

Multivariate extreme value theory in flood mapping

Keef, Caroline
 JBA Consulting
 South Barn, Broughton Hall
 Skipton, UK
 E-mail: caroline.keef@jbaconsulting.co.uk

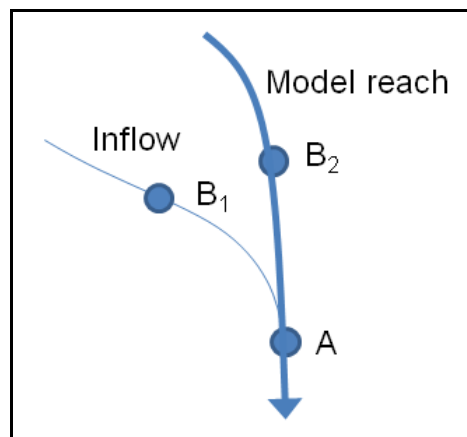
Introduction

Flooding is a natural phenomenon that affects many countries worldwide. In the past year alone major river floods have caused devastation in Pakistan, Australia and the United States. Floods will never be prevented, instead ways of living with flooding must be found. A vital part of living with and managing flooding is estimating the areas at risk of flooding. A common approach to this estimation is to use computational hydraulic models to map the extent of large floods. Flood mapping has two stages, the first is to estimate the volume of water that will flow during floods, the second is to drive this water through a hydraulic model to obtain the extent. A flood flow is typically defined as the flow that will be exceeded once every X years, where X is taken to be a range of values in the region 20 – 1000 years.

Hydraulic modeling engines vary in complexity and hence the time they take to spread the water. Simple models use approximate techniques based wholly on volume and floodplain topography and can spread water across the floodplain in seconds. More complex models solve the full hydrodynamic two dimensional shallow water equations (Stelling, 1984) and generally take hours to run. In the UK, and many other countries, the use of these complex hydrodynamic models is standard practice in high risk areas.

Many flood mapping projects simply map the flood flows from a single stretch of river, and so it is only necessary to calculate the flood flows at a single point. However in mapping flood flows around a confluence likely flows on both tributaries must be estimated. For example, in the confluence modeled in Figure 1 to estimate the extent of flooding that will be exceeded once every 100 years downstream of the confluence, sensible simultaneous flows at B1 and B2 must be entered so that the flow at A is equal to the 100 year flow.

Figure 1: Example of a confluence



A common engineering solution is to simply enter flows with the same exceedance probability, however these are unlikely to occur in reality, and so flows (and floods) downstream of the tributary

are likely to be overestimated. An improved approach to estimating these simultaneous flows is to use simultaneously observed data and a statistical model for simultaneous extreme values.

An alternative would be to use a continuous simulation approach, where long rainfall time series are generated from a stochastic rainfall generator and this simulated rainfall is used as an input to a hydrological rainfall runoff model to estimate flows for the catchment, an examples is Calver et al. (2005). Although this method has been used successfully in the past the results are subject to a large amount of uncertainty; and some statistical analysis must be undertaken on the resulting simulated flows in order to answer the questions posed in this study.

Because the river flows requiring modelling are often larger than those that have been observed it is necessary to use statistical dependence models that have been developed for extreme values. Many extreme value dependence models have been developed for use with block (e.g. annual) maxima data (for an overview see Bierlant et. al., 2004). However these approaches cannot predict likely concurrent flows at tributaries, as annual maxima are not guaranteed to be observed concurrently at different locations.

One dependence measure that is used in flood mapping is χ which is equal to the limiting probability that one variable is above a high probability threshold conditional on another variable being above the same probability threshold (for full details see Coles et al 1999). This measure has been applied by Svensson and Jones (2002, 2004), who mapped χ for pairs of flood risk variables (sea surge, river flow and precipitation) around the coast of Great Britain. For the purposes of the present study this dependence measure has two disadvantages: the first is that it is only possible to use it for pairs of variables so it would not be possible to assess likely flows at B_1 and B_2 when A has an extreme flow. The second disadvantage is that it is a measure of dependence rather than a description of the full joint distribution function.

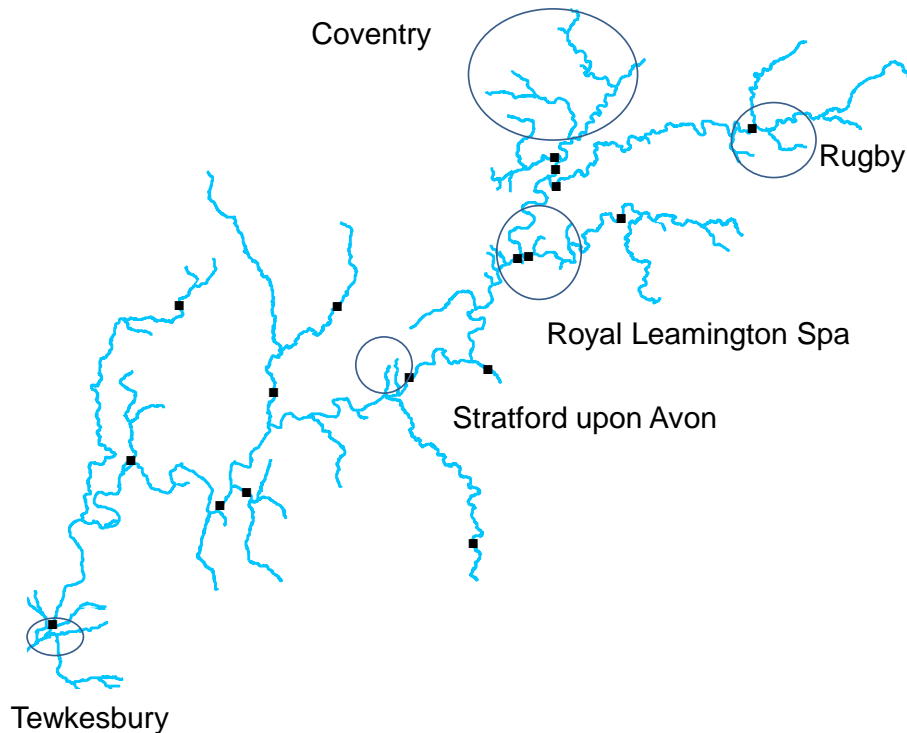
A possible choice for a statistical model is that of Heffernan and Tawn (2004). This describes the dependence between a set of variables conditionally on one of these variables having an extreme observation, where extreme is defined as being above a certain (high) threshold. This is a semi-parametric model in which the conditional joint distribution of a set of variables \mathbf{Y} is modelled with relation to another conditioning variable X . The conditional distributions of $Y_j|X$, $Y_j \in \mathbf{Y}$, are modeled parametrically and the conditional joint distribution of $\mathbf{Y}|X$ is modeled non-parametrically. Due to the semi-parametric nature of the model the simplest way of using the model to set inflows to hydraulic model is to simply simulate a Monte-Carlo sample from a fitted distribution. This approach has been used in Neal et al (2011) where a sample of Monte-Carlo simulated events were used to assess the likely extent of flooding around Carlisle, a town in north west England. The focus of that study was to demonstrate a possible method to estimate the uncertainty in the estimation of this extent. However, the hydraulic model used in this study is considerably simpler than those used in detailed modeling, and so cannot produce the detailed maps used in flood management of high risk areas.

The approach that we describe here is suitable for practical use with detailed hydraulic models, although it is not guaranteed to deliver a physically meaningful answer. Instead the results must be viewed as a starting point for the modelling validation and calibration process. The way we use the Heffernan and Tawn model is to estimate the likely flows upstream of a tributary that will combine to produce the 100 year flow downstream. We demonstrate this approach using the real example of a hydraulic model that was built for the River Avon, the location of which is shown in Figure 2.

The flows entered to the hydraulic model are estimated by fitting a distribution to the annual maxima observations. This is in contrast to the marginal distributions fitted in the dependence

modelling. This causes some difficulty in the modelling and we discuss reasons for this and possible solutions in this paper.

Figure 2: Location of study area, square points show gauging stations used in the analysis, blue circles rough locations of towns, blue line river network



Data and study area

In total flow or stage (water depth) data is available from 43 gauging stations, however only 17 are suitable for use in this analysis. The data we use in the analysis are daily maximum observations from these stations. The gauge records we have are a mixture of flow and stage records which, because we standardize the marginal distributions, does not cause any difficulty in the modeling procedure. Stations were rejected due to short record lengths, presence of artificial influences such as abstractions and censored data records (i.e. all high observations given the same value). We focus on the most downstream confluence of the model and illustrate how some difficulties with data can be solved using two other sites.

Statistical model

Within the Heffernan and Tawn model the joint distribution of a set of variables (in our case daily maxima gauging station records) is modelled in two stages. First a distribution is fitted to each individual record; these records are then transformed so that they have the same distribution function. A suitable distribution function to use for this standard distribution is the Gumbel distribution.

The second step of the modelling procedure is to model the dependence structure between the variables. This dependence structure is modelled by assuming that it is possible to express the conditional distribution of a set of variables, \mathbf{Y} , given another variable, X , takes a certain value in the form of $\mathbf{Y} = \alpha X + X^\beta \mathbf{Z}$. The parameter sets α and β are constants that describe the dependence

between Y and X and the variables Z are residuals which play a part in describing the variation of the conditional distribution of Y given a certain value of X .

The dependence model describes the whole distribution of Y for a certain value of X . So in addition to simply stating the most likely value of Y for a given value of X we can also give the range of likely values of Y for a given value of X . For each confluence, we state the median (most likely) value of Y when X has a 100-year flow and the 25th and 75th percentiles of the distribution of Y when X has a 100-year flow. We also state the level of correlation between the variables in Y given that X has a 100 year flow. This can inform the modeller if there are any combinations of Y that should be kept together, or if the inflows from the different upstream catchments can be varied independently.

Timing between peaks

In addition to estimating the likely flow/stage at upstream points given a 100 year return period downstream of a confluence we also examine the time differences between the peaks of the events at different locations. The first stage of this analysis is to identify events at each individual gauging station. We define events as periods of time when the observed flow or stage is above the 0.99 probability threshold, so periods of time when the observations are in the top 1% of all observations. In order to separate events we say that if two exceedances of the threshold are separated by five days then they belong to separate events.

We do not attempt to model the time differences between flood peaks at different locations. Instead we report the median time difference (in hours) and also the 25th and 75th percentiles. We only report these differences when both gauging stations observe an event within +/- 1 day, so when the downstream station observes a peak on either the same day, the day before, or the day after a peak on the upstream station.

Distribution of individual gauge records

The first step in the modelling procedure is usually to fit a distribution to the highest observations from each gauge record in turn. This is done by extracting all the observations above a threshold and fitting a suitable distribution. We fitted a Generalised Pareto Distribution (GPD) as this choice of distribution is theoretically shown to be suitable. However when we performed diagnostic checks on the fitted distribution we found that for two records the fitted distribution did not provide a good fit to the data. These were the Leam at Leamington and Badsey Brook at Offenham. From past experience poor fitted distributions often signal irregularities with the original data, although poor fitting choices also affect the fit. For both sites, the two highest observations appear to be much higher than the rest of the observations, these were recorded during the Easter 1998 and July 2007 flood events. Because of the difficulties in measuring high flows there are a number of reasons why these observations may be inaccurate. So it is our opinion that, certainly for the 1998 observations, and the 2007 observation at Badsey Brook, the recorded flows for these dates at these two sites are higher than the true flows. However it is still likely that the flows on these dates were high and that they were the top two flows on record at these two locations which means that they are valuable observations and should not be disregarded.

One solution in this situation is to use the empirical distribution function at each location to transform the data to standard scales. This is equivalent to using only the ranks of the observed flows/stages to inform about the probability of observing that particular flow/stage. Using the empirical distribution function eliminates most concerns about data irregularities, but introduces additional uncertainty in the resulting dependence analysis and makes it difficult to express the likely values at upstream gauges of confluences in any way other than the probability scale.

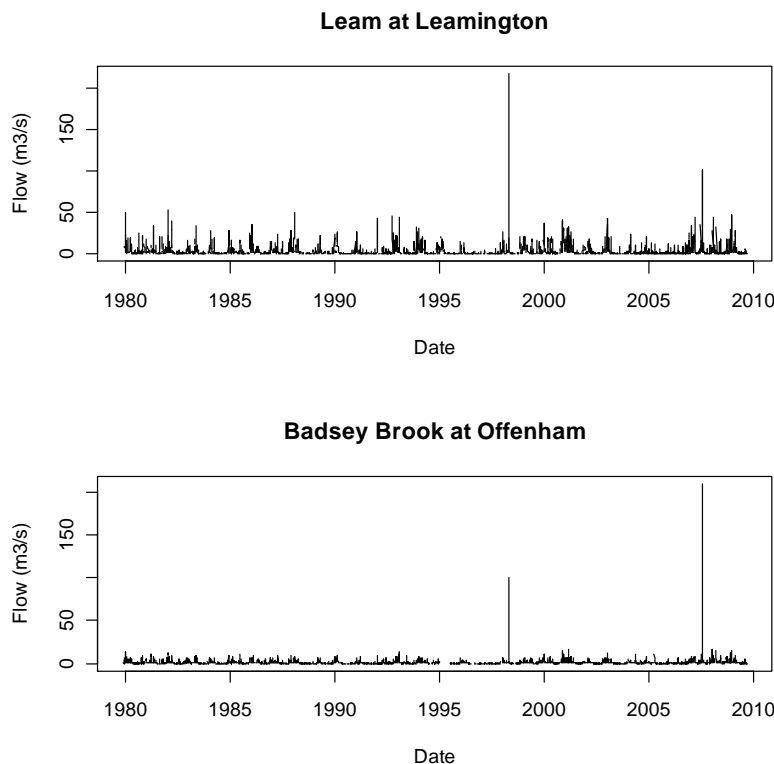
An alternative solution is to use the fitted GPD distributions, but replace the erroneous values in the two flow records that appear to have a poor fit. This is the approach we took. We replaced the 1998 event peak at the Leam at Leamington with 75 m³/s, which is consistent with the record of a nearby upstream gauge, the Leam at Eathorpe, we did not replace the 2007 peak record as this observation is consistent with that observed at the Leam at Eathorpe, and it is also not as clear that it does not fit the rest of the flow record. There is no gauging station upstream of Offenham on Badsey Brook, so we replaced the observation from the 1998 event with 50 m³/s, and the observations from the 2007 events with 95 m³/s, these flows are consistent with the fitted distribution.

The final stage in obtaining the distribution function at the individual gauging stations is to calculate the relationship between how likely it is to exceed a particular value on a single day, and how likely it is to exceed a particular value in a year. The approach that we took here is to exploit a theoretical relationship between the Generalised Extreme Value distribution and the GPD distribution. If we assume that the number of events per year follows a Poisson distribution then the return period T of a flow with the probability p of being exceeded on any given day can be estimated by

$$T = \frac{k}{365.25(1-p)}$$

where k is the average number of days each event lasts for. The return periods estimates obtained using this approach are approximations and are typically better for higher return periods, but are essentially meaningless below 1 year. Because of this fact we also state the flows/stages used for the 100-year return period at the downstream station of the confluence, and also the flows/stages at the upstream gauging stations given the 100-year return period at the downstream point.

Figure 3: Time series plots for sites with poor fitting GP distributions



Results

The confluence we consider here is the most downstream gauge on the Avon within the area of interest. This is the Avon at Tewkesbury and is a level only station and has recorded 50 events. The two gauging stations upstream of this that we consider are the Avon at Evesham and Piddle Brook at Wyre Piddle. Table 1 summarizes the time differences between the Avon at Tewkesbury and peaks at these two stations and Table 2 summarizes the likely flows at the upstream stations given a 100 year stage at Tewkesbury, the stage taken to be the 100-year at Tewkesbury is 6.41m. One consideration that should be taken into account when setting flows at these two upstream gauges is that there is a high level of residual dependence between them when there is a 100 year flow at Tewkesbury, which is illustrated in

Figure 4. This residual dependence means that if a high flow is set at Evesham then a high flow should also be set on the Piddle Brook.

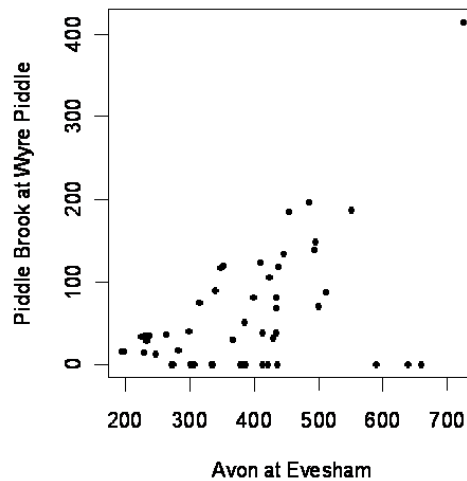
Table 1: Time differences between peak flows positive time differences show flood peaks occurring at the Avon at Tewkesbury after flood peaks at the upstream stations

Gauging station	25th percentile of time difference (hours)	Median time difference (hours)	75th percentile of time difference (hours)	Mean time difference (hours)	Number of simultaneous events
Avon at Evesham	14	19	23	16	9
Piddle Brook at Wyre Piddle	17	25	30	23	4

Table 2: Summary of flows at upstream stations given 100 year flow at the Avon at Tewkesbury

Gauging station	25th percentile of return period	Median return period	75th percentile of return period	25th percentile of flow (m ³ /s)	Median flow (m ³ /s)	75th percentile of flow (m ³ /s)
Avon at Evesham	21	72	132	298	386	437
Piddle Brook at Wyre Piddle	3	13	34	14	35	89

Figure 4: Simulated data for Avon at Evesham and Piddle Brook at Wyre Piddle, conditional on the Avon at Tewkesbury having a 100-year flow in m³/s



Daily to annual

In using the results from this study the estimated flows from the daily maximum analysis were not used. Instead the industry standard approach was used, the distribution function of the annual maxima was estimated based on annual maxima data and the method described in the Flood Estimation Handbook (FEH) (Robson and Reed, 1999) for combining flows from different locations. The FEH method is based on the index flood method (Dalrymple, 1960) and is based on annual maxima data from a set of sites and has two stages. The first is to estimate an 'index flood', (the median maxima in the FEH), for each site, and then estimate a 'growth curve' for the whole set of sites. The growth curve describes how flows change with increasing non-exceedance probability. The distribution function of flood flows for a single site can be obtained by multiplying the growth curve for the set of sites by the index flood of that particular site.

The simplest way to match the flow distribution fitted to daily maxima data to the flow distribution fitted to annual maxima data would be to fit these separately and then match them up by working out the exceedance probability in days of the flow with a 1 in X year exceedance probability. However, typically these two distributions do not match up very well. An alternative approach would be to combine a model for the peak flows in a flood event on the daily scale with a separate distribution for the number of events per year to produce a distribution for the number of events per year. For this we would to determine suitable forms for both of these distributions.

The Extremal Types Theorem gives the Generalised Extreme Value (GEV) Distribution as the limiting distribution of the maxima of samples where each observation, within each sample and for different samples, is independent and identically distributed. If the number of flow peaks a year follow a Poisson distribution then the exceedances of a (high) threshold then follow a Generalised Pareto (GP) distribution. For river flow data from a single site hourly, daily, or even monthly observations are neither independent nor identically distributed. River flow observations are time series which exhibit both temporal dependence and seasonality. This suggests that the assumption of the GP or GEV distribution may not be valid, particularly for time series which exhibit very high levels of temporal dependence or seasonality. An alternative distribution commonly used by hydrologists is the Generalised Logistic (GL) distribution which has been found (Robson and Reed,

1999) to provide a good fit to annual maxima of flow records in the UK. The GL distribution has a formulation similar to the GEV but there is little theoretical justification for its use.

The fact that seasonality means that flow data is not identically distributed is often ignored in fitting a distribution. A possible method to account for this non-stationarity would be to derive a model for the underlying seasonality in the series, for instance by modelling the expected flows on each day of the year, and then to carry out an extreme value analysis on the deviations from this model. However, it can be debated whether or not this type of approach would work for river flow data. The first problem is that of characterisation of the seasonal process, which is unlikely to be a simple sinusoidal variation. The second problem is that it is not unlikely that summer deviations would have different characteristics to winter deviations; the flow processes from rainfall on dry ground with a high level of vegetation are different to those when rain falls on wet ground when many plants are dormant. This means that fitting a suitable distribution to these deviations is likely to be no less complicated than fitting a distribution to the raw flow observations and also that peak observations in summer (which are unlikely to result in flooding) may provide little information about the distribution of peak observations in winter. An approach for this type of modelling was developed by Eastoe and Tawn (2009) and demonstrated using extremes of air pollution data.

An alternative to the strong assumption of seasonality is to assume that once the ground conditions are such that flooding is possible each flood event is independent and has an identical distribution. This assumption is implicit in using peaks over threshold methods. If in addition the number of days in which ground conditions were such that flooding is possible were the same in each year, then the block size for each year would be identical and so the GEV distribution would still be observed as the correct limit distribution for annual maxima. In reality there is between year variation in ground conditions and so the number of days for which flooding is possible varies greatly from year to year.

There is much empirical evidence (Robson and Reed, 1999, Lang, 1999) that a Negative Binomial distribution provides a better fit to the number of flood events per year than the Poisson distribution. Compared to the Poisson distribution the Negative Binomial is over dispersed i.e. the variance is greater than the mean. The Negative Binomial is an attractive distribution to use in modelling the number of events per year for a number of reasons. First a special case of the Negative Binomial when combined with the GP distribution for the size of flood events leads to the GL distribution for AMAX data. The second main reason is that it can be formulated as a combination of the Gamma and Poisson distributions, in that the number of events in any particular year follows a Poisson distribution with parameter λ where λ follows a Gamma distribution. An important point to note here is that although the Gamma distribution is sufficiently flexible that it is likely to provide a suitable fit to most data sets, there is no theoretical justification for its use.

To overcome some of the limitations of classical extreme value theory we present two alternative models developed by Eastoe and Tawn (2010, 2011). In Eastoe and Tawn (2010) the theorem of Leadbetter (1983) is used which allows the assumption that event maxima follow a GP distribution. It is then assumed that the number of events per year follow a specific formulation of the Negative Binomial distribution. This construction leads to the Extended Generalised Logistic distribution, which collapses to the GL distribution. Eastoe and Tawn successfully fitted this distribution to annual maxima from the Thames at Kingston, which has a record length of 123 years, using Bayesian MCMC techniques, and introducing covariates for the between year differences.

There are two difficulties with this approach; the first is that on empirical evidence of flow data from UK gauging stations, the Negative Binomial distribution is, although an improvement on the Poisson, is still not an adequate fit in many locations. The second is that it can be observed that a GP

distribution fitted to cluster maxima is often different to a GP distribution fitted to all exceedances of a threshold. This is consistent with the findings of Fawcett and Walshaw (2007) who found evidence of bias in estimates of the GP distribution when these estimates were obtained using cluster maxima only. To overcome the first problem, when data from a number of sites, or a very long record, is available it may be possible to use the observed empirical distribution of numbers of events per year as an approximation of the true number of events per year. The observed number of events per year can be identified using standard declustering methods, and the two threshold method of Laurini and Tawn closely matches the declustering procedure described in the Flood Studies Report (NERC, 1975) and FEH (Robson and Reed, 2000).

To overcome the limitation in terms of the distribution of cluster maxima Eastoe and Tawn (2011) provide a model for the cluster maxima of exceedances of sub-asymptotic thresholds. The solution provided by this model is to introduce a modifier to the GP distribution that describes how the degree of clustering varies with the level of extremeness. The implication of this model is that the GP is only an appropriate distribution for cluster maxima if the length of time a river can be expected to be above a threshold u in a single cluster is the same as the length of time a river can be expected to be above a threshold x at one time. In the range of the data these cluster lengths can be estimated empirically, however for the range of data for which there are no observations the temporal dependence must be explicitly modelled. Eastoe and Tawn examine two dependence models, those of Ledford and Tawn (1996) and Heffernan and Tawn (2004), and found that both resulted in models that fitted cluster maxima from the river Lune at Caton but that the Heffernan and Tawn model has the advantage of a much lighter computational burden.

An approach to estimating the distribution of annual maximum flows from daily data is described in Keef (2012). In this approach the empirical distribution for numbers of events per year is combined with the distribution of cluster maximum suggested by Eastoe and Tawn (2011). This approach was found to give very similar estimates to a GEV distribution fitted to the annual maxima, but with narrower confidence intervals. This result is unsurprising as it has been shown by Madsen et al. (1997) that a peaks over threshold approach performed better than a annual maxima approach at estimating the distribution of flood flows. Gauging stations where this approach did not fit well were those that had a very high level of temporal dependence and large seasonal variation. However it is for these sites that the underlying assumptions for choosing both the GEV and GP distributions are least safe and it is likely that this is the reason for these poor fits. It is likely that a more sophisticated modelling procedure is required here.

Discussion

The approach to estimating simultaneous extreme flows described in this paper is one that has been used successfully in a number of applications. However that does not mean it cannot be improved. The main area of improvement is in the modelling of flood flows at each location. The current approach is unsatisfactory in that the differences in fitted distributions used for the dependence modelling, and for the hydraulic modelling mean introduce uncertainties in the resulting output. The best approach would be to use the distribution obtained from peaks over threshold modelling to estimate the flood flows used in the hydraulic modelling. However there is considerable resistance to this approach. The index flood method based on AMAX data is regarded as industry standard and its use is defended using a number of arguments such as simplicity and eliminating the need for choosing a threshold and identifying clusters. However advances in statistical extreme value theory have solved these issues.

Political resistance to a new method also exists. Typically the people who have responsibility for flood risk management are typically not statisticians, and so very strong arguments must be made

to convince them that a new method is better. Changing to a new method would also have implications in terms of the confidence placed on previous studies. If a new method is approved for all future studies, how reliable are past studies and what are the cost implications for repeating these studies?

There are two practical arguments that do currently weigh in favour of AMAX approaches. The first is that it is not always possible to obtain daily maxima (or alternatively hourly or 15 minute observations) instead only daily mean data is available, this does not fully describe the flow peaks. The second is that the use of the statistical software that can be used to implement the more advanced approaches is not widespread amongst flood mappers and hydrologists. Tools to make these more advanced methods available would help widespread uptake.

A major disadvantage of the dependence estimation approach described here is that it can only be used in locations where flow records are available. A more general approach was used in Faulkner et al (2010) who used the same dependence model to estimate the joint distribution of pairs of flow variables for a large number of pairs of stations in Ireland. The differences between the catchments of these gauging stations were then compared and provided 'rules of thumb' for setting likely flows upstream of a confluence given a 100-year flow downstream.

References

- Bierlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004) *Statistics of Extremes Theory and Applications*, John Wiley & Sons
- Calver, A., Crooks, S., Jones, D. Kay, A., Kjeldsen, T. and Reynard, N. (2005) National river catchment flood frequency method using continuous simulation modelling. R&D Technical Report FD2106/TR, Department for Environment, Food and Rural Affaris, London, UK.
- Coles, S. G., Heffernan, J. E. and Tawn, J. A. (1999). Dependence measures for extreme value analyses. *Extremes*, 2, 339–365.
- Dalyrymple, T. (1960) Flood frequency analysis. USGS Water-Supply Paper 1543A
- Davison; A.C. and Smith, R.L. (1990). Models for exceedances over high thresholds (with discussion). *J.R. Statist. Soc.* **52**, 393-442
- Eastoe, E.F. & Tawn, J.A. (2010). Statistical models for overdispersion in the frequency of peaks over threshold data for a flow series. *Water Resour. Res.*, 46.
- Eastoe, E.F. and Tawn, J.A. (2011) The distribution for the cluster maxima of exceedances of sub-asymptotic thresholds (preprint)
- Faulkner, D., Keef, C., and Martin, J. 2010. How to calculate design inflows for catchment-wide flood studies. In: *Proc. BHS 2010 International Symp.*, Newcastle, July 2010.
- Laurini, F. And Tawn, J. A. (2003) New estimators for the extremal index and other cluster characteristics. *Extremes*, **6**, 189-211.
- Madsen, H., Rasmussen, P.F. and Rosbjerg, D. (1997) Comparison of annual maximum series and partial duration series methods for modelling extreme hydrologic events 1. At-site modelling. *Water. Resour. Res.* **33(4)** 747-757

Northrop, P. and Jonathan, P., 2010, Modelling spatially dependent non-stationary extremes with application to hurricane induced wave heights. Research Report 306, Department of statistical science University College London

Robson, A. and Reed, D. (1999) Flood Estimation handbook Vol.3: Statistical procedures for flood frequency estimation. Wallingford: Institute of Hydrology

Smith, R.L. and Weissman, I. (1994), Estimating the extremal index. *J.R. Statist. Soc. B* **56**, 515-128

Stelling, G.S. (1984) On the Construction of Computational Methods for Shallow Water Flow Problems Rijkswaterstaat Communications, No. 35/1984, The Hague, The Netherlands.

Svensson, C. and Jones, D. A. (2002) Dependence between extreme sea surge, river flow and precipitation in eastern Britain, *Int. J. Climatol.*, **22**, 1149-1168.

Svensson, C. and Jones, D. A. (2004) Dependence between extreme sea surge, river flow and precipitation in south and west Britain, *Hydrol and Earth System Science*, **8**, 973-992.