tighter confidence intervals around much smaller design event (i.e. the return period of large floods is estimated to be higher when historical information is included in the estimation).

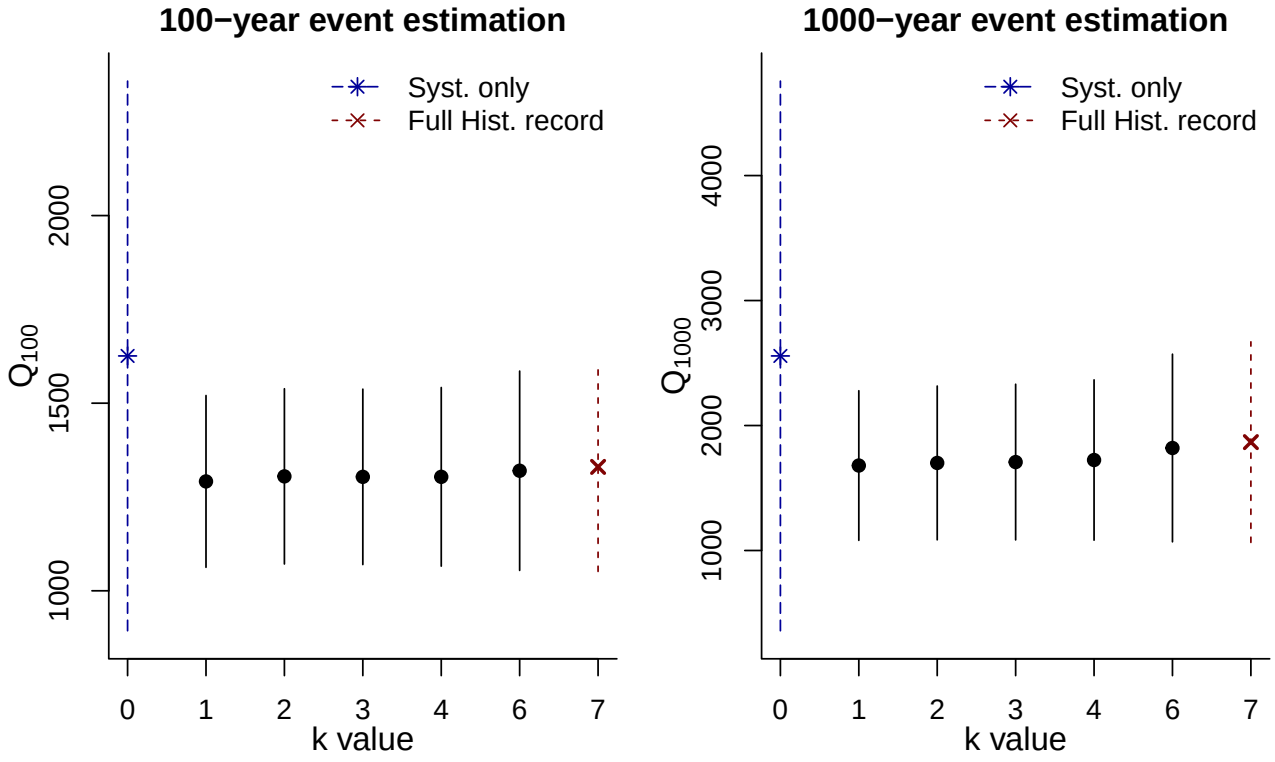

Figure 8: 95% confidence intervals for Q_{100} and Q_{1000} using systematic data only and varying number of historical events k (with corresponding varying perception threshold X_0).

The fact that tighter confidence bands are obtained when using historical data indicates that a higher confidence is attached to the estimated return curve. This is the natural consequence of using more information in the estimation procedure, although the estimated value is very different from what would have been obtained when applying the standard procedures used in the UK. This is mostly due to the fact that there appear to be more extreme events in the systematic record than in the historical period: there are a total of 11 events above the perception threshold in the 46 years of the systematic record, against the seven events in the 87 years of the historical record. Further, the 2015 event was indeed very large and well above any of the systematic and historical events. Similar findings in terms of how excessively high

the Winter 2015 events in Northern Britain were compared with a long historical record, are also discussed in Parkes and Demeritt (2016), which presents the history of flooding of the city of Carlisle from the river Eden. As a reference for how influential the most recent events might be for the estimation of flood risk in the area it is here noted that the return period for the value of 1700 m^3/s when including the peak flow value of December 2015 in the sample would be approximately 60 years (annual exceedance probability equal to 0.0165) when using systematic records only and 240 years (annual exceedance probability equal to 0.0042) when historical data are used in the estimation. These are spectacular increases of the frequency at which very high events might be expected to occur compared to the values one would have obtained before the December 2015 events. This questions the validity of the methods used to estimate the frequency of flooding, which ultimately rely on the working assumptions that past events can be informative of the present and future risk. An attribution of the drivers which caused a higher number of large flooding events in the more recent years is beyond the scope of this investigation, although climate change (van Oldenborgh et al., 2015) and large scale natural cycles (Macdonald, 2014) have been connected with increased likelihood of extreme events. It is finally worth to point out that the higher confidence in statistical terms which is obtained when using historical data can sometimes not coincide with a higher confidence in the final estimate, given that very different results are obtained when using the more recent and systematically gauged record. The decision of which frequency curve to use for design event estimation would lead to very different results: Bayliss and Reed (2001) discuss some practical ways in which decisions could be made on whether to rely on the results which include historical data or how to modify estimates obtained using the standard methods based on the regional frequency analysis approach. One of the possible and simplest ways to assess how including historical data impacts the overall flood risk estimation is to run some sensitivity analysis as those presented above or in Viglione et al. (2013). See also Environment Agency (2017) for a large simulation study which investigates the impact on the overall estimation under different scenarios of historical data availability. For this specific case study it appears that simply including even few historical records already gives large differences in the final estimates and in the confidence intervals around them, and this is partially due to the fact that the largest events in the record have been recorded in the last decades rather than in the historical period. Of course this is not always the case, and sensitivity studies can be useful to

assess how different historical peak flows samples correspond to different final estimates. There is often some trade-off between the length of the historical period for which information can be retrieved, and the level at which the perception threshold can be reliably fixed, as information on very large floods which happened in the far past might be available, but it would then be unclear what perception threshold could be used since it is assumed that all historical events in the sample are higher than X_0. A possible approach to this would be to use varying perception threshold, as done for example in Naulet et al. (2005), although this requires a very thorough study of the history of the catchment. Further, as mentioned earlier in the paper, considerations on the suitability of using information from a time in which the catchment was likely very different from its present form also need to be taken into account. These consideration are made even more complicated by all the possible sources of non-stationarity, as statistical models which rely on the assumption of an underlying stationary process might be not appropriate and thus additional structures would need to be added in the model to account for the impact of changes in the climate and in the catchment (as in Machado et al., 2015).

7 Discussion

The use of historical data can help in reducing the uncertainty around the estimation of the frequency of rare events as testified by the widespread recommendation that they should be used when available (Kjeldsen et al., 2014; Environment Agency, 2017). Caution should be taken in ensuring that the historical records included in the estimation procedure can indeed be deemed representative of the present day risk, as the standard procedure assumes that the data generating process for the whole sample (systematic and non-systematic data) is unchanged in time: this might be a restrictive assumption and the statistical models might need to be adjusted to account for possible non-stationarities. Further, when historical records are used in the estimation, it is generally assumed that the properties of the historical events are correctly characterised and that all information needed to compute the likelihood shown in Equation (1) is available. If the historical events are not properly characterised, there is a risk of actually increasing the uncertainty in the estimation procedure (Strupczewski et al., 2014). The importance of using accurate historical data can therefore not be stressed enough,

and all efforts should be made to collect as much information as possible regarding past large events. It might nevertheless be the case that the start date of the period of time covered by the historical events can not be accurately retrieved in the historical information. In such cases, rather than discarding the historical information, an estimate for the coverage of the historical record can be obtained and plugged in the estimation procedure. Interestingly, the question of estimating the length of time covered by the historical record corresponds to the problem of estimating the size of a population, a classic statistical problem which is often referred to as the German tanks problem. Different estimators of the total size of a population are available in the literature: their theoretical derivation and properties have been presented and their performance has been investigated by means of a simulation study. The simulation study confirmed that the preferred method to estimate h should be the MSP estimator, which is unbiased and has minimal variance. The MSP estimator gives the best results in terms of the estimated value of h itself and in terms of the estimation of the extreme value model, which is the ultimate goal of the estimation in this application. The performance of the estimation improves with increasing sample sizes, as it would be expected. The use of historical data reduces the uncertainty in the estimation of the extreme value modeling, even when the detail of the temporal coverage is estimated and not known *a-priori*.

Acknowledgements

The author wishes to thank Thomas Kjeldsen for his useful comments on a earlier draft of this manuscript. Neil Macdonald provided the Sussex Ouse data and gave some insight on the state of historical data on flooding in England: his help is gratefully acknowledged. The comments of Alberto Viglione and an anonymous associate editor greatly improved the manuscript.

References

Bayliss, A. C. and D. W. Reed (2001). The use of historical data in flood frequency estimation. Technical report, Centre For Ecology & Hydrology, Wallingford, UK.

Bernardara, P., M. Andreewsky, and M. Benoit (2011). Application of regional frequency anal-

ysis to the estimation of extreme storm surges. *Journal Geophysical Research 116*(C02008). doi:10.1029/2010JC006229.

Bulteau, T., D. Idier, J. Lambert, and M. Garcin (2015). How historical information can improve estimation and prediction of extreme coastal water levels: application to the Xynthia event at La Rochelle (France). *Natural Hazards and Earth System Sciences 15*(6), 1135–1147. doi:10.5194/nhess-15-1135-2015.

Cheng, R. and N. Amin (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society. Series B (Methodological)*, 394–403.

Cohn, T. A., W. L. Lane, and W. G. Baier (1997). An algorithm for computing moments-based flood quantile estimates when historical flood information is available. *Water Resources Research 33*(9), 2089–2096. doi:10.1029/97WR01640.

Coles, S. and R. Sparks (2006). Extreme value methods for modelling historical series of large volcanic magnitudes. In H. M. Mader, S. G. Coles, C. B. Connor, and L. J. Connor (Eds.), *The Oxford Handbook of Innovation, Statistics in Volcanology, vol. 1*, Chapter 5, pp. 47–56. London: Geological Society of London, Special Publication of IAVCEI.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Elleder, L. (2015). Historical changes in frequency of extreme floods in Prague. *Hydrology and Earth System Sciences 19*(10), 4307–4315. doi:10.5194/hess-19-4307-2015.

Environment Agency (2008). Improving the FEH statistical procedures for flood frequency estimation. R&D Report SC050050, Environment Agency, Bristol, UK.

Environment Agency (2017). Making better use of local data in flood frequency estimation. R&D Report SC130009/R, Environment Agency, Bristol, UK.

Gaume, E., L. Gaál, A. Viglione, J. Szolgay, S. Kohnová, and G. Blöschl (2010). Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. *Journal of Hydrology 394*, 101–117. doi:10.1016/j.jhydrol.2010.01.008.

Hall, J., B. Arheimer, G. T. Aronica, A. Bilibashi, M. Boháč, O. Bonacci, M. Borga, P. Burlando, A. Castellarin, G. B. Chirico, P. Claps, K. Fiala, L. Gaál, L. Gorbachova, A. Gül, J. Hannaford, A. Kiss, T. Kjeldsen, S. Kohnová, J. J. Koskela, N. Macdonald, M. Mavrova-Guirguinova, O. Ledvinka, L. Mediero, B. Merz, R. Merz, P. Molnar, A. Montanari, M. Osuch, J. Parajka, R. A. P. Perdigão, I. Radevski, B. Renard, M. Rogger, J. L. Salinas, E. Sauquet, M. Šraj, J. Szolgay, A. Viglione, E. Volpi, D. Wilson, K. Zaimi, and G. Blöschl (2015). A european flood database: facilitating comprehensive flood research beyond administrative boundaries. *Proceedings of the International Association of Hydrological Sciences 370*, 89–95. doi:10.5194/piahs-370-89-2015.

Hirsch, R. M. (1987). Analysis of extraordinary flood events probability plotting position formulas for flood records with historical information. *Journal of Hydrology 96*(1), 185 – 199. doi:10.1016/0022-1694(87)90152-1.

Hirsch, R. M. and J. R. Stedinger (1987). Plotting positions for historical floods and their precision. *Water Resources Research 23*(4), 715–727. doi:10.1029/WR023i004p00715.

Hosking, J. R. M. and J. R. Wallis (1986). The value of historical data in flood frequency analysis. *Water Resources Research 22*(11), 1606–1612. doi:10.1029/WR022i011p01606.

Hosking, J. R. M. and J. R. Wallis (1997). *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press.

Johnson, R. W. (1994). Estimating the size of a population. *Teaching Statistics 16*, 50–52. doi:10.1111/j.1467-9639.1994.tb00688.x.

Kjeldsen, T. R. (2007). The revitalised FSR/FEH rainfall-runoff method. NERC, Centre for Ecology & Hydrology, Wallingford, UK.

Kjeldsen, T. R., N. Macdonald, M. Lang, L. Mediero, T. Albuquerque, E. Bogdanowicz, R. Brázdil, A. Castellarin, V. David, A. Fleig, G. O. Gül, J. Kriauciuniene, S. Kohnová, B. Merz, O. Nicholson, L. A. Roald, J. L. Salinas, D. Sarauskiene, M. Šraj, W. G. Strupczewski, J. Szolgay, A. Toumazis, W. Vanneuville, N. Veijalainen, and D. Wilson (2014). Documentary evidence of past floods in Europe and their utility in flood frequency estimation. *Journal of Hydrology 517*, 963–973. doi:10.1016/j.jhydrol.2014.06.038.

Kjeldsen, T. R. and I. Prosdocimi (2015). A bivariate extension of the Hosking and Wallis goodness-of-fit measure for regional distributions. *Water Resources Research 51*(2), 896–907. doi:10.1002/2014WR015912.

Kjeldsen, T. R. and I. Prosdocimi (2016). Assessing the element of surprise of record-breaking flood events. *Journal of Flood Risk Management Published online*. doi:10.1111/jfr3.12260.

Lam, D., C. Thompson, and J. Croke (2016). Improving at-site flood frequency analysis with additional spatial information: a probabilistic regional envelope curve approach. *Stochastic Environmental Research and Risk Assessment*, 1–21. doi:10.1007/s00477-016-1303-x.

Macdonald, N. (2014). Millennial scale variability in high magnitude flooding across britain. *Hydrology and Earth System Sciences Discussions 11*, 10157–10178. doi:10.5194/hessd-11-10157-2014.

Macdonald, N., T. R. Kjeldsen, I. Prosdocimi, and H. Sangster (2014). Reassessing flood frequency for the Sussex Ouse, Lewes: the inclusion of historical flood information since AD 1650. *Natural Hazards and Earth System Science 14*(10), 2817–2828. doi:10.5194/nhess-14-2817-2014.

Machado, M. J., B. A. Botero, J. López, F. Francés, A. Díez-Herrero, and G. Benito (2015). Flood frequency analysis of historical flood data under stationary and non-stationary modelling. *Hydrology and Earth System Sciences 19*(6), 2561–2576. doi:10.5194/hess-19-2561-2015.

Merz, R. and G. Blöschl (2008a). Flood frequency hydrology: 1. Temporal, spatial, and causal expansion of information. *Water Resources Research 44*(8). doi:10.1029/2007WR006744.

Merz, R. and G. Blöschl (2008b). Flood frequency hydrology: 2. Combining data evidence. *Water Resources Research 44*(8). doi:10.1029/2007WR006745.

Met Office (2016). Flooding in Cumbria December 2015. http://www.metoffice.gov.uk/climate/uk/interesting/december2015. Accessed: 2016-09-09.

Mudersbach, C., J. Bender, and F. Netzel (2015). An analysis of changes in flood quantiles at the gauge neu darchau (elbe river) from 1875 to 2013. *Stochastic Environmental Research and Risk Assessment*, 1–13. doi:10.1007/s00477-015-1173-7.

Natural Environment Research Council (1975). Flood Studies Report, NERC, London, UK.

Naulet, R., M. Lang, T. B. Ouarda, D. Coeur, B. Bobée, A. Recking, and D. Moussay (2005). Flood frequency analysis on the Ardéche river using French documentary sources from the last two centuries. *Journal of Hydrology 313*, 58–78. doi:10.1016/j.jhydrol.2005.02.011.

Neppel, L., B. Renard, M. Lang, P.-A. Ayral, D. Coeur, E. Gaume, N. Jacob, O. Payrastre, K. Pobanz, and F. Vinet (2010). Flood frequency analysis using historical data: accounting for random and systematic errors. *Hydrological Sciences Journal 55*(2), 192–208. doi:10.1080/02626660903546092.

Parent, E. and J. Bernier (2003). Bayesian POT modeling for historical data. *Journal of Hydrology 274*(1-4), 95–108. doi:10.1016/S0022-1694(02)00396-7.

Parkes, B. and D. Demeritt (2016). Defining the hundred year flood: A bayesian approach for using historic data to reduce uncertainty in flood frequency estimates. *Journal of Hydrology 540*, 1189 – 1208. doi:10.1016/j.jhydrol.2016.07.025.

Parry, S., L. Barker, I. Prosdocimi, M. Lewis, J. Hannaford, and S. Clemas (2016). Hydrological summary for the united kingdom: December 2015.

Ranneby, B. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics*, 93–112.

Reis, D. S. J. and J. R. Stedinger (2005). Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology 313*, 97–116. doi:10.1016/j.jhydrol.2005.02.028.

Rice, J. (2006). *Mathematical statistics and data analysis*. Nelson Education.

Ruggles, R. and H. Brodie (1947). An Empirical Approach to Economic Intelligence in World War II. *Journal of the American Statistical Association 42*, 72–91. doi:10.1080/01621459.1947.10501915.

Salinas, J. L., A. Castellarin, A. Viglione, S. Kohnová, and T. R. Kjeldsen (2014). Regional parent flood frequency distributions in Europe - Part 1: Is the GEV model suitable as a pan-European parent? *Hydrology and Earth System Sciences 18*(11), 4381–4389. doi:10.5194/hess-18-4381-2014.

Silva, A. T., M. Naghettini, and M. M. Portela (2016). On some aspects of peaks-over-threshold modeling of floods under nonstationarity using climate covariates. *Stochastic Environmental Research and Risk Assessment 30*(1), 207–224. doi:10.1007/s00477-015-1072-y.

Silva, A. T., M. M. Portela, M. Naghettini, and W. Fernandes (2015). A bayesian peaks-over-threshold analysis of floods in the itajaí-açu river under stationarity and nonstationarity. *Stochastic Environmental Research and Risk Assessment*, 1–20. doi:10.1007/s00477-015-1184-4.

Stedinger, J. R. and T. A. Cohn (1986). Flood Frequency Analysis With Historical and Paleoflood Information. *Water Resources Research 22*(5), 785–793. doi:10.1029/WR022i005p00785.

Strupczewski, W. G., K. Kochanek, and E. Bogdanowicz (2014). Flood frequency analysis supported by the largest historical flood. *Natural Hazards and Earth System Science 14*(6), 1543–1551. doi:10.5194/nhess-14-1543-2014.

U.S. Interagency Advisory Committee on Water Data (1982). Guidelines for determining flood flow frequency, Bulletin 17-B of the Hydrology Subcommittee. Technical report, Reston, Virginia.

van Oldenborgh, G. J., F. E. L. Otto, K. Haustein, and H. Cullen (2015). Climate change increases the probability of heavy rains like those of storm Desmond in the UK - an event attribution study in near-real time. *Hydrology and Earth System Sciences Discussions 12*, 13197–13216. doi:10.5194/hessd-12-13197-2015.

Viglione, A., R. Merz, J. L. Salinas, and G. Blöschl (2013). Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research 49*(2), 675–692. doi:10.1029/2011WR010782.