E3S Web of Conferences 7, 01011 (2016) DOI: 10.1051/e3sconf/20160701011
FLOOD*risk* 2016 - 3rd European Conference on Flood Risk Management

Improving statistical models for flood risk assessment

Ross Towe[1,2,a], Jonathan Tawn[2], Rob Lamb[1,4], Chris Sherlock[2] and Ye Liu[3]

[1]*JBA Trust, Broughton Hall, Skipton, BD23 3AE, United Kingdom*
[2]*Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, United Kingdom*
[3]*JBA Risk Management, Broughton Hall, Skipton, BD23 3AE, United Kingdom*
[4]*Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, United Kingdom*

Abstract. Widespread flooding, such as the events in the winter of 2013/2014 in the UK and early summer 2013 in Central Europe, demonstrate clearly how important it is to understand the characteristics of floods in which multiple locations experience extreme river flows. Recent developments in multivariate statistical modelling help to place such events in a probabilistic framework. It is now possible to perform joint probability analysis of events defined in terms of physical variables at hundreds of locations simultaneously, over multiple variables (including river flows, rainfall and sea levels), combined with analysis of temporal dependence to capture the evolution of events over a large domain. Critical constraints on such data-driven methods are the problems of missing data, especially where records over a network are not all concurrent, the joint analysis of several different physical variables, and the choice of suitable time scales when combining information from those variables. This paper presents new developments of a high-dimensional conditional probability model for extreme river flow events conditioned on flow and rainfall observations. These are: a new computationally efficient parametric approach to account for missing data in the joint analysis of extremes over a large hydrometric network; a robust approach for the spatial interpolation of extreme events throughout a large river network,; generation of realistic estimates of extremes at ungauged locations; and, exploiting rainfall information rationally within the statistical model to help improve efficiency. These methodological advances will be illustrated with data from the UK river network and recent events to show how they contribute to a flexible and effective framework for flood risk assessment, with applications in the insurance sector and for national-scale emergency planning.

1 Introduction

Understanding the risks of widespread flooding is a vital part of flood risk management. Given that we are then interested in understanding the chance of multiple locations experiencing flooding concurrently, we can use a joint probability analysis to understand the statistical characteristics of previous events. This joint probability analysis then can provide us with predictions of plausible future flood events, subject either to stationarity assumptions or explicit modelling of any underlying changes in the environment.

Statistical extreme value models provide a natural way of parametrising complex relationships between environmental variables at a number of different locations.

There are a number of possible statistical models that can be used to characterise the spatial and temporal distribution of flood events. Exploring one potential approach, some recent research has explored the development of models based on max-stable processes for river network data [1]. However max-stable processes are only suitable for spatial extreme data sets that exhibit asymptotic dependence [1], a property where the largest values in each variable can occur concurrently. Therefore they are practically infeasible for higher-dimension problems such as models of national scale river flow or rain gauge networks, because simultaneous extreme flows are unlikely in any one event over all sites.

Instead, the statistical model that we adopt is the conditional exceedance model of Heffernan and Tawn [2]. The conditional exceedance model is adopted as it is able to handle the range of extremal dependence structures, including both asymptotic independence and asymptotic dependence, which are observed within the extremes of a river flow data set [3]. Previous extensions to the conditional exceedance model were developed by Keef et al. [4], to account for missing data and temporal dependence and Keef et al. [5], to obtain an improvement in the estimation of the dependence parameters under negative dependence.

Our proposed extensions to the conditional extreme value model are illustrated through a case study, which uses data from the River Severn in the UK.

Stochastic realisations generated from the conditional exceedance model hold for the d flow variables used in the analysis; however many reaches of a

[a] Corresponding author: ross.towe@jbatrust.org

river network will typically not have a measurement gauge, therefore a spatial extension is needed in order to characterise the full extent of a flood event. It is possible that the most severe river flow in a flood event may occur at such an ungauged location, and as a result events need to be interpolated so as to produce realistic predictions of extremes at ungauged locations along the river network.

Many existing interpolation methods use information solely about river flow, however we propose to also use information about rainfall. Rainfall is the main driver of river flow, so exploiting this known relationship is likely to be productive in improving the current interpolation methods.

The paper is structured as follows, Section 2 provides a review of the available river flow and rainfall data. Section 3 provides a review of the spatial extreme model we use to generate flood events; with details of the simulation algorithm as well as the different extensions that are adopted to handle missing values. Section 4 illustrates these different methods and the benefits of our proposed approach.

Section 5 details different interpolation techniques that can be used to obtain realisations along the river network as well as applications of the techniques to the River Severn data. Existing interpolation methods are first presented and then a new method is developed, which incorporates information about the behaviour of rainfall. Rainfall data are likely to help in the explanation of flood events at ungauged locations, as these gauges are likely to have smaller catchments, and are typically affected by flash flooding caused by heavy rainfall events. Finally conclusions and potential extensions are given in Section 6.

2 Data

For our study, the River Severn was chosen; the Severn is one of the largest rivers in the UK. The Severn has several gauging stations along its length, which allows us to explore how the extremal dependence between the gauges changes as a function of time and distance.

Daily mean river flow data and catchment boundaries were accessed from the National River Flow Archive (NRFA) at CEH Wallingford. The catchment boundaries were used to determine whether the river flow gauges lie within the catchment of the River Severn as well as to determine the nested structure of the gauging stations.

Daily rainfall data were obtained from CEH-GEAR (Centre for Ecology & Hydrology – Gridded Estimates of Areal Rainfall), a 1km resolution data set containing gridded estimates of daily rainfall [6]. For this analysis, a crossover period of the river flow and rainfall data was considered for the period 1990-2010.

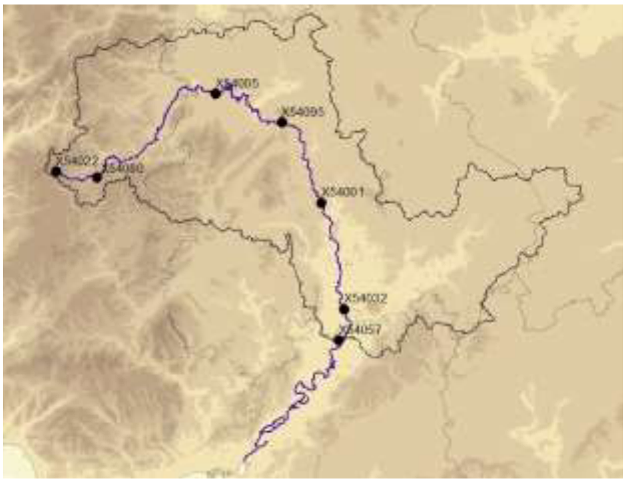

Figure 1: Outline of the Severn catchment. The solid black line shows the catchment boundary as defined by the most downstream gauge X54057. The purple line shows the main stem of the River Severn catchment. Gauging stations used in the analysis are labelled with their National River Flow Archive catalogue numbers. Shading represents terrain elevation (darker is higher ground).

Figure 1 provides an illustration of our study region and the gauges used on the main River Severn. We do not include gauges that are on tributaries within the catchment of gauge X54057 (Haw Bridge, near Gloucester) to simplify presentation.

3 Methodology

3.1 Statistical model

Standard multivariate extreme value models, corresponding to families of copula, typically only handle one class of extremal dependence (either asymptotic dependence or asymptotic independence) and this dependence structure has to be pre-determined before the model is fitted.

Asymptotic dependence means that there is a non-zero probability that extreme events of the same size occur simultaneously, whereas for asymptotic independence the probability of two extreme events of the same size occurring simultaneously is zero.

For example consider two sites, in the case of asymptotic dependence if one site observed a 1 in 100 year event, it is possible that the other site will also observe a 1 in 100 year event. However, for the case of asymptotic independence if one site observed a 1 in 100 year event, the other site is likely to observe an event with a much reduced level, for example at worst 1 in 10 year event. If an asymptotically dependent copula model is fitted to asymptotically independent data then the probability that two sites concurrently observe their 100 year levels will be over-estimated, and vice versa.

An alternative approach is the conditional extreme value model of Heffernan and Tawn [2], which estimates the form of extremal dependence structure as part of the fitting procedure so covers both asymptotic dependence and asymptotic independence.

Like other multivariate extreme value models, the conditional extreme value model is a two-step approach that models the marginal and dependence characteristics of a data set separately.

3.1.1 Marginal transformation

Consider n independent and identically vectors of $\boldsymbol{R} = (R_1, \dots, R_d)$ representing values of flow variables at d sites at different times.

Our model for the marginal distributions of $\boldsymbol{R}$ has two components, separated using a predetermined threshold level u_i for variable R_i. For those points below the threshold u_i, the empirical distribution of R_i is used and above the threshold, the generalised Pareto distribution (GPD) is adopted. Thus we have

$$F_i(r) = \begin{cases} \tilde{F}_i(r) & , \quad r < u_i \\ 1 - \varphi(u_i)[1 + \xi_i(r - u_i)/\sigma_i]^{-1/\xi_i}, & \quad r \ge u_i \end{cases}$$

equation 1

where $\tilde{F}_i(r)$ is a kernel smoothed empirical cumulative distribution function and $\varphi(u_i) = 1 - \tilde{F}_i(u_i)$ is the probability of an exceedance above the threshold u_i. The GPD is a two parameter distribution, with the scale parameter $\sigma_i > 0$ and the shape parameter $\xi_i \in \mathbb{R}$.

To estimate the dependence structure of the random variables $\boldsymbol{R}$, the data are transformed componentwise to common Laplace margins via the transform,

$$Y_i = \begin{cases} \log\{2F_i(R_i)\} & , \ F_i(R_i) < 0.5 \\ \text{-}\log\{2[1 - F_i(R_i)]\}, & \ F_i(R_i) \ge 0.5 \end{cases}, \text{ for } i = 1 \dots, d$$

equation 2

where F_i is given in equation 1. The resulting variable $\boldsymbol{Y} = (Y_1, \dots, Y_d)$ has Laplace margins, e.g. marginal densities

$$f(s) = \frac{1}{2} e^{-|s|}, \text{for } -\infty < s < \infty.$$

The transformation to Laplace margins means that $\mathbb{P}(Y_i > y + v | Y_i > v) = e^{-y}$ for $y > 0$, and $v > 0$. Therefore, the random variables $\boldsymbol{Y} = (Y_1, \dots, Y_d)$ now have an exponential upper tail, similarly lower tail, which is of importance when we consider the convergence of the conditional distribution.

3.1.2 Conditional dependence model

After making the transformation given in equation 2, the extremal behaviour of the joint tail of the random variables $\boldsymbol{Y}$ can now be determined. A key element of the conditional extreme value model is the conditioning variable to fit the model. This variable we denote by Y_1 and consider the extremal dependence of the $(d-1)$ remaining variables $\boldsymbol{Y}_{-1}$ conditional on Y_1 being above a sufficiently large value v, which we call the dependence threshold. The approach is motivated by the following asymptotic result, we assume that there exists normalising functions $\boldsymbol{a}_{|1}(Y_1)$ and $\boldsymbol{b}_{|1}(Y_1)$>0 such that for y_1>0 the following limit probability holds for $\boldsymbol{Y}_{-1} = (Y_2, \dots, Y_d)$

$$\lim_{v \to \infty} \mathbb{P}\left(\frac{\boldsymbol{Y}_{-1} - \boldsymbol{a}_{|1}(Y_1)}{\boldsymbol{b}_{|1}(Y_1)} \le \boldsymbol{z}, Y_1 - v > y_1 | Y_1 > v \right) \\ = \exp\{-y_1\} G_{|1}(\boldsymbol{z})$$

equation 3

where vector algebra is interpreted as componentwise and $G_{|1}(\boldsymbol{z})$ is non-degenerate in each margin. The first term in the limit given in equation 3 arises from the fact that Y_1 follows a standard Laplace distribution. The second term in the limit characterises the behaviour of $\boldsymbol{Y}_{-1}|Y_1 > v$ in terms of the limiting distribution function $G_{|1}(\boldsymbol{z})$ along with the location $\boldsymbol{a}_{|1}(Y_1)$ and scale function $\boldsymbol{b}_{|1}(Y_1)$>0.

As a result of equation 3, $G_{|1}(\boldsymbol{z})$ is the limiting conditional distribution of

$$\boldsymbol{Z}_{|1} = \frac{\boldsymbol{Y}_{-1} - \boldsymbol{a}_{|1}(Y_1)}{\boldsymbol{b}_{|1}(Y_1)}, \text{ given } Y_1 > v \text{ as } v \to \infty,$$

equation 4

where $\boldsymbol{Z}_{|1} \sim G_{|1}$ and we call $\boldsymbol{Z}_{|1}$ the residual of the conditional extreme value model. The result of the limits given in equation 3 and equation 4 is that $\boldsymbol{Z}_{|1}$ and Y_1 are independent given that $Y_1 > v$ in the limit as $v \to \infty$.

Heffernan and Tawn [2] found that although the different classes of extremal dependence have different forms for $\boldsymbol{a}_{|1}(Y_1)$ and $\boldsymbol{b}_{|1}(Y_1)$>0, they all can be written in a simple parametric form. Through assuming Laplace margins, this form simplifies to $\boldsymbol{a}_{|1}(Y_1) = \boldsymbol{\alpha}_1 y_1$ and $\boldsymbol{b}_{|1}(Y_1) = y_1^{\boldsymbol{\beta}_1}$ where $\boldsymbol{-1} \le \boldsymbol{\alpha}_1 \le \boldsymbol{1}$ and $-\infty < \boldsymbol{\beta}_1 < \boldsymbol{1}$ [5]. We further assume that the limiting assumptions hold exactly above a sufficiently large dependence threshold v. This leads to the following model:

$$\boldsymbol{Y}_{-1} = \boldsymbol{\alpha}_1 Y_1 + y_1^{\boldsymbol{\beta}_1} \boldsymbol{Z}_{|1} \text{ for } Y_1 > v, \\ \text{where } -1 \le \boldsymbol{\alpha}_1 \le 1 \text{ and } -\infty < \boldsymbol{\beta}_1 < 1,$$

where $G_{|1}(\boldsymbol{z})$ is a marginal non-degenerate distribution function and the $\boldsymbol{Z}_{|1}$ is independent of Y_1. When $\alpha = 1, \beta = 0$ the data are asymptotically independent.

There is no known general distributional form for $\boldsymbol{Z}_{|1}$, so we adopt the same approach as Heffernan and Tawn [2] by estimating the distribution of $\boldsymbol{Z}_{|1}$ non-parametrically. In order to do this we assume that $\boldsymbol{Z}_{|1}$ has a mean $\boldsymbol{\mu}_1$ and variance $\boldsymbol{\sigma}_1^2$.

As a result, the following expressions for the conditional expectation and variance of Y_i can be determined

$$\mathbb{E}[\boldsymbol{Y}_{-1} | Y_1 = \boldsymbol{y}_1] = \boldsymbol{\alpha}_1 y_1 + y_1^{\boldsymbol{\beta}_1} \boldsymbol{\mu}_1 \\ \mathbb{V}ar[\boldsymbol{Y}_{-1} | Y_1 = \boldsymbol{y}_1] = \left(y_1^{\boldsymbol{\beta}_1} \boldsymbol{\sigma}_1\right)^2$$

equation 5

for $y_1 > v$.

3.1.3 Inference

In order to estimate the dependence parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$, the working assumption is made that the $\boldsymbol{Z}_{|1}$'s stated in equation 4 follow a Gaussian distribution with mean and variance stated in equation 5 and are independent across components of $\boldsymbol{Z}_{|1}$. The estimation of these dependence parameters is performed through pairwise maximum likelihood for the n_v pairs with $y_1 > v$, where $i = 2, \ldots, d$. The likelihood is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{j=1}^{n_v} \frac{1}{\sqrt{2\pi}(y_{ij}^{\beta_i}\sigma_{\boldsymbol{i}})} exp\left\{-\frac{\left(y_{ij} - \left[\alpha_i y_{1j} + \mu_{\boldsymbol{i}} y_{1j}^{\beta_i}\right]\right)^2}{2\pi(y_{ij}^{\beta_i}\sigma_{\boldsymbol{i}})^2}\right\}$$

equation 6

where $-\infty < \mu_i < \infty$, $\sigma_i > 0$, $-1 \leq \alpha_i \leq 1$, and $-\infty < \beta_i < 1$, where y_{ij} denotes component i for the j^{th} exceedance of v by y_1. The forms of the mean and variance parameter given in equation 6 arises from the conditional expectation and variance given in equation 5.

Once maximum likelihood estimates for the parameters μ_i, σ_i, α_i and β_i are determined, realisations of Z_i, the i^{th} component of $\mathbf{Z}_{|1}$ corresponding to Y_i can obtained by using equation 7,

$$Z_{ij} = \frac{y_{ij} - \hat{\alpha}_i y_{1j}}{y_{1j}^{\hat{\beta}_i}}, \text{for } y_{1j} > v, where\, j = 1, \ldots, n_v.$$

equation 7

In order to account for spatial dependence the following expression is used to generate a temporally coherent vector of residuals.

$$\boldsymbol{Z}_{|1}^{(j)} = \frac{\boldsymbol{y}_{-1,j} - \hat{\alpha}_1 y_{1j}}{y_{1j}^{\hat{\beta}_i}}, \text{for } y_{1j} > v, \text{where } j = 1, \ldots, n_v.$$

equation 8

Once a sample of $\boldsymbol{Z}_{|1}$ is obtained from equation 8, this is used to obtain an empirical estimate of the joint distribution function $G_{|1}$. This estimated conditional model now enables us to model the distribution of $\boldsymbol{Y}_{-1}|Y_1 > v$. The same inference procedure holds for any $\boldsymbol{Y}_{-j}|Y_j > v$ for $j = 1, \ldots, d$. Consequently we have a model for the joint tail behaviour of $\boldsymbol{Y}$, when at least one component is large. This enables us to make predictions beyond the range of the observed data.

3.2 Handling missing values in the conditional extreme value model

In what follows, different ways of handling the residual distribution are discussed in Section 3.2.1. A detailed review of the set-up of the Keef et al. [4] modelled infill approach is given in Section 3.2.2 and this is compared with our approach in Section 3.2.3, which uses a Gaussian copula. The general simulation procedure of the conditional extreme value model is given in Section 3.2.1; this simulation procedure holds only for the d sites included in the analysis.

3.2.1 Handling missing values in the residual distribution of the conditional extreme value model

The standard methods given in Heffernan and Tawn [2] only consider vectors of complete observations (referred to as the **Heffernan** method in Tables 1 and 2). In many environmental applications, data are likely to be missing. As observed data can be very sparse, the Heffernan and Tawn [2] method is therefore restrictive and inefficient when modelling the extremes of network river flows.

Keef et al. [4] developed a strategy to estimate the distribution of the missing variables; we call this approach the **Modelled Infill**. We extend this approach by adopting what is known as a **Gaussian copula** based approach.

The two extensions of the methods, the **Modelled Infill** and the **Gaussian copula** have some similarities as both model the residual distributions copula by using a Gaussian copula, missing data and all data respectively. A Gaussian copula is chosen because it is computationally feasible in higher dimensions and the dependence between variables requires only pairwise data. It is easiest to understand the model for the joint distribution of the resulting $\boldsymbol{Z}_{|1}$ if we marginally have to transform the $(\boldsymbol{Z}_2 \ldots, \boldsymbol{Z}_d)$ to have standard Normal margins. The probability integral transform for Z_{i} is

$$Z_i^N = \Phi^{-1}[\tilde{F}_i(Z_i)] = \Phi^{-1}\left[\frac{1}{n}\sum_{j=1}^{n} \Phi\left(\frac{Z_i - z_{ij}}{h_i}\right)\right],$$
$$i = 2, \ldots, d$$

equation 9

where Z_i^N represent the residuals on Normal margins, and Φ is the cumulative distribution function of a standard Normal, $\tilde{F}_i$ is the kernel smoothed marginal distribution function of standard Normal residuals $\boldsymbol{Z}_{\mathrm{i}}$. Here $Z_{1i}^N, \ldots, Z_{ni}^N$ denote the n realisations of $\boldsymbol{Z}_{\mathrm{i}}$ and h_{i} is the bandwidth [7]. Unlike the standard Heffernan and Tawn [2] approach the residuals are no longer restricted to the sample as the kernel smoothing allows both interpolation and limited extrapolation of the residuals.

The use of the probability integral transformation in equation 9 results in each $Z_i^N \sim \mathrm{N}(0,1)$ and we make the assumption that the copula is Gaussian with distribution

$$\begin{pmatrix} Z_2^N \\ \vdots \\ Z_d^N \end{pmatrix} \sim MVN_d\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \cdots & \rho_{2d} \\ \vdots & 1 & \vdots \\ \rho_{d2} & \cdots & 1 \end{bmatrix}\right),$$

equation 10

the assumption that the copula is Gaussian gives us a pairwise relationship between each pair of residuals. A formal test to check this whether the copula is Gaussian is given in Section 3.2.5.

The concurrent pairs of observations of $\boldsymbol{Z}^N = (\mathrm{Z}_2^{\mathrm{N}}, \dots, \mathrm{Z}_{\mathrm{d}}^{\mathrm{N}})$ are used to estimate the correlation parameters provided that data exists for a given $\mathrm{Z}_{\mathrm{i}}^{\mathrm{N}}$ and $\mathrm{Z}_{\mathrm{j}}^{\mathrm{N}}$. This gives the following estimated correlation matrix $\hat{\Sigma}$

$$\hat{\Sigma} = \begin{bmatrix} 1 & \dots & \hat{\rho}_{1d} \\ \vdots & 1 & \vdots \\ \hat{\rho}_{d1} & \dots & 1 \end{bmatrix}$$

where

$$\hat{\rho}_{ij} = \frac{\Sigma_k (z_{i,k} - \bar{z}_i)(z_{j,k} - \bar{z}_j)}{\sqrt{\Sigma_k (z_{i,k} - \bar{z}_i)^2 \Sigma_k (z_{j,k} - \bar{z}_j)^2}}$$

k relates to the sum over the parts of the time series when both z_i and z_j are observed, an estimate of the correlation. The sets over which k is summed will change for different pairs of sites.

As the correlation matrix is estimated for non-overlapping data sets, there is a possibility that the matrix is not positive semi-definite, however there are eigen-decomposition methods that can solve this problem [8].

3.2.2 Modelled infill approach

If the residual data set $\boldsymbol{Z}^N$ can be modelled by using a Gaussian copula, we can use standard results to impute missing values. For example, we partition the incomplete data such that the observed subset of data $\mathbf{Z}_1^{\mathrm{N}}$ is of dimension q and the remaining missing data in the incomplete data set $\mathbf{Z}_2^{\mathrm{N}}$ are of dimension $(d-1-q)$. This results in the mean and covariance matrix of the Gaussian copula being partitioned as follows,

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, where μ has $\begin{bmatrix} q \\ (d-1-q) \end{bmatrix}$ dimensions with the following covariance matrix,

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

equation 11

where

Σ has $\begin{bmatrix} q \times q & q \times (d-1-q) \\ (d-1-q) \times q & (d-1-q) \times (d-1-q) \end{bmatrix}$,

equation 12

dimensions. Using the definitions in equation 11 and equation 12 the conditional distribution is defined as follows,

$$\mathbf{Z}_2^{\mathrm{N}} \,|\mathbf{Z}_1^{\mathrm{N}} = \mathbf{z}_1^{\mathrm{N}} \sim \mathrm{MVN}(\bar{\mu}, \bar{\Sigma}),$$

equation 13

where $\bar{\mu} = \Sigma_{21}\Sigma_{11}^{-1}\mathbf{z}_1^{\mathrm{N}}$ and $\bar{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. The result given in equation 13 allows us to model those observations that are missing, and hence to infill their values.

Using the conditional Gaussian copula given in equation 13 produces a simulated sample of $\mathbf{Z}_2^{\mathrm{N}}$ together with the observed $\mathbf{Z}_1^{\mathrm{N}}$. Before these simulated residuals can be used to simulate samples from the Heffernan and Tawn model we need to convert the residual back to their original margins. In order to do this, we solve equation 14 to find $\mathrm{Z}_{\mathrm{i}}, i = 2, \dots, d$.

$$\Phi(Z_i^N) = \frac{1}{n}\sum_{j=1}^{n} \Phi\left(\frac{Z_i - z_{ij}}{h_i}\right).$$

equation 14

This sample of residuals $\boldsymbol{Z}$ can be used as part of the simulation procedure given in Section 3.2.4.

3.2.3 Gaussian copula approach

The methodology in Section 3.2.2 details the set-up of the **Modelled Infill** approach as the missing values within the incomplete data set are modelled by using the conditional Gaussian copula as defined in equation 13. This Modelled Infill approach can become computationally costly when large scale data sets are considered as there will be a large number of missing values.

The **Gaussian copula** instead uses the estimated covariance matrix defined in equation 12 to simulate a new residual distribution on Normal margins. This residual distribution $\boldsymbol{Z}^N$ is based on the observed dependence between pairs of residuals by using the distributional assumption given in equation 10. Equation 14 is used to transform $\boldsymbol{Z}^N$ to $\boldsymbol{Z}$, variables on their original margins.

Simulating directly from the Gaussian copula rather than the conditional distribution of the observed residuals becomes increasingly computationally efficient when a larger percentage of data are missing. Furthermore, the Gaussian copula allows us to simulate plausible combinations of residuals that we have not observed.

To verify the assumption that the dependence of the residuals on Normal margins $\boldsymbol{Z}^N$ can be represented by a Gaussian copula, the following tests in Section 3.2.5 can be performed. The Gaussian copula has an asymptotically independent extremal dependence structure, however this assumption of asymptotic independence is not restrictive as the tails of $\boldsymbol{Z}$ are not vital for determining the joint tails of $\boldsymbol{Y}$.

3.2.4 General simulation procedure with missing values

The general simulation procedure detailed is an adaptation of the algorithm in Keef et al [4]. Simulations are typically generated such that the conditioning variable Y_1 is above a certain level, for example the 1 in 100 year level. We denote this level by v_p. The steps of the simulation procedure are outlined as follows:

1. Draw a value of residual $\boldsymbol{Z}_{|1}$ this is where the different methods deviate as in some cases there are missing values in the observed residual data. This step changes depending on the method used:
   a. Block resampling from the residual distribution $\mathbf{Z}$ and infilling the missing values using the conditional multivariate Normal distribution: this is the **Modelled Infill** approach
   b. Simulate from a Gaussian copula with correlation matrix $\hat{\Sigma}$, to produce a sample $\boldsymbol{Z}^N$, which is transformed to

Laplace margins to give **Z**: this is used in the **Gaussian copula** approach

2. Draw a value of conditioning variable Y_1 from a standard Exponential distribution above the level of interest $v_p > v$, where v is the dependence threshold. For example, $Y_1 = v_p + \text{Exp}(1)$.
3. Derive the simulated value of the unconditioned variates $\boldsymbol{Y}_{-1}$, which is a function of $Y_1, \boldsymbol{Z}_{|1}$ and the estimate of the dependence parameters $\boldsymbol{\alpha}_{|1}$ and $\boldsymbol{\beta}_{|1}$. The formula is given below,
$$\boldsymbol{Y}_{-1} = \hat{\alpha}_{|1} Y_1 + Y_1^{\hat{\beta}_1} \boldsymbol{Z}_{|1}, \text{ for } Y_1 > v$$
4. The sample of $\boldsymbol{Y}$ have common Laplace margins; the probability integral transform as given in equation 2 can be used to transform the sample back to its original margins.

3.2.5 Testing whether the residual distribution can be characterised by a Gaussian copula

In order to assess whether the residual distribution $\boldsymbol{Z}^N$ can be characterised using a Gaussian copula a formal test can be conducted, which can handle both complete and incomplete data. In both cases the null hypothesis is that on Normal margins the residual distribution follows a multivariate Normal distribution. In order to assess the higher order dependence structure of the residual distribution, we adopt methods developed by Bortot et al. [9].

Consider the set of observations $\mathbf{Z}^N = (\mathrm{Z}_2^{\mathrm{N}}, \dots, \mathrm{Z}_{\mathrm{d}}^{\mathrm{N}})$, which have standard Normal margins and covariance Σ, with the square of the Mahalanobis distance defined

$$T = \mathbf{Z}^N \Sigma^{-1} \mathbf{Z}^{N\prime}.$$

If $\boldsymbol{Z}^N$ follows a multivariate Normal distribution with covariance matrix Σ then T follows a $\boldsymbol{\chi}^2_{d-1}$ distribution, where $d-1$ is the dimension and with $\mathbb{E}[\boldsymbol{\chi}^2_{d-1}] = d-1$ and $\mathbb{V}\boldsymbol{ar}[\boldsymbol{\chi}^2_{d-1}] = 2d-1$.

In reality, missing values are present in the residual distribution of $\boldsymbol{Z}^N$ and the percentage of missing values is not consistent across locations. Therefore the test statistic T has to account for the different record lengths of data. If we consider a particular vector of the data $\mathbf{Z}_{\mathrm{i}}^{\mathrm{N}}$ with dimension $d_i \leq d$. Then $\mathbf{Z}_{\mathrm{i}}^{\mathrm{N}} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\dim(\boldsymbol{\Sigma}_i) = d_i \times d_i$ corresponding to the rows/columns of $\boldsymbol{Z}^N$ that are observed in $\mathbf{Z}_{\mathrm{i}}^{\mathrm{N}}$. Then

$$T_i = \boldsymbol{Z}_i^N \Sigma_i^{-1} \boldsymbol{Z}_i^N$$

equation 15

follows a $\boldsymbol{\chi}^2_{d_i}$ distribution with $\mathbb{E}[\boldsymbol{\chi}^2_{d_i}] = d_i$ and $\mathbb{V}\boldsymbol{ar}[\boldsymbol{\chi}^2_{d_i}] = 2d_i$. Through using equation 16, we can define the test statistic to be

$$T_{miss} = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i - d_i}{\sqrt{2d_i}}$$

equation 16

where n corresponds to the number of observations of $\boldsymbol{Z}^N$. If a particularly large value of T_{miss} is observed then there is a deviation away from the assumption of multivariate Normality. The sampling distribution of T_{miss} under the null hypothesis for a given pattern of missing data is easily derived by Monte Carlo methods.

4 Application of the missing values methods to generate values of the conditioned variates

4.1 Outline

We now apply the methodology given in Section 3 to the River Severn flow data described in Section 2. Conditional probabilities relating to certain scenarios are calculated to illustrate how these missing value methods aid in the calculation of probabilities for flood risk management. Finally, the section also illustrates the added benefit of modelling the residual distribution in the Heffernan and Tawn model with a Gaussian copula.

The application given here is an illustration of the developed methodology given in Section 3.2. As the application is an illustration, we only focus on spatial dependence and consider concurrent observations. The derived methodology can easily be extended to incorporate temporal dependence, details of this can be found in Keef et al. [4] and Lamb et al. [3].

The seven daily mean flow gauges as given in Figure 1 were used in the analysis. The most upstream gauge, X54022, was chosen as the conditioning location Y_1. The daily mean flow data set has very few missing values (1% of the total observations) for the time period of interest.

4.2 Statistical analysis of the River Severn flow data

4.2.1 Calculations of the conditional probabilities for flood risk management

The main benefits of modelling the missing values is best illustrated when we calculate probabilities of particular events. For example, one question that might be of interest is as follows: if the flow is large at the most upstream gauge (Y_1), what is the probability that one of the remaining river flow gauges is also large? To answer this, let $Y_{(1)} > \dots > Y_{(6)}$ be the ordered values of $(Y_2, \dots, Y_7)$. Then we are interested in the following probability,

$$p_p = \mathbb{P}(Y_{(1)} > y_p \mid Y_1 > y_p),$$

equation 17

whereby y_p corresponds to a sufficiently large value such as the 1 in 100 year event. As the $\boldsymbol{Y}$'s are on the same scale, if Y_1 exceed a 100 year event, it is easy to determine whether any of $\boldsymbol{Y}_{-1}$ also exceeds the same level. The probability given in equation 17 corresponds to there being at least one of the sites $(Y_2, \dots, Y_7)$ being larger than y_p given $Y_1 > y_p$. Here y_p is chosen to be either a 100, 500, 1000 or 10000 year return period with corresponding conditional probabilities $p_{0.01}, p_{0.002}, p_{0.001}, p_{0.0001}$. The corresponding estimates and 95% confidence intervals are given in Table 1.

Prob	Obs	Heffernan	Modelled Infill	Gaussian copula
$p_{0.01}$	0 (NA)	0.053 (0.026, 0.116)	0.060 (0.025,0.132)	0.059 (0.027,0.118)
$p_{0.002}$	0 (NA)	0.036 (0.017,0.080)	0.043 (0.017,0.098)	0.039 (0.018,0.085)
$p_{0.001}$	0 (NA)	0.031 (0.014,0.070)	0.037 (0.015,0.088)	0.036 (0.016,0.076)
$p_{0.0001}$	0 (NA)	0.023 (0.005,0.031)	0.025 (0.007,0.041)	0.023 (0.007,0.033)

Table 1: The estimates with the 95% confidence intervals for the conditional probability given in equation 16.

An estimate of the 95% confidence interval conditional probability given in equation 17 cannot be obtained empirically. This clearly emphasises the need to fit a statistical model and in particular the use of extreme value theory to provide us with reliable predictions of events we have not yet observed.

4.2.2 Second example of the calculation of conditional probability for flood risk management

The conditional probability given in equation 17 is again calculated, but for this particular example, a larger percentage of missing values are observed. A naive application of Heffernan and Tawn [2] would consider only the times in which all of the variables are observed. However from Figure 2 note that there are no periods of time in which all seven river flow gauges have observed data, and so the existing method of Heffernan and Tawn [2] cannot be applied because, for example, if we condition on gauge X54022 (Y_1), we would need the empirical joint distribution of the remaining gauges, which cannot be estimated for these data.

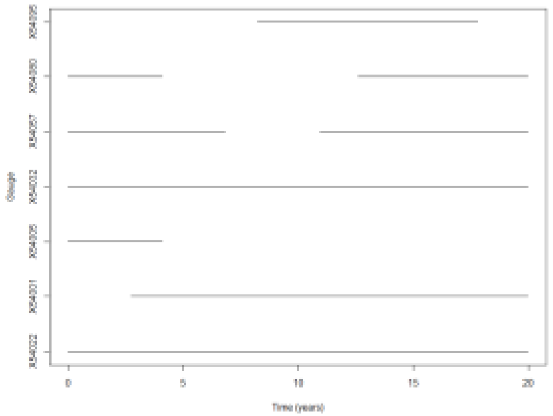

Figure 2: The time series plot shows the seven river flow gauges for the River Severn, with a larger percentage of missing

For many applications this naive approach is therefore highly restrictive as it leaves a small proportion of data from which to make predictions of extreme events.

Prob	Obs	Heffernan	Modelled Infill	Gaussian copula
$p_{0.01}$	NA	NA	0.026 (0.007,0.978)	0.029 (0.007,0.091)
$p_{0.002}$	NA	NA	0.014 (0.002,0.076)	0.017 (0.002,0.066)
$p_{0.001}$	NA	NA	0.011 (0.001,0.068)	0.012 (0.001,0.057)
$p_{0.0001}$	NA	NA	0.006 (0.000,0.036)	0.006 (0.000,0.024)

Table 2: The estimates with the 95% confidence intervals for the conditional probability given in equation 16.

4.2.3 Further benefits of the Gaussian copula based approach

The existing modelled infill approach of Keef et al. [4] uses the empirical distribution to model the marginal distribution of the observed residuals, i.e. it assumes the only future values for residuals are the observed values. The Gaussian copula based approach instead uses a kernel smoothed distribution function to capture the marginal behaviour and the Gaussian copula marginal and joint residuals can both differ from observed values.

This smoothed distribution function allows smooth interpolation between observed data points as well as limited extrapolation. Furthermore, extreme events simulated under the Heffernan and Tawn [2] model are no longer restricted to deterministic functions of events already observed and instead a cloud of possible combinations can be simulated. The effects of this development can be seen in Figure 3. Here we show simulated values for gauges X54032 and X54057 given that gauge X54095 observed a 100 year event from using both the Keef et al. [4] approach and our method. Pairwise plots of the two gauges are compared in Figure 3 with the gauges plotted on the same scale to make the comparison easier. In Figure 3 the new method is able to capture the variability in the extremes unlike the existing method. The starkest contrast between the two methods in Figure 3 is that the spread of the simulated points in the right hand plot is much wider. This additional variation seems realistic given the extremal behaviour of the observed data set. In the existing method only 7.1% of the simulated sample are unique, whereas for the new method 100% of the points are unique.

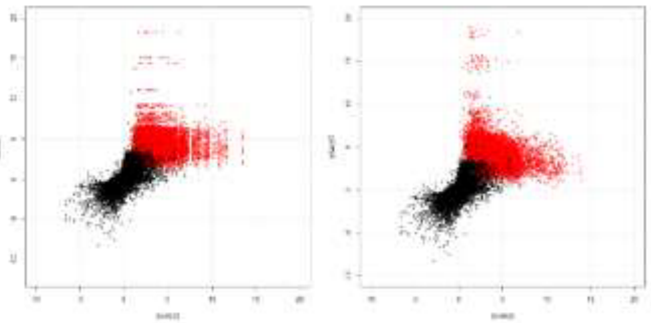

Figure 3: Observed (black) and joint behaviour of gauge X54032 and X54057 and simulated (red) given that a 1 in 100 year event was observed at X54095. Left: the existing method (modelled infill); right the newly developed approach (Gaussian copula). In both figures the data are shown after transformation to Laplace margins.

5 Interpolation of flood events using information about rainfall

The River Severn catchment is well gauged. However, there are still many reaches within the catchment that are not gauges, and many smaller river catchments are likely to be ungauged. Therefore, a methodology needs to be developed that produces reliable predictions of river flow in both gauged and ungauged catchments.

If a particular catchment is ungauged or the only upstream flow gauge is contained in the headwaters of the catchment then an alternative source of information is to use rainfall data. This study aims to show that an increase in the predictive performance of a spatial interpolation scheme for extremes modelled at gauges can be obtained by using the extra information provided by the rainfall data.

The interpolation scheme study in this paper focuses on river flow gauges contained with the catchment of gauging station of X54057. The aim of the analysis in Section 5 is to develop a flexible methodology to produce reliable predictions of future river flow at X54057 and ultimately for future simulated events at a number of sites along the river network.

The methodology for the interpolation scheme is constructed with the data being on common margins, the choice of Laplace margins is the same marginal distribution as used in the dependence model of Heffernan and Tawn [2]. The choice of modelling the data on Laplace margins is approximately equivalent to modelling the data on the log-return period scale.

5.1 Interpolation methods

There are three interpolation methods that are primarily compared; the nearest neighbour approach, an inverse distance weighted approach and finally the proposed approach, which combines information about river flow and rainfall. Each of these three interpolation methods are introduced in turn in Sections 5.1.1 to 5.1.3.

5.1.1 Method 1: Nearest neighbour approach

The simplest interpolation method would be to use the river flow from the nearest gauging station. The nearest river flow gauge is typically determined by a metric such as the Euclidean or the hydrological distance.

5.1.2 Method 2: Inverse distance weighted approach

An alternative approach was adopted by Keef et al. [4], this approach is otherwise known as an inverse distance weighted (IDW) approach, The approach uses a weighted sum of the nearby gauges. For example consider that $\boldsymbol{Y}^*$ is the flow gauge that we wish to predict with observations $Y_1^*, \dots, Y_m^*$ and that there are n gauges, $\boldsymbol{Y_1}, \dots, \boldsymbol{Y_n}$, which provide covariate information; the prediction of the river flow at gauge Y_j^* observed at time j becomes

$$\hat{Y} = \mathbb{E}[Y_j^*] = \gamma \sum_{i=1}^{n} w_i \, Y_{ij}$$

equation 18

where $w_1, \dots, w_n$ are the corresponding positive weights that satisfy the constraint $\sum_{i=1}^{n} w_i = 1$ and $\gamma > 0$ is an estimated parameter. The weights defined in equation 18 are obtained as follows

$$w_i = \frac{{d_i}^{-1}}{\sum_{i=1}^{n} {d_i}^{-1}},$$

where d_i is a distance metric between each gauge Y_i and Y^*. There are also extensions to the IDW approach that account for the proportion of the catchments that overlap [3].

5.1.3 Method 3: Proposed approach

The interpolation approaches given in Section 5.1.1 and 5.1.2 only use available information about river flow. However, other important information can be obtained from the rainfall that falls within the catchment of the flow gauging station that defines the catchment. For example, let R_j be the rainfall that fell within the catchment at time j, with each term R_j having common Laplace margins (equation 1). Therefore, the statistical model becomes

$$\hat{Y} = \mathbb{E}[Y_j^*] = \alpha + \gamma \sum_{i=1}^{n} w_i \, Y_{ij} + \sum_{l=0}^{l_n} \beta_l R_{j-l},$$

equation 19

where $\alpha \in \mathbb{R}$ is the intercept term and the $\boldsymbol{\beta}_l \in \mathbb{R}$ are the contribution of each of the days of rainfall. The value l_n represents the number of lags of rainfall days included in the regression model.

5.1.4 Estimation procedure of the proposed approach

The unknown parameters in equation 18 have to be estimated. Least absolute deviations (LAD) is used to obtain parameter estimates of $(\alpha, \gamma, \beta_0, \dots, \beta_{l_n})$. The LAD is similar to the least squares estimation but instead minimises the sum of absolute errors. For the $j = 1, \dots, m$ observations of river flow at site Y^*, the estimation is defined as follows

$$LAD = \sum_{j=1}^{m} \left| Y_j^* - \left(\alpha + \gamma \sum_{i=1}^{n} w_i \, Y_{ij} + \sum_{l=0}^{l_n} \beta_l R_{j-l} \right) \right|$$

equation 20

the minimisation given in equation 20 is equivalent to a maximum likelihood estimation as the errors of the regression model given in equation 19 have a Laplace distribution.

The methodology stated so far in Section 5.1.3 has dealt with statistical models that are fitted to the whole distribution of Y^*, however the main focus of the interpolation technique is to improve the predictions of flood event footprints.

In order to do this we only focus on a subset of the largest Y^*'s and repeat the optimisation given in equation 20. For example let $Y^*_{(1)} > \cdots > Y^*_{(m)}$ be the ordered values of Y^* and we only consider those m_u points Y^* greater than an arbitrarily large threshold u. The subset of points $Y^*_{(1)} > \cdots > Y^*_{(m_u)}$ satisfy the constraint that $Y^*_{(m_u)} > u$ and are taken along with the respective values of Y and R_t such that LAD is now minimised above the value u

$$LAD_u \ = \sum_{j=1}^{m_u} \left| Y^*_{(j)} - \left(\alpha + \gamma \sum_{i=1}^{n} w_i \, Y_{ij} + \sum_{l=0}^{l_n} \beta_l R_{j-l}\right)\right|.$$

equation 21

The estimated parameters given in equation 21 are likely to be different to those in equation 20.

5.2 Statistical summaries to assess the predictive performance of the interpolation methods

Section 5.1 proposed a number of interpolation techniques, an initial comparison is to plot up the proposed predictions $\hat{\boldsymbol{Y}}$ against the response of interest $\boldsymbol{Y}^*$. However, other more robust test statistics need to be considered to fairly compare between the three interpolation techniques. Suitable statistics include the Mean Absolute Error and the coefficient of determination otherwise known as the R^2.

5.2.1 Mean Absolute Error

The mean absolute error (MAE) is a typical statistic that is used to measure the difference between a set of predictions and outcome. The mean absolute error is defined as follows

$$MAE = \frac{1}{m_u} \sum_{j=1}^{m_u} |y_j^* - \hat{y}_j|$$

equation 22

where $\hat{y}$ are predictions obtained from the regression model. If $m_u = m$ the MAE is calculated for the entire data set. The optimal value of the MAE is obtained through the objective function given in equation 20.

5.2.2 Coefficient of determination

The MAE given in Section 5.2.1 gives us an indication of the performance of our proposed regression model, however it is insufficient in determining the relative performance against the other candidate regression models in Section 5.1. In order to fairly compare the performance of the proposed methods, statistics such as the coefficient of determination can be calculated, which is denoted by R^2. The coefficient of determination is defined as

$$R^2 = 1 - \frac{\sum_{j=1}^{m_u} \left(y_j^* - \hat{y}_j\right)^2}{\sum_{j=1}^{m_u} \left(y_j^* - \bar{y}\right)^2}$$

equation 23

where y^* are the m_u observations of flow at the gauging station location of interest. The $\hat{y}$ are the estimated predictions from the fitted regression model. In the denominator, the mean of the observations $\bar{y}$ is subtracted from the predictions y^*. If $R^2 = 1$ then the fitted regression model perfectly fits the data; if $R^2 = 0$ the regression model does not fit the data at all. The coefficient of determination R^2 given in equation 23 gives a statistical summary of the performance of the fitted regression model. An issue is that the statistic does not account for if the model is over parameterised.

To account for this possibility, the adjusted coefficient of determination $\bar{R}^2$ is used and is typically defined as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{m_u - 1}{m_u - p - 1},$$

equation 24

where p is the number of explanatory variables and m_u the number of observations.

If an explanatory variable is unnecessarily incorporated into the regression model, the adjusted $\bar{R}^2$ will decrease. The optimal regression model will have the largest adjusted $\bar{R}^2$ but not necessarily the largest R^2.

The unadjusted and adjusted R^2 given in Section 5.2.2 is defined only for data on the L_2 norm, in other words for data that are Normally distributed. As our dependence modelling is performed on Laplace margins, this assumption does not hold and we need to consider test statistics that hold on the L_1 norm.

5.2.3 Coefficient of determination for Laplace data

The previous R^2 given in equation 23 is defined only for data on the L_2 norm and results in considering squared differences between the response and corresponding predictions. As we are now working on the L_1 norm, we instead consider absolute deviations. This change in metric in determining the distance between points, results in the R^2 becoming $|R|$, in reference to the change in norm,

$$|R| = 1 - \frac{\sum_{j=1}^{m_u} \left| y_j^* - \hat{y}_j \right|}{\sum_{j=1}^{m_u} \left| y_j^* - \tilde{y} \right|}$$

equation where $\tilde{y}$ is the median of the observed responses. The adjusted $|\bar{R}|$ is then similarly as in equation 24 but with R^2 replaced with $|R|$.

5.3 Exploratory analysis of the different interpolation methods

5.3.1 River Severn case study

In order to compare the predictive performance of the three different interpolation methods. The furthest downstream gauge, X54057 is selected as our prediction location Y^*. The combination of the other river flow gauges are used as explanatory variables. From exploratory analysis, it was found that it was beneficial to use the intervening area catchment to define R_t instead of the entire catchment of X54057. The map given in Figure 4 illustrates what we define as the intervening area if we had river flow observations at gauge X54032.

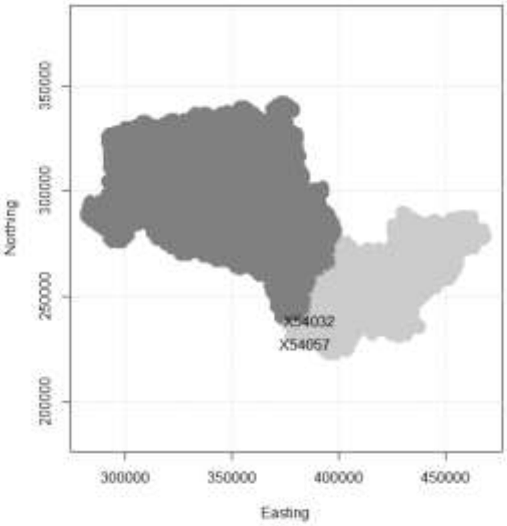

Figure 4: Illustration of the River Severn catchment as defined by gauge X54057. The solid grey area shows the catchment as defined by the upstream gauge X54032. The light grey area is what we define as the intervening area, in other words the part of X54057's catchment that is not gauged by X54032.

An initial comparison of the three different interpolation techniques is shown as a scatterplot of the predictions against the response of interest. The third method requires an upfront determination of the number of rainfall days, for this example l_n was chosen to be equal to 30. More consideration into the choice of $l_n$is given in Section 5.2.2. The scatterplots shown in Figure 5 are shown for a range of quantiles, specifically for the 0th, 50th, 90th and 95th percentiles; the parameters are obtained through the optimisation given in equation 21. The pattern of the scatterplots given in Figure 5 clearly changes across the range of quantiles, for the whole data set (q_0) there seems to very little difference between the three methods, however for the highest threshold ($q_{0.95}$), the new method shown by the blue is outperforming the other two methods. In order to formally assess the behaviour seen in Figure 5 the MAE (defined in equation 22) is calculated for each of the four percentiles. The estimated values of MAE are given in Table 3, for the lowest percentile, the nearest neighbour approach performs best. However, when predictions of the largest values of $Y^*$are produced the approach (method 3) that combines both river flow and rainfall performs best.

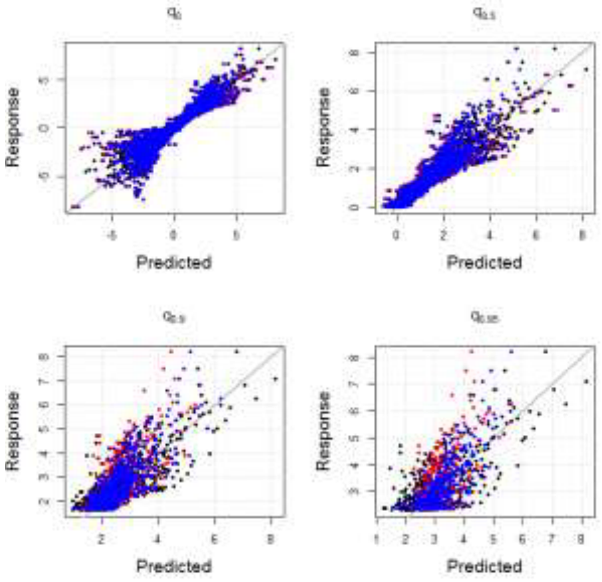

Figure 5: The four scatterplots left to right show the predictions above the 0th, 50th, 90th and 95th percentile. The three interpolation methods are the nearest neighbour approach (black), the inverse distance weighted approach (red) and the combined river flow and rainfall approach (blue).

The predictions obtained from method 3 in both Figure 5 and Table 3 were estimated from 31 lagged days if rainfall information. The choice of 31 days' worth of rainfall information was purely an arbitrary choice, in fact another regression model may be preferred with fewer days of rainfall information.

Percentile	Method 1	Method 2	Method 3
q_0	0.247	0.266	0.265
$q_{0.5}$	0.190	0.211	0.205
$q_{0.9}$	0.451	0.452	0.417
$q_{0.95}$	0.629	0.572	0.461

Table 3: The three columns show the three different methods as given in Section 5.1. The four different rows correspond to the 0th, 50th, 90th and 95th percentile.

5.3.2 Coefficient of determination

The newly proposed model given in equation 19 is investigated further to determine the optimal number of days of rainfall information. This comparison is performed by fitting the regression model given in equation 19 for $l_n = 0, \ldots, 30$ and the value of the unadjusted and adjusted $|R|$ is recorded (given in Section 5.2.3). The latter will help to determine the optimal number days of rainfall to include in the regression model.

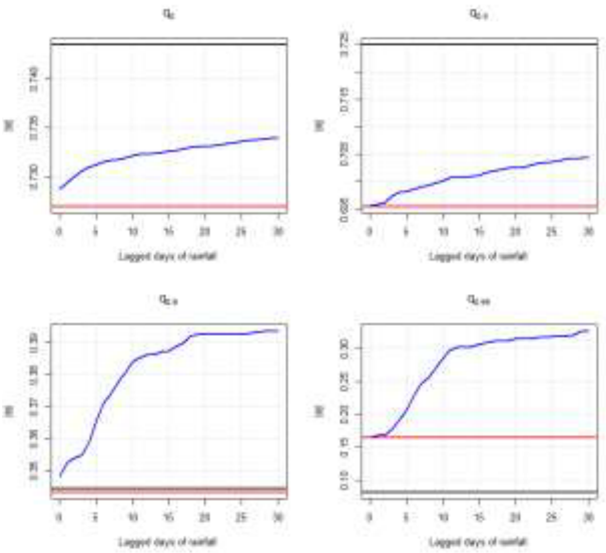

Figure 6: The four figures left to right show the unadjusted coefficient of determination $|R|$ for percentiles 0th, 50th, 90th and 95th. In each figure, the black line shows method 1, the red line shows method 2 and the blue method 3.

The coefficient of determination $|R|$, given in Figure 6 illustrates the pattern seen in Table 3 as for the lowest percentiles, the nearest neighbour approach (method 1) performs best). However, when we consider the higher percentiles, the addition of rainfall (method 3) is most beneficial and outperforms the two existing methods. Reassuringly the rainfall and river flow method outperforms the method that solely uses information about river flow. A general pattern also arises in the estimated $|R|$, this is that as the percentile increases, the value of $|R|$ decreases. This is to be expected as there is likely to be more unexplained variability in the extremes that relate to pluvial flood events.

The comparisons of the statistics given in Figure 6 are useful, however they do not account for the possibility that superfluous covariates have been included in the regression model given in equation 19. As a result, the adjusted $|\bar{R}|$ is calculated and compared. The pattern in Figure 7 is similar with that in Figure 6, as method 3 is still consistently outperforms method 2. For the lower percentiles (0th and 50th), there is still evidence to include more days of rainfall information. However, for the higher percentiles (90th and 95th), accounting for the number of parameters in the regression model has had a clear impact. In both cases, there is evidence to include rainfall information for a period of up to 11 days. Ultimately, for the 5% largest values observations of river flow, the inclusion of these rainfall data results in an extra 10% of the variability of Y^* being explained.

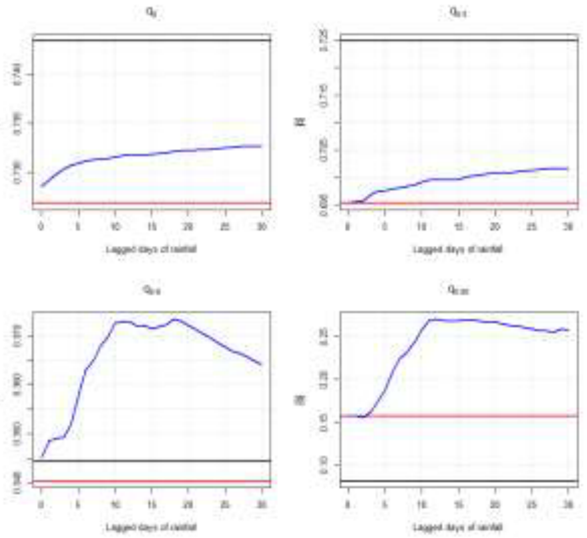

Figure 7: The four figures left to right show the adjusted coefficient of determination $|R|$ for percentiles 0th, 50th, 90th and 95th. In each figure, the black line shows method 1, the red line shows method 2 and the blue method 3.

6 Concluding remarks

This paper has illustrated how the conditional exceedance model of Heffernan and Tawn [2] can be used to aid flood risk assessment. Extensions to the model have been outlined and illustrated by using a case study related to the River Severn.

The conditional probabilities calculated in Section 3 consider equally extreme events at different sites, however the methodology can easily be extended to consider more general extreme events.

The existing approach of Keef et al. [4] resampled from the residual distribution and simulated missing values by conditioning on the gauges that were observed. This modelled infill approach becomes incredibly computationally expensive when the dimension of the data increases. The Gaussian copula, proposed here, offers a more efficient alternative as it is still relatively easy to simulate from the residual distribution even for data sets of a high dimension.

An issue of the methodology is that if there are no time overlap between data at different gauges there is no clear way of estimating the covariance matrix. A solution to this is to create an artificial time series from neighbouring time series and use that to estimate the correlation with the other sites in the analysis.

The benefits of the Gaussian copula based approach are improved marginal distributional modelling by using a kernel smoothed instead of an empirical distribution function. This allows for the generation of the physically realistic events that we have not yet observed. So we are no longer restricted to resampling from the residual distribution and the rays present in the left hand plot of Figure 3 will no longer occur.

The assumption of the Gaussian copula means that the residual distribution can be represented by a number of correlation parameters. In higher dimensional examples, the number of correlation parameters can become increasingly large, for example for $n = 100$ gauges there would be a total of $\sum_{i=1}^{n-1} i = 4950$ parameters. There exists ways of simplifying this correlation matrix, for example if we assume the residual distribution is a Gaussian process.

The methodology given in Section 5 provides an initial outline into how existing interpolation techniques can be extended to include information about rainfall. Using rainfall data from intervening areas of the catchment rather than the entire catchment proved the most beneficial, as it does not result in double counting the same covariate information. A possible extension is to simplify the regression model given in equation 19, as the model has an individual parameter for each lagged day of rainfall.

In conclusion, this paper has proposed a new more computationally efficient to deal with missing values in the residual distribution of the conditional extreme value model as well as illustrations of how rainfall data can be used to interpolate flood events across the river network.

Acknowledgments

Ross Towe would like to thank Jeremy Benn Associates Ltd. and Innovate UK for funding KTP009454. Ross would also like to thank Lancaster University, JBA Consulting, JBA Trust and JBA Risk Management for their support. We gratefully acknowledge the use of the daily mean river flow and the catchment boundary data, which were provided by the National River Flow Archive (NRFA). The CEH-Gear gridded rainfall data were provided by the Centre of Ecology and Hydrology.

7 References

1. Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *arXiv preprint arXiv:1501.02663*
2. Heffernan J. E. and Tawn J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **66(3)**, 497–546
3. Lamb, R., Keef, C., Tawn, J., Laeger, S., Meadowcroft, I., Surendran, S., Dunning, P. and Batstone, C. (2010). A new method to assess the risk of local and widespread flooding on rivers and coasts. *Journal of Flood Risk Management*, ***3*(4)**, 323-336
4. Keef, C., Tawn, J.A and Svensson, C. (2009). Spatial risk assessment for extreme river flows. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **58(5)**, 601-618
5. Keef, C., Papastathopoulos, I., and Tawn, J. A. (2013). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis*, **115**, 396-404
6. Keller, V.D.J., Tanguy, M., Prosdocimi, I., Terry, J.A., Hitt, O., Cole, S.J., Fry, M., Morris, D.G. and Dixon, H. (2015). CEH-GEAR: 1 km resolution daily and monthly areal rainfall estimates for the UK for hydrological use. *Earth System Science Data Discussions*, ***8*(1)**
7. Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press
8. Towe, R.P., Tawn, J., Lamb, R., Sherlock, C., and Liu, Y., (2016). Efficient conditional extreme value modelling to handle missing values and better marginal distributions. *Unpublished internal document, JBA Trust*
9. Bortot, P., Coles, S., and Tawn, J. (2000). The multivariate gaussian tail model: An application to oceanographic data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, ***49*(1)**, 31-049