MODELLING INTER-SITE DEPENDENCE
IN REGIONAL ESTIMATES OF HYDROLOGICAL EXTREMES

S.J. Neil
University of Surrey
4.3.1990
CONTENTS

1. Introduction

2. Regional Estimation Techniques
   2.1 Why Regionalise
   2.2 Estimation Techniques
   2.3 "Flood Like" Distributions
   2.4 Homogeneous Regions
   2.5 Regional Heterogeneity

3. Inter Site Dependence
   3.1 The Importance of Inter Site Dependence
   3.2 Summary of Previous Work

4. Max Stable Processes
   4.1 Definition
   4.2 Representation
   4.3 Motivation
   4.4 Min Stable Processes
   4.5 Construction of Min-Stable Processes
   4.6 The Gaussian Extreme Value Distribution
   4.7 Estimation

5. Data Analysis
   5.1 Objectives
   5.2 Simulation Methodology
   5.3 Simulation Results
   5.4 Analysis of Real Data

6. Conclusions
ABSTRACT

In order to help the design of flood protection schemes and so on it is often required that predictions should be made about the extreme levels of river flows or rainfall over a very long time. For example, it may be required to predict the rainfall level that may be exceeded once in perhaps 500 years. To make these predictions stochastic hydrologists tend to use the annual maxima series for river flow or rainfall level. However, these records tend to be of short duration and they may not be taken particularly close to the site of interest. To overcome this lack of data it is now usual to use some form of regional estimation technique. These techniques generally assume that data from a hydrologically homogeneous region may in some way be pooled and used to produce regional estimates of the river flow and rainfall quantiles. The quantiles at any particular site may then be found by scaling the regional quantiles using at-site data. The main points of the most commonly used regional estimation techniques are briefly discussed. One of the problems with these techniques is that they assume that within a region all the sites are independently distributed. After a review of previous work examining this problem of inter-site dependence a new method for modelling and examining inter-site dependence using the properties of min-stable processes is introduced. The problem of how to generate min-stable distributions is considered and the techniques required to fit these distributions are investigated using both simulated data and annual maxima rainfall data. Although the estimators are found to be rather inaccurate the model itself is shown to have great potential for use as a simulation tool for studying the effects of inter-site dependence.
1. INTRODUCTION

In many engineering applications it is important that we have some information about the behaviour of hydrological variables at extremely large or extremely small values. Typically, if we need information about the chance of floods then we need to examine extreme values for river flows and rainfall. The engineers would like to know what river flow or rainfall level is likely to be exceeded once in 100 years or perhaps even once in 500 or 1000 years. The objective of statistically modelling hydrological extremes is to provide the engineers with this information.

The techniques used to model hydrological extremes are the subject of much on-going research. Section 2 of this thesis will give a very brief outline of the major models and techniques that have been developed so far. These techniques are becoming increasingly sophisticated but, of course, they can never approach the subtleties of the real hydrological processes. Therefore, the assumptions that are required to set up any model need to be examined closely. Section 2 of this thesis will also contain a brief summary of the work that has been done to validate the assumptions used. In Section 3 the assumptions of inter-site independence is examined in rather more depth. The previous work done in this area is examined and it is suggested that more work is necessary to model the inter-site dependence structures. To further investigate these dependence structures in Section 4 we introduce max stable and min stable processes and show how these processes can be used to model dependence structures. The results of data analysis performed using these processes are given and discussed in Section 5. The modelling was performed using both real rainfall data and simulation studies. The final Section discusses how this work helps our understanding of inter-site dependence and what effect this may have on the modelling of hydrological extremes.
2. REGIONAL ESTIMATION TECHNIQUES

2.1 Why Regionalise?

To model the behaviour of extremes accurately naturally requires a large amount of data. Extensive data sets are required in order that we can be confident that the large observations are drawn from the upper tail of the underlying distribution. Unfortunately, in practice the data record at most sites is quite short, with reliable data for perhaps 30 years or less. Therefore, a way of increasing the amount of data available was looked for and regional estimation was an obvious method of achieving this goal. All these techniques are based on selecting a region, not necessarily a geographical region, within which the annual maxima series from each of the sites can be regarded as independent and, except for a scaling factor at each site, identically distributed. To fit this regional distribution the annual maxima series from the individual sites are pooled - either directly or by taking a weighted sum of certain statistics calculated at each site. The pooled data are then used to fit the regional distribution. When the regional distribution has been fitted the quantiles for a particular site may be estimated from the regional quantiles and the estimated scaling factor for that site.

The way in which the at-site quantiles are found points to the second advantage of the regional estimation techniques. In real life we will seldom need to know the extremal behaviour of, say, the rainfall at a place where there is a rain gauge. Typically we will be interested in ungauged sites. Using a regional model the problem of estimating the quantiles at an ungauged site requires only that we may allocate the site to an already defined region and that we may in some way estimate the scaling factor for an ungauged site. It is not proposed to study these methods for estimating the at-site scaling factors for ungauged sites. Instead we move on to discuss the main regional estimation techniques and the distributions that are currently in use.
2.2 Estimation Techniques

There is a wide range of regional estimation techniques in use around the world. Since each method has its advantages and disadvantages and the results of numerous comparative studies have been inconclusive the diversity of methods used is unlikely to diminish. Therefore, in this Section we will only look at the two techniques most used in the U.K. Further techniques - used, for example, in America - are discussed in the summary paper by Cunnane [1986].

The simplest regional estimation technique is the so-called "station-year method". This method involves taking the data from each site, standardised to have mean 1, and simply treating all the standardised data as being one sample from the regional distribution. The regional distribution can then be fitted using a standard estimation technique. The at-site quantiles are found by multiplying the estimated regional quantiles by the estimated site mean. The procedure in use in the U.K. works in this way. The Flood Studies Report, N.E.R.C. [1975], defines regions for rainfall and for river flows and calculates a "growth curve" for each region. The growth curve gives the regional quantiles from which the quantiles at any site may be estimated. Since the growth curves are fixed, within each region, in practice we only need to estimate the mean of the sites annual maxima series in some way. This makes applying the method comparatively easy. The disadvantages of this simple approach are reasonably clear.

A rather more sophisticated approach now in use is the method of regionalised probability weighted moments. Probability weighted moments (PWM) were introduced by Greenwood et al. [1979] and their theoretical properties are considered in Hosking [1986]. We may define the PWM of a random variable, with distribution function $F$, as:
\[ M_{prs} = E[x^p F'(x)(1 - F(x))^r] \]

for \( p, r, s \) real numbers. Notice that \( M_{prs} \) exists if and only if \( E |X|^p \) exists.

We may usually restrict attention to the simpler quantities:

\[ \alpha_r = M_{10r} = E[x(1 - F(x))^r] \]  
\[ \beta_r = M_{110} = E[x F'(x)] \]

It can be shown - see Hosking [1986] - that \( \alpha_r \) and \( \beta_r \) are linear functions of the expected values of extreme order statistics. Hence:

\[ r \alpha_{r-1} = E[x_{1:r}] \]
\[ r \beta_{r-1} = E[x_{r:r}] \]

where \( x_{k:r} \) is the \( k^{th} \) largest observation from a sample of size \( r \). It may also be shown that the set of functions \( \{ \alpha_r, r=1,2,3,... \} \) \( \{ \beta_r, r=1,2,3,... \} \) uniquely define a real valued random variable. In some cases it is helpful to examine certain linear combinations of the PWM. These combinations, which Hosking called “L-moments”, are defined as follows:

\[ \lambda_{r+1} = \left\{ (-1)^r \sum_{k=0}^{r} p_{r,k} \alpha_k \right\} \]

where \( p_{r,k} = (-1)^r - k \binom{r}{k} \binom{r+k}{k} \). \( r = 0, 1, 2,... \)
It is noted that $\lambda_2$ provides a measure of scale and this leads to the definition of the dimensionless ratios:

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \quad r = 3, 4, 5, \ldots$$  \hspace{1cm} (2.7)

These functions are called "L-moment ratios". The first two L-moment ratios, $\tau_3$ and $\tau_4$, can loosely be seen as measures of skewness and kurtosis respectively. Therefore, PWM allow us to summarise the properties of a random variable using $\lambda_1$ (location), $\lambda_2$ (scale), $\tau_3$ (skewness) and $\tau_4$ (kurtosis).

The application of PWM that is of interest in stochastic hydrology is as a parameter estimation technique. This is done in a similar way to the method of (conventional) moments: the PWM are expressed as functions of the unknown parameters and a system of equations is solved for the parameters but replacing the true PWM by their sample estimates. The sample PWM may be estimated either by a graphical method, using plotting positions, or a numerical method using the ranked sample maxima. Hosking [1986] discusses the merits of both methods. Hosking [1986] also establishes the asymptotic normality of the estimates and shows that their asymptotic efficiency is reasonably high. To investigate the small sample properties a number of simulation studies have been performed - see, for example, Hosking et al [1985] - and the results have been encouraging.

Having described the general use of PWM we now outline the regional PWM estimation technique. Suppose we have a region containing N sites and that the $j^{th}$ site has $n_j$ years of record. Suppose also that the regional distribution we wish to fit has k parameters and that its first $k+1$ PWM exist and can be found in terms of these unknown parameters. We then proceed as follows:

i) At each site estimate the first $k+1$ PWM, say, $\hat{\beta}_{ij}, \quad r = 0, 1, 2, \ldots, k, \quad j = 1, 2, \ldots, N$. Equivalently we could use
\[ \hat{\alpha}_{rj} \text{ - see 2.2 and 2.3.} \]

ii) Scale each of the site PWM by dividing by the site mean, \( \hat{\beta}_{oj}, \) say \( t_j = \hat{t}_{rj} / \hat{\beta}_{oj} \) \( r = 1, 2, 3, \ldots k \) and \( j = 1, 2, 3, \ldots N. \)

iii) The regional PWM are found as weighted sums of these site, scaled, PWM. Say:
\[
T_r = \left[ \sum_{j=1}^{N} n_j t_{rj} \right] \left[ \sum_{j=1}^{N} n_j \right]^{-1} \quad r = 1, 2, 3, \ldots j
\]
\[
T_0 = 1.
\]

iv) The regional distribution is then fitted \( \omega \ln \sigma_k T_r = 0, 1, 2, \ldots, k. \)

v) From the fitted regional distribution any regional quantile can be found, say \( \hat{q}. \) The equivalent at - site quantile can be estimated as
\[
\hat{q}_{j} = \hat{\beta}_{oj} \hat{q}.
\]

This technique is quite general and can be applied to any distribution for which the first \( k+1 \) PWM can be found analytically. Notice that, unlike the station-year approach the data are only pooled via a set of sample statistics. This renders the method less susceptible to the influence of outliers. Techniques of this kind are sometimes referred to as “index flood” methods.

2.3 “Flood Like” distributions

Having described the most popular regional estimation techniques
now in use in the UK we now examine some of the models that are used for the regional distributions. Since we are interested in the annual maximum rainfall or river flow a commonly used family of distributions is the extreme value family. It can be shown, that if the distribution of the maximum of a sample of size \( n \), after the maximum has been normalized by a function of \( n \), tends to a non-degenerate distribution as \( n \) tends to infinity then this distribution must be a member of the extreme value family. A good description of extreme value theory may be found in Galambos [1987]. There are three separate types of distributions in the extreme value family. However, Jenkinson [1955] showed that they may be combined into the following distribution function:

\[
F(x) = \begin{cases} 
\exp \left[ 1 - k(x - u) / \alpha \right] & k \neq 0 \\
\exp \left\{ \exp \left[ (x - u) / \alpha \right] \right\} & k = 0 
\end{cases} \quad (2.8)
\]

This distribution is called the Generalised Extreme Value distribution (GEV). The GEV forms a plausible model for the behaviour of hydrological extremes and, in fact, it is the distribution currently in use - see the Flood Studies Report, NERC [1975].

The Wakeby distribution - derived in Houghton [1978] - is a distribution that is becoming increasingly popular. This distribution has no explicit formula for its distribution function but is defined in the inverse form:

\[
x(F) = \zeta + \alpha \left[ 1 - (1 - F)^\beta \right] / \beta + \gamma \left[ 1 - (1 - F)^{-\delta} \right] / \delta \quad (2.9)
\]

This form of parameterisation is used since it simplifies the equations used to estimate the parameters using PWM - see Hosking [1986]. In fact the PWM estimating technique was developed for just this problem. The Wakeby distribution has succeeded in modelling a wide range of “flood like” data - both in real applications and in simulation studies. It has noticeably more flexibility than the GEV but, with two extra parameters...
in the Wakeby distribution, this is not altogether surprising. This increased flexibility means that the Wakeby distribution is extremely useful for producing data to be used in simulation studies.

Rossi et al [1984] introduced the "Two Component Extreme Value distribution" (TCEV). This distribution has the following distribution function:

$$ F(x) = \exp - \left\{ \Lambda_1 \exp \frac{-x}{\theta_1} + \Lambda_2 \exp \frac{-x}{\theta_2} \right\} . \quad (2.10) $$

One way of deriving this distribution is to suppose that the annual maximum rainfall or flow is drawn from one of two Gumbel distributions. A Gumbel distribution is a GEV with $k=0$, see (2.8). The first distribution tends to give lower results and the second gives the larger values. These distributions are known as the basic component and outlying component respectively. The TCEV is derived as a mixture of these two components. Practically we may reason that this situation may apply, taking rainfall as an example, if the annual maximum rainfall might be a moderately small value if it were caused by frontal rain or a noticeably larger value if caused by a convective storm. The TCEV can be fitted using the standard regional estimation techniques. However, the PWM estimation procedure can fail to give a valid result in a noticeable minority of cases. There is however, a regional estimation technique particular to the TCEV distribution. The details of this method may be found in Fiorentino et al [1987] and the references that paper contains. Since the method applies to only one distribution it will not be described here. The TCEV has been used extensively and has been found to be reasonably flexible. However, in a number of studies problems have arisen through the failure of the estimation technique used to converge to a valid distribution.

These three distributions - GEV, Wakeby and TCEV - are the most commonly used distributions in this country. Other distributions are in
use, perhaps the most important of which is the log-Pearson type 3 distribution since this is the distribution in official use for flood prediction in the U.S.A. - see U.S.W.R.C. [1967]. The question of how to choose between the various models and estimation techniques is still an open one. There are very few theoretical results to guide this choice. However, a lot of work has been done to derive results from simulation studies. Cunnane [1986], in a general review of regional estimation techniques, gives a table showing what simulation studies have been performed, which distribution and estimation techniques they use and if they contain any misspecification in the model. The results from the various simulation studies are not clear but in general the results may be summarised as follows. The regional PWM technique is now generally favoured as the estimating procedure. However, this technique can fail when a TCEV distribution is fitted and so for this distribution other fitting methods are better. For all the studies it was found that the estimates of the quantiles were very variable and while they were nearly unbiased for moderate quantiles, as the quantiles increased so did the bias. The differences between the methods and distributions were mainly in quite small changes in the sizes of the variability and bias. The usual technique in the simulation studies is to take a selection of distributions and use these to fit a set of true regional distributions. In this way the performance of a distribution is assessed both when it is the true regional distribution and when it is misspecified. Not surprisingly all the distributions perform best when they are the true regional distribution. Due to their greater flexibility the Wakeby and TCEV distributions are generally superior to other distributions under misspecification. The nature and level of this superiority does depend on the true regional distribution, as does the comparison between the Wakeby and TCEV distributions. Given that these two distributions tend to be more accurate than, say, the GEV distribution in a wide range of cases a question that has remained unanswered is whether the increase in accuracy that we obtain is significant given the extra parameters used and hence the loss of degrees of freedom.
We observe, then, that there is no combination of fitting technique and distribution that is generally superior to the others. This superiority may be judged to the accuracy of the method - its bias and variability - or its being calculationally simple or having a particularly elegant derivation. Undoubtedly in the future further methods will be proposed and the choice of method will remain, at least in part, a matter of convenience or prejudice.

2.4 Homogeneous Regions

To use these regional estimation techniques we need to be able to define regions within which the annual maxima series are independently and, except for an at-site scale factor, identically distributed. In this section we examine some of the ways that have been proposed for defining these "homogeneous regions".

Perhaps the most obvious approach is to divide the country into geographical regions. This is indeed what was done in the Flood Studies Report, N.E.R.C. [1975], where the division was made partly to produce regions within which the coefficients of variation of the annual maxima series were similar and partly on pragmatic grounds. This approach is probably adequate for problems of rainfall extremes - where a spatial element in the process is strongly indicated. However, the situation is more complicated if we wish to study river flow. In this case the fact that two river basins are close together may not mean that their annual maxima series are similar if the catchment characteristics of the two basins are not similar. Hence in order to define homogeneous regions for river flows the current trend is to move away from geographical regions and instead to define the regions either on the basis of the statistics of the flood frequency distribution - sample mean, coefficient of variation, skew and so on - or on the basis of catchment characteristics - soil types, basin
area, average slope, urban coverage mean annual rainfall and so on. Both methods will use clustering algorithms to assign the catchments to regions. Recent applications of these methods may be found, for example, in Mosley [1981] - this paper also contains an account of the principles that should be used in any regionalisation scheme - or in Wiltshire [1986a] or in White [1975]. The regions defined by these methods will depend critically on which factors are used in the clustering algorithm. If significant factors have been overlooked then the regions may not be particularly useful. This problem will clearly increase in importance as fewer factors are used for the partitioning. However, assuming that the key factors may be identified, the regions produced by these partitioning methods should be an improvement over those defined by the more simplistic geographical methods.

2.5 Regional Heterogeneity

If we have a region defined by one of the above methods there are two questions that we need to know the answers to. Firstly how can we measure the level of heterogeneity in a region - and so devise a test for a homogeneous region - and secondly what effect will a given amount of heterogeneity have on any regional estimation technique that we might use.

Measures of regional heterogeneity are quite rare for the simple reason that it is a very difficult thing to measure. A particular measure may detect departures from all the sites having identical distributions quite well but be confounded by the dependence between the sites, for example. Dalrymple [1960] proposed a test for homogeneity based on the 90 percentile - 10 year return period - flood at each site in the region. More recently Wiltshire [1986b] proposed two tests, one a distribution free
test based on the observed site coefficients of variation, the other assumed a regional GEV distribution. This test used the fact that if for each member of a random sample \( x_j \), we calculated a non-exceedance probability, \( G_j \), then \( \{G_j, j=1,2,...,n\} \) should be a random sample from a uniform distribution on \([0,1]\). The test assumes we have correctly specified the parent distribution from which the random sample was drawn. Wiltshire [1986c] used simulations to study the properties of these two tests. Both tests tend to be conservative so that a region is found to be homogeneous unless there is very strong evidence that it is not. The powers of both tests depend on the numbers of sites in the region and the length of record at each site, the power increasing with these numbers. In general it was found to be better to have a few sites with long data records rather than many sites with short records. For the length of record typical in real applications neither test had high power. The tests were used by Wiltshire [1986c] to examine the geographical regions for river flows defined in the Flood Studies Report N.E.R.C. [1975]. These regions were all found to be heterogeneous. Similarly Dales and Reed [1986] examined the regions for rainfall from the same source. In this case most of the regions were not significantly heterogeneous and only two - Wales and South West England - were found to be heterogeneous. Notice, of course, that any heterogeneity detected by the distributionally dependent test may merely indicate that the regional distribution has been misspecified. However, since both tests tend to overstate the level of homogeneity, if the tests show significant heterogeneity then we may be sure that in some way the region is misspecified.

Awareness of the problems of regional heterogeneity has been growing recently and this has led to quite a lot of work being done on the question of how robust the regional estimation techniques are. In many of the simulation studies comparing the various estimation techniques some form of regional heterogeneity has been included. Cunnane [1986, table 2] gives a list of those simulation studies incorporating
heterogeneity and comments on the results. In most of these studies the heterogeneity is introduced as a departure from the sites being identically distributed - the sites are still independent. A good example of these methods - and the results obtained - is Hosking et al. [1985]. The results from these studies depend very much on the way in which the heterogeneity is modelled. It is therefore important that the heterogeneity is modelled in a plausible manner. The results from the simulation studies show that in general if the site distributions are all drawn from the same family - but not necessarily the same distribution - and that the sample properties - mean, coefficient of variation, skew and so on - are evenly spread around some average set of values then the effects of regional heterogeneity are not great. The main effects will be to introduce bias into the site quantile estimates and to increase the variability of these estimates. However, as Hosking et al. [1985] note, it appears that any loss of accuracy in the regional estimates is insufficient to make it preferable to rely simply on the single sites estimates alone. This reassuring conclusion that, even in the presence of heterogeneity, regional estimators are more accurate than single-site estimates depends on the modelled heterogeneity structure. Hence if the site distributions are less well behaved than described above the relative accuracy of the regional and single site estimators may change. Another criticism of many of the simulations is that they still tacitly assume inter-site independence. It is the problem of inter-site dependence that we examine in the next section.
3. INTER-SITE DEPENDENCE

3.1 The importance of Inter-Site Dependence

In order to set up the regional estimation techniques we needed to assume that the annual maxima series at each site were independent of each other. For any real hydrological process this will not be exactly the case. Indeed, especially for rainfall data, our knowledge of the processes involved would suggest quite strong inter-site dependence. Hence it is necessary to investigate the magnitude and effect of inter-site dependence under a variety of realistic dependence structure models.

The way in which inter-site dependence may influence the flood frequency estimates can be illustrated as follows - see Reed and Dales [1989]. Suppose there is a region containing N sites and suppose that in some way we may defined the “equivalent number of independent sites”, $N_e$, with $N \geq N_e$. For any site the probability of observing a value greater than the $T$ year return period value is $P_T = \frac{1}{T}$. Hence for the whole region the probability of at least one such exceedance is:

$$P_T = 1 - \left(1 - \frac{1}{T}\right)^{N_e}.$$  

(3.1)

If $N/T$ is small then $P_T \approx N_e/T$, but if all the sites were independent then the probability of at least one exceedance would be $P_T \equiv N/T$. Therefore the risk of there being an exceedance of the T year return period value is reduced by a factor of $N_e/N$. However, by a similar argument we find that if any $N_e$ sites exceed the T year return value then the number of sites with an exceedance is increased by a factor of $N/N_e$. The time between extremes is therefore predicted to be increased but when exceedances occur they will tend to occur in clusters. This may have important practical applications if it means that instead of isolated floods
(for example) the engineers need to design for far more widespread extremes.

The main problem with studying inter-site dependence is to define the dependence structure. The choice of a model for the dependence structure is almost limitless - beginning with simple linear correlation and increasing in complexity as we introduce spatial dependence or trends and cycles in the data and so on. Clearly results that are obtained will depend critically on the dependence structure used. Therefore a matter of prime concern is to produce a dependence model that can reproduce a wide variety of dependence structures of a type that are observed in real hydrological data.

### 3.2 Summary of Previous Work

An obvious starting point for studying inter-site dependence is transform the data at each site so that each site’s data has a standard normal distribution. We then assume that over the whole region the set of N sites forms an N dimensional multivariate normal distribution with a mean vector of zeros and covariance matrix $\Sigma$. The problem of defining the dependence structure is therefore reduced to defining the covariance matrix $\Sigma$. This method certainly makes studying inter-site dependence a tractable problem. Stedinger [1983] investigated the effects of inter-site correlations and derived some theoretical results. These results depend on the assumption that linear correlation is an adequate model for the true dependence structure. The main result derived by Stedinger [1983] was that under inter-site correlation the estimates of the regional probability weighted moments remained unbiased. The correlation merely increased the variability of these estimates and, in some cases, could increase the variability dramatically. Hosking and Wallis [1988] have reported the results of some simulation studies. The authors generated multivariate normal data and transformed it so that each marginal distribution had a GEV distribution. These transformed
marginal distributions were then used as the at-site data in order to fit a regional GEV distribution by the method of probability weighted moments. The results of these studies were generally encouraging. The biases of the quantile estimates were virtually unaffected and the increase in the variability of the estimates was not great compared to the level of variability measured in the independent case.

While using linear correlations as the inter-site dependence structure is the obvious, simple, approach it may not be appropriate for real hydrological data. We know that linear correlation is best suited to measuring dependence in distributions that are approximately normal. For real hydrological data this is seldom the case - the data tends to be heavily skewed and also relatively long tailed. Hosking and Wallis [1988] recognised this problem and commented that they had found that when they used a dependence structure that increased the level of dependence with the quantile levels the accuracy of the regional estimates could decline much more rapidly than they had previously observed.

An alternative approach that has been tried towards modelling the dependence structure was developed by Buishand [1984]. Consider two sites each with distribution functions $F$ and with the joint distribution function $G$. If the two sites are independent then $G(x,y) = F(x).F(y)$. However, if the sites are dependent then define the dependence function $p(x,y)$ such that

$$G(x, y) = [F(x).F(y)]p(x,y)$$

(3.2)

Now consider the maximum value from the pair of sites. The probability that a pair of results is less than any value is $G(x,x)$. Now if we define $q(x) = 2p(x,x)$ then we see that from a sample of $N$ pairs of observations the distribution function for the maximum, $M_N$, is

$$P(M_N \leq x) = [G(x,x)]^N.$$  

(3.3)
Notice that we assume that the N pairs are independent but the values within each pair are dependent. Now from 3.2 we see that:

$$P(M_N \leq x) = [F(x)]^{Nq(x)}$$

(3.4)

Buishand [1984] proposed the use of q(x) as a dependence function. Notice that for independent sites q(x) = 2 and for completely dependent sites q(x) = 1 (for all values of x). Typically we do not know the form of either of the distribution functions and so Buishand proposed a non-parametric estimate for the function q(x) using the empirical distribution functions. This procedure is quite appealing because of its simplicity. However, the estimate of the dependence structure is quite crude and it is limited to bivariate data. An ad hoc method for finding q(x) over a region is proposed but its validity is questionable and in any case its sampling properties are shown to be quite poor. This approach was the first time that a measure of inter-site dependence was used beyond linear correlations. Notice how Buishand attempts to use the fact that we are interested in annual maxima series in defining the dependence function. An interesting idea that arises naturally from this scheme is the “equivalent number of independent sites”. Hence if q(x) is estimated a region containing k sites it can take any value between 1 - a completely dependent region - and k - full independence. The concept is certainly useful in explaining inter-site dependence to non-statisticians.

A similar measure of dependence may be obtained using extreme value theory. Again consider the case of only having two sites and suppose that their joint distribution function is one of the bivariate extreme value family. Then without loss of generality it can be shown that, if the margins have Gumbel distributions, the joint distribution function has the following form:

$$G(x, y) = \exp - \left[ (e^{-x} + e^{-y})A(w) \right]$$

(3.5)
where \( w = \gamma/\lambda + y \) and \( A(w) \) is called the dependence function. See, for example Tawn [1983]. Notice that in this case \( q(x) = 2A(\frac{1}{2}) \). Multivariate extreme value distributions have attracted some recent work - see Tawn [1983] - but their usefulness is limited by the lack of progress made for dimensions higher than the bivariate case.

Another model for inter-site dependence, proposed by Reed and Dales [1989], has been developed directly from considering the regional estimation techniques. The model was developed for use with rainfall data but in principle could be applied to river flow data. However, as the authors state, inter-site dependence is, by the nature of the processes involved, more likely to be a serious problem in rainfall data. Suppose there is a region containing \( N \) sites and that we fit a regional GEV distribution. This defines the regional growth curve. By repeated resampling of a subset of the sites we fit a “typical” GEV distribution and hence find the “typical” growth curve. The objective is to estimate the equivalent number of independent sites, \( N_e \), by comparing the regional and the typical growth curves. The authors propose making this comparison using the estimated first probability weighted moments - \( \beta_{o}^{(r)}, \beta_{o}^{(t)} \) - of the regional and typical distributions. If the curvature parameter for the typical GEV distribution is \( h \) and the distribution has lower bound \( X_b \), then, as Dales and Reed [1988] show, \( N_e \) may be estimated as:

\[
N_e = \left[ (\beta_{o}^{(r)} - X_b)(\beta_{o}^{(t)} - X_b)^{-1} \right]^{\frac{1}{h}}.
\]  

(3.6)

The authors also show that if the sites are independent then the expected distance between the regional growth curve and the typical growth curve is \( \ln N \) (when the graphs are plotted on the Gumbel scale - where \( y = -\ln -\ln F(x) \)). If the sites are dependent then this distance is \( \ln N_e \). Hence the ratio \( \ln N / \ln N_e \) can be seen as a measure of the level
of dependence within the region. This relationship is illustrated in Figure 1. The equivalent number of independent sites may be estimated by either of these methods. Reed and Dales estimated $N_e$ for the rainfall regions defined in the Flood Studies Report, N.E.R.C. [1975]. The authors also attempt to generalise the method by fitting a model of the type

$$\ln N_e / \ln N = a + b \ln (\text{AREA}) + c \ln N + d \ln D, \quad (3.7)$$

where AREA is an estimate of the area spanned by the subset of sites used in the re-sampling, N is the number of sites in the region and D is the period over which the rainfall results are taken, in days - i.e. daily, two-daily, weekly data. The constants $a, b, c, d$ are to be fitted. To fit this model the authors repeatedly estimated $N_e$ for the various rainfall regions and estimated $a, b, c, d$ by weighted regression - the weights were designed to give more importance to data sets with long records. The authors comment that the estimated parameters do not vary greatly between the regions. This is quite surprising given that the regional growth curves differ greatly. For example consider the two regions differing the most in regional growth curves - the west country (Somerset and Dorset) and the Lake District. The west country has the most rapidly upward sloping growth curve while the Lake District's growth curve is nearly a straight line. However, the index of spatial dependence

$$\ln N_e / \ln N$$

found from 3.7 for these two regions were:

West Country: \[ \ln N_e / \ln N = 0.0 + 0.101 \ln (\text{AREA}) - 0.085 \ln N + 0.0 \ln D \]

Lake District: \[ \ln N_e / \ln N = 0.0 + 0.109 \ln (\text{AREA}) - 0.076 \ln N - 0.021 \ln D \]

The consistency of this measure of spatial dependence - for set AREA, N and D - may be a genuine result, showing the spatial factors are quite constant nationwide. Alternatively it may be indicating that the spatial effects are being dominated by the general variability of the growth curves with the regions so that their overall effect is only minor and cannot be reliably measured.
This work by Reed and Dales gives a reasonably simple and usable measure of the spatial dependence in a region. The sampling properties of this measure have not been studied yet and it is unlikely that elegant theoretical results will exist. Therefore simulation studies will be required to assess the sampling properties of this measure. This observation brings us back to the problem of what kind of multivariate model should be used for the regional dependence structure. Let us consider what properties are required for a suitable inter-site dependence structure. The first important property that any model should have is flexibility. We need to be able to model a very wide range of dependence structures if the results from simulation studies are to be reliable. This is especially true because we are mostly ignorant about the true inter-site dependence structures. However, any model should produce plausible dependence structures - either by being generated by a process that is hydrologically "reasonable" or by showing, if possible, that such dependence structures can be found in real data. We also note that, because we are attempting to establish a tool for simulations, it is not required that a dependence structure model should also be capable of being used as a regional estimation technique in its own right, although this would of course be extremely useful. Similarly calculational elegance is not a prime requirement. When a dependence model meeting these criteria has been produced it should be possible to assess the effect of inter-site dependence on the regional estimation techniques with a greater degree of confidence that the results have not been distorted by the dependence structures used.
4. MAX STABLE PROCESSES

4.1 Definition

Max-stable processes were independently introduced by de Haan [1984] and Vatn [1987]. The processes were further studied by de Haan and Pickands [1986]. However, the authors used the alternative form of the process - Min Stability. The method of transforming from Min Stability to Max Stability and the relationship between the two types of process are considered below in Section 4.4. In this section we consider only Max Stable processes.

Consider a process \{Y_t, t \in J\} and take \(r\) independent copies of this process. Let these copies be \(\{Y^{(i)}_t, t \in J\} i = 1, 2, ..., r\). Define the process \(\{X_t, t \in J\}\) as:

\[
\{X_t, t \in J\} = \left\{ \max_{1 \leq i \leq r} \left[ \frac{Y^{(i)}_t - a^{(i)}_t}{b^{(i)}_t} \right], t \in J \right\}
\]

(4.1)

where \(a^{(r)}_t\) and \(b^{(r)}_t\) are normalising constants dependent on \(r\), the number of copies of the original process taken. If \(\{X_t, t \in J\}\) has the same distributional form as \(\{Y^{(i)}_t, t \in J\}\) then the process \(\{Y_t, t \in J\}\) is said to be Max-Stable.

4.2 Representation

The properties of processes of this kind were studied by de Haan and Pickands [1986]. One result of particular practical interest is that all max-stable processes have a particular representation. It can be shown - by transforming the result for min-stable processes found by de Haan and Pickands [1986] - that the marginal distributions of this process are required to be in the extreme value family. For simplicity we will assume that
each margin has a unit Frechet distribution. Hence:

\[ P(Y_t \leq y) = \exp \left( -\frac{1}{y} \right), \forall t \in J, \quad y > 0, \quad (4.2) \]

Define a measurable set \( S \) over which a positive measure \( v(ds) \) is defined. Also define an index set \( J \). Consider a positive function \( f(s,t) \) over \( S \) and \( J \) such that:

\[ \int_S f(s,t)v(ds) = 1, \quad \forall t \in J \quad (4.3) \]

If the process \( \{Y_t, t \in J\} \) is max-stable then the distribution function of this process can be written as:

\[ P(Y_t < y, \ t \in J) = \exp \left\{ -\int_S \left[ \max_{t \in J} \left( \frac{f(s,t)}{y_t} \right) v(ds) \right] \right\} \quad (4.4) \]

for suitable choices of \( S, J, v(ds) \) and \( f(s,t) \). This representation is not unique. de Haan and Pickands [1986] considers the sets of functions \( f(s,t) \) that lead to the same distribution function. These sets are referred to as "pistons".

It can be shown, de Haan and Pickands [1986], that if the index set, \( J \), is a set of \( p \) distinct points then the distribution function defined in (4.4) becomes a member of the \( p \)-dimensional extreme value family. Hence extreme value distributions are a subset of max-stable processes.

4.3 Motivation

We now examine the motivation for using max-stable processes for modelling hydrological extremes. In practice we already use the
univariate extreme value family of distributions - i.e. the GEV - in many cases. It is then consistent to use multivariate extreme value distributions when we wish to model the dependencies between the sites or other regional properties that require a model of more than one dimension. Max-stable processes provide a possible method for generating multi-variate extreme value distributions.

A max-stable process allows us, at least conceptually, to model a hydrological process over the whole of a region. The data from a particular set of p sites can be regarded as a sample from the p-dimensional marginal distribution of the process. The regional process then gives us a method for extrapolating the model to sites where there is no data. Similarly the overall process may allow us to address questions about the regional behaviour. For example, the total rainfall in a region or the maximum river flow at any site in the region may be found. Notice that the current regional methods cannot answer questions like these without assuming complete inter-site independence.

If we consider only rainfall maxima it is possible to give a physical interpretation to a max-stable process. Suppose that a region is hit by a random number of rainstorms over a year. Each storm has random magnitude, $x_i$, and random location within the region $s_i$. The proportion of rain falling at a site, $t$, from the $i$th storm is defined using a “spread function”, $f(s_i, t)$. We assume that all the rain from each storm falls inside the region. The actual rainfall at site $t$ from storm $i$ is therefore:

$$r_{it} = x_i f(s_i, t)$$  \hspace{1cm} (4.5)

Now suppose that the process generating the storm magnitudes and locations, $((X_i, S_i), i=1,2,...)$, is a suitable poisson process. It can be shown, de Haan and Pickands [1986], that if the index set, $J$, contains the locations of all the sites in the region then the process:
\{Y_i, t \in J\} = \left\{ \max_{\text{all storms}} \{ x_i f(s, t), t \in J \} \right\}, \quad (4.6)

is max-stable. This is, of course, an artificial derivation but it is interesting to note that Rodriguez-Iturbe et al. [1986] use a similar basic approach for modelling total daily rainfall in South America.

The key theoretical result derived by de Haan and Franses [1994] is that any max stable process may be generated in this way. However, in this context, it is the converse that is immediately important - that any process generated in this way is max stable.

dependence structures. Recall that taking the index set as a set of distinct points - such as the set of sites in a region - defines a multi-variate extreme value distribution. For any \(p\)-dimensional extreme value distribution we may write the distribution function as:

\[
P(Y_i \leq y_i, i = 1, 2, \ldots, p) = \exp - \left\{ \sum_{i=1}^{p} (y_i^{-1}) A(w_1, w_2, \ldots, w_{p-1}) \right\}, \quad (4.7)
\]

where  \( w_i = y_i^{-1} / \sum_{j=1}^{p} (y_j^{-1}) \).

The function \( A(w_1, \ldots, w_{p-1}) \) is known as the dependence function. The properties of this dependence function are discussed in Tawn (1988). Notice that this is not the only definition used for the dependence function, indeed several different authors have used their own function. For a summary of the various dependence function definition see Deheuvels [1984] or Weissman [1985]. The important features of this dependence function that we need to note are:

(i) The limits of the function are \(p^{-1} \leq A \leq 1\) .
(c) We may relate this dependence function directly to the equivalent number of independent sites, $N_0$, as:

$$N_e = \rho A(\frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p})$$
(ii) For independent variables $A = 1$.

(iii) For completely dependent variables $z$

$$A = \max \left[ \frac{w_1}{w_2}, ..., \frac{w_{p-1}}{w_p}, 1 - \sum_{i=1}^{p-1} w_i \right].$$

Figure 2 shows these limits for the bi-variate case. It is sometimes of use to consider, as a kind of summary of the level of dependence, the value of the $A$ function along the line $Y_1 = y$, $i=1,2,...,p$. Hence we examine $A(1/p, 1/p,...,1/p)$. This may be useful in many dimensions where a complete evaluation of \[4.4\] is impractical, but we need some measure of overall regional dependence, since calculating \[4.4\] for one particular value of $Y_1$ is more straightforward.

Notice that this definition of a dependence function arises naturally from the nature of the extreme value problem. This may be a justification for using dependence structures generated in this way rather than using correlations which are based on the normal distribution. We therefore wish to study max-stable processes in order to demonstrate their usefulness in generating dependence structures that can model the intersite dependence we find in hydrological regions.

4.4 Min Stable Processes

Just as max-stable processes preserve the distribution under maximisation a related type of process preserves distribution under minimisation. These are the min-stable processes. A process, $(Z_t, t \in J)$, is said to be min-stable if it has the same distribution as:

$$\left\{ \min_{1 \leq i \leq n} \left[ \frac{z_i^{(0)}}{b_0^{(0)}}, t \in J \right], \quad 0 \right\}$$
It is easy to show that if \( \{Z_t, t \in J\} \) is a min-stable process and if we set
\[ Y_t = Z_t^{-1} \]
for all \( t \in J \), then the process \( \{Y_t, t \in J\} \) is max-stable. Therefore, we may transform between the two types of processes. However, it is often easier to use min-stable processes since the marginal distributions of a min-stable process are required to be exponential distributions. The representation of a min-stable survivor function can be given in a form similar to the distribution function given in [4.4]. Hence for a min-stable process \( \{Z_t, t \in J\} \):

\[
P[Z_t \geq z_t, t \in J] = \exp - \int_S \left[ \max(z_t f(s, t))v(ds) \right] \]  \tag{4.9}

where, as in section 4.2, \( S \) is a measurable set, \( v(ds) \) is a positive measure over \( S \), \( J \) is an index set and \( f(s,t) \) is a positive, real function over \( S \) and \( J \). The function \( f \) should integrate to unity over the set \( S \) for any point in \( J \).

The results discussed in Sections 4.1 and 4.2 were originally derived for min-stable processes by de Haan and Pickands [1986]. Max-stable processes were introduced first in this work in order to present the physical interpretation of the process. This physical motivation does not apply to min-stable processes. The equivalent construction for a min-stable process is to let \( X \) be the interval \((0,1)\) and \( S \) be the region \([R^k] - \) for some integer \( k \). Suppose the points \((x_i, s_i)\) \( i=1,2,... \) form a Poisson process on \((0,1) \times [R^k] \). If we let \( J \) be an index set and define the “spread functions” \( f(s,t) \) over \([R^k] \) and \( J \) such that \( f \) integrates to unity over \([R^k] \) for all \( t \in J \). Then define \( z_t \) such that:

\[
z_t = \min \left[ \frac{x_i}{f(s_i, t)} \right] \quad \text{for all } t \in J, \quad \text{all } i \]  \tag{4.10}
The process \( \{Z_t, t \in J\} \) defined in this way is min-stable.

Notice that if \( J \) is a set of distinct points then a multi-variate extreme value distribution is again defined, in this case with exponential rather than Frechet marginal distributions. Hence the survivor function of such a process may be written as:

\[
P(Z_i > z_i, i = 1, 2, \ldots, p) = \exp \left[ -\sum_{i=1}^{p} z_i A(w_1, w_2, \ldots, w_{p-1}) \right]. \quad (4.11)
\]

where \( w_j = z_j / \sum_{i=1}^{p} z_i \). This again shows that by setting \( Z_t = Y_t^{-1} \) for all \( t \in J \) we may transform between max-stable and min-stable processes.

Since, in general, it is more convenient to work with exponential distributions rather than Frechet distributions we will now restrict attention to min-stable processes. For the task of modelling dependence structures the choice is not of great importance since the dependence function, \( A \), is unchanged by the transformations.

### 4.5 Construction of Min-Stable Processes

In this section the representation of a min-stable process, [4.9], is used in an attempt to generate examples of min-stable distributions. To construct a min-stable distribution we need to define the region \( S \), its measure \( v(ds) \), the index set \( J \) and the function \( f(s,t) \) - this function will generally be referred to as the "spread function". In practical terms we will usually wish to impose the further condition that any distribution generated should possess a continuous density function. Therefore, we need to discover when the min-stable representation, [4.9], generates a
distribution with a continuous density function.

Consider the simplest case in which we wish to generate a bi-variate distribution. Hence we set J = \{1,2\}. Define S as an interval on the real line, for convenience set S = (0,1). Let the measure be \( v(ds) = ds \). The spread function is defined as \( f(s,t) \), \( t = 1,2 \). Assume that \( f(s,1) \) and \( f(s,2) \) are both continuous functions. We need to evaluate the integral in [4.9]. Let this be:

\[
I = \int_0^1 \text{MAX}[z_1 f(s,1), z_2 f(s,2)] ds . \tag{4.12}
\]

To evaluate this integral we need to identify the intervals in which \( z_1 f(s,1) > z_2 f(s,2) \). Now consider the "cross over points" where:

\[
z_1 f(s_o,1) = z_2 f(s_o,2) , \tag{4.13}
\]

for \( 0 < s_o < 1 \). The crossover points are found by solving:

\[
\frac{f(s_o,1)}{f(s_o,2)} = z_2 / z_1 . \tag{4.14}
\]

For any pair of values \((z_1, z_2)\) this equation will define \( c \) cross-over points in the interval \((0,1)\) with \( c \geq 0 \). If \( c > 0 \), order these points so that \( 0 < s_{01} < s_{02} < \ldots < s_{0c} < 1 \) and set \( s_{00} = 0 \) and \( s_{0c+1} = 1 \). Without loss of generality assume that \( z_1 f(0,1) > z_2 f(0,2) \). The cross-over points define \( c+1 \) intervals that partition \((0,1)\) and so we may divide \( I \) into \( c+1 \) integrals, \( I_j \), where:

\[
I_j = \begin{cases} 
\int_{s_{oj}}^{s_{o(j+1)}} z_1 f(s,1) ds & \text{for } j \text{ even and for } j = 0 \\
\int_{s_{oj}}^{s_{o(j+1)}} z_2 f(s,2) ds & \text{for } j \text{ odd}
\end{cases} \tag{4.15}
\]
for \( I = 0, 1, 2, \ldots, c \) and we find \( I = \sum_{j=0}^{c} I_j \).

From [4.9] we see that the density function for this process, \( g(z_1, z_2) \), is:

\[
\frac{\partial^2 I}{\partial z_1 \partial z_2} - \frac{\partial^2 I}{\partial z_1 \partial z_2^2} \exp(-I).
\]

It is therefore sufficient for \( g(z_1, z_2) \) to be continuous that the integral, \( I \), should be twice continuously differentiable by both \( z_1 \) and \( z_2 \).

Now consider inserting equation [4.14] so that:

\[
s_0 = F(z_2 / z_1).
\]

By considering the terms in \( \frac{\partial^2 I}{\partial z_1 \partial z_2} \) we can see that if the spread function is continuous then this differential is continuous provided that the function \( F \) is continuously differentiable by both \( z_1 \) and \( z_2 \). The result is demonstrated using \( S = \mathbb{R} \) but it can be extended to \( S = \mathbb{R}^p \).

In practice it was found that the simplest way to ensure that any distribution generated using [4.9] possessed a continuous density was to define the spread function such that \( f(s, 1) \) was a monotone increasing continuous function with \( f(0, 1) = 0 \) and that \( f(s, 2) \) was a monotone decreasing continuous function with \( f(1, 2) = 0 \), where we have set \( S = (0, 1) \), \( v(ds) = ds \) and \( J = \{1, 2\} \). We shall now examine some examples of distributions generated by spread functions of this form.

Example 1. Linear spread function.

Set \( S = (0, 1) \) and take \( 1/2 \leq \gamma \leq 1 \). Define
\[
    f(s, 2) = \begin{cases} 
        2(y - s) / \gamma \quad & 0 < s < \gamma \\
        0 & \gamma < s < 1 
    \end{cases} \\
    f(s, 1) = \begin{cases} 
        0 & 0 < s < 1 - \gamma \\
        2(s - 1 + \gamma) / \gamma^2 & 1 - \gamma < s < 1 
    \end{cases} 
\]

Then the crossover point is \( s_0 = \frac{z_1 \gamma + z_2 (1 - \gamma)}{z_1 + z_2} \).

This leads to the survivor function

\[
    P(z_1 > x, z_2 > y) = \exp \left[ x + y - (2\gamma - 1)xy(x + y)^{-1} \right]. 
\]

We may also write the dependence function, \( A(w) \):

\[
    A(w) = 1 - \theta w(1 - w) \quad (4.20) 
\]

where \( w = \frac{x + y}{x + y} \) and \( \theta = (2\gamma - 1) \). Notice that this distribution is in fact Gumbel's mixed model - see Tawn [1988]. Therefore using this simple spread function we have generated one of the standard bi-variate extreme value models. If in equation (4.20) we set \( \theta = 0 \) then the variables are independent. This model cannot represent complete dependence. The maximum dependence we can generate from this model is found when we set \( \theta = 1 \) for which we find \( A \left( \frac{1}{2} \right) = \frac{3}{4} \). In Figure 2 the dependence function for the mixed model with parameter \( \theta = 1 \) is shown.

**Example 2**  Powered spread function.

Set \( s = (0,1) \) and take \( 0 < \alpha \leq 1 \). Define:
The crossover point is found to be \( s_0 = \frac{1}{z_1} \left( \frac{1}{2} \right) z_1^{1-\alpha} \left( \frac{1}{2} \right) z_2^{1-\alpha} \).

The survivor function is:

\[
P(z_1 > x, z_2 > y) = \exp \left( - \left( \frac{1}{x^\alpha} + \frac{1}{y^\alpha} \right)^{1-\alpha} \right)
\]

which gives the dependence function:

\[
A(w) = \left[ w^\alpha + (1-w)^\alpha \right]^{\alpha-1}
\]

Examining the survivor - or dependence - function we notice that we have generated another of the standard bi-variate extreme value distributions. This is the logistic model. The properties of logistic and mixed models are discussed in Tawn [1988]. Here we simply comment that if we set \( \alpha = 1 \) in equation (4.23) then the variables are independent. If we let \( \alpha \to 0 \) then the variables tend towards complete dependence. For intermediate values of \( \alpha \) the full range of dependence can be modelled.

Example 3. Positive powered spread function.

Set \( S = (0,1) \) and take \( \alpha > 0 \). Define:

\[
f(s, 1) = (\alpha + 1)s^\alpha
\]
This gives a crossover point at \( s_0 = \frac{1}{2} \left( \frac{1}{z_1^a} + \frac{1}{z_2^a} \right)^{-1} \).

The survivor function is found to be:

\[
P(z_1 > x, z_2 > y) = \exp \left[ -xy - xy \left( \frac{1}{x^a} + \frac{1}{y^a} \right)^{-a} \right],
\]

with the dependence function:

\[
A(w) = 1 - w(1 - w) \left[ w^a + (1 - w)^a \right]^{-a}.
\]

Notice that as \( \alpha \to \infty \) then \( A(w) \to 1 \) so the variables tend towards independence. The most dependent the variables may become is when

Example 4. Hyperbolic spread function.

Set \( S = (0,1) \) and take \( \alpha > 0 \). Define:

\[
f(s, 1) = \gamma \text{h}(\alpha s)
\]

\[
f(s, 2) = \gamma \text{h}(\alpha(1 - s))
\]

where \( \gamma = \alpha(\text{ch}\alpha - 1)^{-1} \). The crossover point for these functions is:
This leads eventually to the survivor function

\[ S_0 = \alpha^{-1} th^{-1} \left[ \frac{z_1 sh\alpha}{z_2 + z_1 ch\alpha} \right], \]

where \( \theta = ch\alpha \).

This distribution has the dependence function:

\[ A(w) = \theta(\theta - 1)^{-1} - (\theta - 1)^{-1} \left[ 1 + 2(\theta - 1)(1 - w)w \right]^{\frac{1}{2}}, \quad (4.29) \]

From this we see that as \( \theta \to \infty \) then \( A(w) \to 1 \) and so the variables trend towards independence. For \( \theta \to 1 \) we find that \( A(w) \to 1 - w(1-w) \) and so this model is another to have this function as the upper limit for the amount of dependence.

Therefore using the min-stable representation and some obvious spread functions we can both recover the two well-known bi-variate extreme value distributions and generate some new models. Neither of the new models generated here are in fact of much immediate interest except as a demonstration of manipulating the min-stable representation. The dependence functions offer no greater flexibility than the well-known mixed model and it is unlikely that we would have any real reason to prefer one of these models to the others. However, it is always of interest to see new bi-variate extreme value distributions.

For completeness we now give an example of a spread function that does not generate a distribution with a continuous density.
Example 5.

Set $S = (0,1)$ and take $\alpha > 0$. Define:

$$f(s, 1) = \theta e^{-\alpha s}$$

$$f(s, 2) = \theta e^{-\alpha(1-s)}$$

where $\theta = \alpha(e^\alpha - 1)^{-1}$. The crossover point is $s_0 = \alpha - ln(z_1/z_2)(2\alpha)^{-1}$. Notice that this point only falls in the interval $(0,1)$ if:

$$e^{-\alpha} \leq (z_1/z_2) \leq e^\alpha$$

(4.31)

Otherwise one of the variables dominates the other over the whole interval. Hence we need to evaluate the survivor function for three separate cases. The survivor function is found to be:

$$P(z_1 > x, z_2 > y) = \begin{cases} 
\exp - (x) & \text{for } x > ye^\alpha \\
\exp - \left[\frac{(x+y)e^\alpha - 2e^{\frac{\alpha}{2}}(xy)^{\frac{1}{2}}}{[e^\alpha - 1]^{-1}}\right] & \text{otherwise} \\
\exp - (y) & \text{for } y > xe^\alpha 
\end{cases}$$

(4.32)

We may also derive the dependence function:

$$A(w) = \begin{cases} 
1 - w & \text{for } w < (1 + e^{-\alpha})^{-1} \\
w & \text{for } w > (1 + e^{-\alpha})^{-1} \\
\frac{e^\alpha - e^{\frac{\alpha}{2}}(w(1-w))^{\frac{1}{2}}}{[e^\alpha - 1]^{-1}} & \text{otherwise} 
\end{cases}$$

(4.33)
Notice that as $\alpha \to 0$ the range defined in (4.31) vanishes and so the variables become completely dependent. As $\alpha \to \infty$ the regions outside the central range tend to vanish. In this case the dependence function $A(w) \to 1$ and so the variables tend towards independence.

The exercise of choosing and manipulating various spread functions shows that using the min-stable representation to generate distributions is possible. However, even with the relatively simple examples demonstrated above, it tends to be a rather laborious process. The next step in developing min-stable processes is to try and use the representation to generate distributions in more than two dimensions. In general this has not proved possible using various trial and error methods. However, it is possible to establish an extremely limited class of multi-variate models which we construct as follows. Define $S$ as the interval $(a,b)$ on the real line and $v(ds) = ds$. Let $J = \{1,2,\ldots,k\}$, i.e. we create a $k$ dimensional distribution. Partition the interval $(a,b)$ into $k-1$ non-overlapping intervals and define the spread function such that in the $j^{\text{th}}$ interval $f(s,i) = 0$ unless $i=j$ or $j+1$. This will define a model in which a variable $z_j$ depends on, at most, $z_{j-1}$ and $z_{j+1}$. The model is, of course, very restrictive in the type of dependence structures that it will allow. However, this “nearest neighbour” type of model may be of some use in hydrological applications - perhaps relating the flow as we move down a river. To illustrate this class of models consider the following simple three dimensional example.

Example 6. A “nearest neighbour” model.

Set $S = (-1, 1)$ and $J = \{1, 2, 3\}$. Define:
We can see that the only crossover points will occur when $z_1 f(s,1) = z_2 f(s,z)$ and $z_2 f(s,2) = z_3 f(s,3)$. Let these points be $s_1$ and $s_2$ then:

$$s_1 = -z_2 / (2z_1 + z_2) \quad \quad \quad s_2 = z_2 / (2z_3 + z_2) \quad \quad \quad (4.35)$$

Hence the integral $I$, becomes:

$$I = \int_{s_1}^{s_2} z_1 f(s,1) ds + \int_{s_1}^{s_2} z_2 f(s,2) ds + \int_{s_2}^{s_3} z_3 f(s,3) ds \quad \quad (4.36)$$

which produces the survivor function:

$$P(z_1 > x, \ z_2 > y, \ z_3 > z) = \exp \left[-\frac{x y}{2(2x + y)} - \frac{y z}{2(2y + z)}\right]. \quad (4.37)$$

From this survivor function we may extract the dependence function $A(w_1, w_2)$ - where $w_1 = x/(x+y+z)$ and $w_2 = y/(x+y+z)$. We may construct further models of this kind using different spread functions and including more variables. The restriction that is required is that only two parts of the spread function may be non-zero in any interval.
4.6 The Gaussian Extreme Value Distribution

The distributions defined in the above example, with the exception of
the logistic model (4.22), are incapable of modelling the full range of
dependence and so for representing inter-site dependence functions their
use will be limited. This fact led to the examination of more complicated
spread functions in an attempt to find a new model capable of
representing any dependence level from independence to complete
dependence. The first choice of spread function was the normal density
function. Smith (personal communication) showed how a bi-variate
min-stable distribution could be constructed using the normal density as
the spread function. This distribution was called the "Gaussian Extreme
Value" distribution.

To generate this distribution take $S = \mathbb{R}^P$, for some integer $p \geq 1$, and
set $v(ds) = ds$. Let $J$ be a pair of position vectors in $\mathbb{R}^P$ so that $J = \{t_1, t_2\}$. Define the spread function

$$f(s, t_j) = f_0(s - t_j), \quad j = 1, 2,$$

(4.38)

where $f_0$ is the density function for the $p$-dimensional normal
distribution with mean vector zero and positive definite covariance
matrix $\Sigma$:

$$f_0(x) = (2\pi)^{-\frac{P}{2}} |\Sigma|^{-\frac{1}{2}} \exp -\frac{1}{2} (x^{T} \Sigma^{-1} x)$$

(4.39)

To generate the survivor function we need to find the regions, in
$\mathbb{R}^P$, where:

$$\sum f_0(s - t_i) > \sum f_0(s - t_2)$$

(4.40)
We find that (4.40) is true if and only if:

$$\frac{1}{2}(s - t_2)^T \Sigma^{-1} (s - t_2) - \frac{1}{2}(s - t_1)^T \Sigma^{-1} (s - t_1) > \ln(z_2/z_1).$$ (4.41)

From this equation we may define a region, $V$, on $S$ where (4.40) is true. $V$ will be defined as a function of $z_2/z_1$. The survivor function can be found as

$$p(z_1 > x, z_2 > y) = \exp\left[ x \int_{t_1} f_0(s - t_1) ds + y \int_{s/y} f_0(s - t_1) ds \right].$$ (4.42)

This is calculated to be

$$p(z_1 > x, z_2 > y) = \exp\left[ x\Phi\left( \frac{a}{2} + \frac{1}{a} \ln\left( \frac{x}{y} \right) \right) + y\Phi\left( \frac{a}{2} - \frac{1}{a} \ln\left( \frac{x}{y} \right) \right) \right].$$ (4.43)

where $a = (t_1 - t_2)^T \Sigma^{-1} (t_1 - t_2)$ and $\Phi$ is the standard normal distribution function.

We may write down the dependence function:

$$A(w) = (1 - w)\Phi\left[ \frac{a}{2} + \frac{1}{a} \ln\left( \frac{1 - w}{w} \right) \right] + w\Phi\left[ \frac{a}{2} - \frac{1}{a} \ln\left( \frac{w}{1 - w} \right) \right].$$ (4.44)

Notice from (4.44) that as $a \to \infty$ then $A(w) \to 1$ and so the model tends to independence. Also if $a \to 0$ then if $w < 1/2$, $A(w) \to (1 - w)$ otherwise $A(w) \to w$. Hence in this case we tend towards complete dependence.

Notice how the definition of $a$, (4.43), leads the "natural" interpretation of these limits as the sites being a very large distance apart - independence - or extremely close together - complete dependence.

Independently of this derivation of the Gaussian Extreme Value
distribution Husler and Reiss (1989) developed the same form of distribution in a different way. It can be shown - Sibuya (1960) - that if we take samples from a multi-variate normal distribution with covariance matrix $\mathbf{\Sigma}$, then, if all elements of $\mathbf{\Sigma}$ are less than 1, as more samples are taken the maxima of the marginal distributions tend towards independence. Husler and Reiss proposed that instead of a fixed covariance matrix $\mathbf{\Sigma}$, a covariance matrix depending on sample size, $\mathbf{\Sigma}_n$, should be used and that we should assume that $\mathbf{\Sigma}_n \rightarrow \mathbf{\Sigma}$ as $n \rightarrow \infty$ where $\mathbf{\Sigma}$ is a positive definite matrix. The Gaussian Extreme Value distribution may also be derived in this way.

4.7 Estimation

The min-stable representation, (4.9), has therefore, been used to generate some bi-variate distributions. The problem is now to find a method for fitting these distributions. Clearly if the distribution possesses a density function then we may use the standard method of maximum likelihood estimation (MLE). Tawn (1988) uses MLE to fit the mixed and logistic models (see (4.19) and (4.22) respectively) and this work shows the feasibility of this approach. However, problems can arise if the sites are nearly independent as in this case the Fisher information tends to infinity. This means we cannot find the usual estimates of the variability of the parameters and so construct a test for independence. Tawn goes on to show that at least for the mixed and logistic models, the parameter estimates are sufficiently regularly behaved to allow us to construct alternative estimates of their variability. These problems near to independence appear to be common to most techniques that attempt to fit a dependent distribution. Below we will introduce a method of moments style estimator and show that this can give poor results near to independence. Note also that in their study of inter-site dependence for
rainfall data Reed and Dales (1989) also encountered some problems with the estimate of regional dependence exceeding its theoretical limit.

Hence we may use MLE to fit the min-stable distributions. If we have long data sets and the sites are at least moderately dependent then we may expect the usual asymptotic properties of MLE to apply - the consistency, efficiency and normality of the estimates. However, we seldom have long data sets and we usually wish to believe that the sites form a homogeneous region and so are independent. In this situation it is unlikely that any estimator will perform particularly well. Thus it may be appropriate to look at a crude estimator.

In personal communication Smith considered a simple estimator of the following form. Suppose we have a pair of sites and each site has, or can be transformed to have, a unit exponential distribution. If the joint distribution of these sites follows some min-stable process then the joint survivor function for the sites can be written as

\[ P(z_1 > x, z_2 > y) = \exp\left(-[(x + y)A(w)]\right) \]  \hspace{1cm} (4.45)

in the usual way. The dependence function, \( A \), is defined by the particular min-stable process that we assume fits the data. The data consist of pairs of observations, \( \{(x_i, y_i), i=1,2,\ldots,\} \). Take the minimum of each pair and then consider the distribution of these minima. Notice that:

\[ P(\text{Min}(z_1, z_2) > z) = P(z_1 > z, z_2 > z) \]  \hspace{1cm} (4.46)

But

\[ P(z_1 > z, z_2 > z) = \exp\left[-2xA\left(\frac{1}{2}\right)\right] \]  \hspace{1cm} (4.47)

Therefore, if we define:
\[ \theta = 2A \left( \frac{1}{2} \right), \] (4.48)

Then from (4.47) we see that \( \text{Min}(z_1, z_2) \) has an exponential distribution with mean \( \theta^{-1} \). Hence we may estimate \( \theta \) as follows:

\[ \hat{\theta} = \frac{N}{\sum_{i=1}^{N} \text{Min}(x_i, y_i)}, \] (4.49)

If the dependence function has only a single parameter to be estimated then we calculate \( \theta = 2A \left( \frac{1}{2} \right) \) as a function of the unknown parameter and equate this to the estimated value \( \hat{\theta} \). This equation is then solved for the unknown parameter. For example, consider the logistic model - (4.23) - we see that \( 2A \left( \frac{1}{2} \right) = 2^\alpha \). Hence we may estimate the parameter \( \alpha \), as \( \hat{\alpha} = \frac{\ln(\theta)}{\ln(2)} \).

Notice that if the observations are standardised so that the samples at each site have unit means then if the sites are completely dependent then \( \hat{\theta} \) will automatically take the correct value - i.e. \( \hat{\theta} = 1 \) (since for complete dependence \( A \left( \frac{1}{2} \right) = \frac{1}{2} \)). No such result exists if the sites are in fact independent. Indeed it is simple to show that \( \hat{\theta} \) is not bounded above. Consider two sites. At the first site the results are clustered about the mean, hence after standardization we find all the results are approximately equal to 1. The other site has one large outlier so that, after standardization, the outlier is of the order of \( N-1 \) and the other observations are approximately, \( (N-1)^{-1} \). If we take the minimum of each pair of observations then \( N-1 \) of these minima will come from
the second site, the first site only contributing one observation. In this case from equation (4.49) we find that \( \hat{\theta} = N / 2 \). Problems such as this are expected to occur more frequently if the sites are nearly independent. We may therefore predict that this estimation technique may fail to give valid parameter estimates when the sites are close to independence. A similar problem with the estimate \( \hat{\theta} \) violating the theoretical limit may occur if the distribution used does not allow the full range of dependence. If, for example, the mixed model - see equation (4.19) - is used then the minimum value allowed for \( \theta = 2A \left( \frac{1}{2} \right) \) is \( \frac{3}{2} \). However, there is no reason why the estimated value, \( \hat{\theta} \), should not by chance be below this limit. Therefore, the estimator may also provide invalid parameter estimates if the distribution used cannot model the full range of dependence.

The crude parameter estimates based on \( \hat{\theta} \) may not therefore be ideal. The method outlined above does, however, have some important advantages. The first is its simplicity. The calculations involved are simple enough for computation by hand and this may allow a range of models to be fitted as a matter of course. Secondly, notice that in constructing the dependence estimate \( \hat{\theta} \) we assume only that the joint distribution of the sites belongs to the bi-variate extreme value family. Hence we may, without specifying the distribution that the sites follow, use \( \hat{\theta} \) as an estimate of the inter-site dependence. Finally we may extend this estimator very easily to distributions in higher dimensions. We know that for a p-dimensional extreme value distribution with unit exponential marginal distributions we may write the survivor function as:
\[ P(z_i > z_i, i = 1, 2, \ldots, p) = \exp \left\{ -\left( \sum_{i=1}^{p} z_i \right) A(w_1, w_2, \ldots, w_{p-1}) \right\} \quad (4.50) \]

where \( w_j = z_j / \sum_{k=1}^{p} z_k, \quad j = 1, 2, \ldots, p - 1 \). Now define:

\[ \theta_p = pA \left( \frac{1}{p}, \frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p} \right) \quad (4.51) \]

Then, using similar steps to the bi-variate case, we see that we may estimate \( \theta_p \) by:

\[ \hat{\theta}_p = N \left[ \sum_{i=1}^{N} \min(z_{1i}, z_{2i}, z_{3i}, \ldots, z_{pi}) \right]^{-1} \quad (4.52) \]

where \( N \) is the number of years data, as before. Again we may use \( \hat{\theta}_p \) to estimate the parameter of the regional distribution or simply as a measure of the regional inter-site dependence. The theoretical limits for \( \theta_p \) are

\[ 1 \leq \theta_p \leq p \quad (4.53) \]

The problem, mentioned for the bi-variate case, of the estimated value exceeding the upper limit for \( \theta_p \) will also occur in higher dimensions. Given the shortage of min-stable distributions that are known to have differentiable survivor functions, the main use of the estimator \( \hat{\theta}_p \) will probably be as a measure of the inter-site dependence taken over the whole region.

The lack of suitable min-stable distributions in more than two
dimensions presents a practical problem. If we wish to fit a min-stable process to a region of many sites so that the dependence function as a whole may be studied - rather than merely a single value from it - what is the best way of achieving this objective? Since we cannot yet study the full model but, for suitable processes, we can derive the bi-variate margins of the process, this suggests that a plausible approach would be to study all the bi-variate marginal dependence functions. Therefore, if we assume that the regional distribution follows some min-stable process then we may find the bi-variate dependence function for this process, \( A(w) \). We then find the value of \( 2A(\frac{1}{2}) \) as a function of parameters of the process, call this value \( \theta_0 \). The for each pair of sites in the region calculate an estimate of \( \theta_0 \) using equation (4.49), call this value \( \hat{\theta}_{ij} \) for sites i and j. Notice that \( \hat{\theta}_{ij} = \hat{\theta}_{ji} \) and \( \hat{\theta}_{ii} = 1 \). Finally we solve for the parameters of the process by minimising the differences between \( \hat{\theta}_{ij} \) and \( \theta_0 \) in some way. Using this methodology Smith (personal communication) proposed the following algorithm for fitting a min-stable process. Suppose that the region contains K sites each with N years record of annual maxima. Then we proceed as follows:

(i) At each site fit a Gumbel distribution. Recall that the Gumbel distribution has the distribution function:

\[
P(X \leq x) = \exp\left( -\exp\left( -\frac{x - \alpha}{\beta} \right) \right) .
\]

(ii) At each site standardize the data to be approximately exponentially distributed and to have mean 1. Hence for each site set:
\[ z_{ij} = \exp - \left[ (x_{ij} - \hat{\alpha}_j) / \hat{\beta}_j \right] \quad i = 1, 2, \ldots, N \nonumber \]
\[ j = 1, 2, \ldots, K \]  \hspace{1cm} (4.55)

where \( \hat{\alpha}_j, \hat{\beta}_j \) are the parameters fitted in stage (i) for each site.

Note that in stage (i) we are not restricted to only fitting the Gumbel distribution. We may select any distribution provided that a transformation from the chosen distribution to the unit exponential distribution exists.

(iii) For each pair of sites calculate the dependence estimator:

\[ \hat{\theta}_{ij} = N \left[ \sum_{n=1}^{N} \min(z_{ia}, z_{jn}) \right]^{-1} \]
\[ i = 2, 3, \ldots, k \text{ and } j = 1, \ldots, i-1. \]  \hspace{1cm} (4.56)

(iv) Calculate the dependence function and find \( \theta_0 = 2A\left(\frac{1}{2}\right) \) as a function of the unknown parameters. Notice that for some processes \( \theta_0 \) can vary between pairs of sites - for instance if we attempt to fit the Gaussian Extreme Value process. Here we will assume \( \theta_0 \) is constant.

(v) Calculate the parameters to minimise

\[ \sum_{i=2}^{k} \sum_{j=1}^{i-1} \left[ \left( \hat{\theta}_{ij} - \hat{\theta}_0 \right)^2 w_{ij} \right] \]
\[ j \]  \hspace{1cm} (4.57)

where \( \hat{\theta}_0 \) is the value of \( \theta_0 \) calculated at the current parameter estimates and \( \{w_{ij} \mid i=2, \ldots, k, j=1, \ldots, i-1\} \) is a set of appropriately
chosen weights.

Therefore using an algorithm such as this it is possible to fit min-stable processes to hydrological annual maxima data. Using these processes we hope to gain an insight into the inter-site dependence structures within the hydrological regions.

5. DATA ANALYSIS

5.1 Objectives

In the previous section it was shown that min-stable processes may provide a useful tool for investigating the level of inter-site dependence. Having also shown a possible estimation procedure we should now attempt to discover if these processes are practically useful. Two steps are necessary to establish this: first, we need to have some idea of the properties of the estimated dependence - bias, variability etc, - and second we need to apply the technique to real data to see how closely min-stable processes can model real dependence structures. For the first step, because of the lack of theoretical results, we will need to assess the properties of the estimators using simulation studies. It must be stressed again that this work has the limited objective of studying the ability of min-stable processes to model the type of dependence structures that are found in real hydrological regions. The problems of using a min-stable process model to predict flood or rainfall quantiles fall outside the scope of the current work.
5.2 Simulation Methodology

We chose to restrict attention in the simulation study to one particular min-stable process. Therefore, all the data will be generated using the Gaussian Extreme Value process. This process is discussed in section 4.6, the bi-variate survivor function and dependence function are given in equations (4.43) and (4.44). This process has the property that the dependence is a function of the distance between the sites which seems a reasonable assumption to make, at least for modelling rainfall data. If we wish to model river flow data, especially if we wish to use the new style non-geographical regions, it may be appropriate to choose a different process to use.

Having chosen the model the first stage in the simulation is to create the “annual maxima” series for each site in the region. One of the advantages of using min-stable processes in a simulation study is that generating data from such a process is simple. Consider the way that a general min-stable process was derived in section 4.4. From this we see that the data are derived by using a Poisson process to generate a random number of locations and magnitudes which give a value at each site via the spread function. Hence to produce data for all sites in the region we simply need to follow the standard procedures for generating points from a Poisson process. In general the procedure is as follows:

(i) In defining the min-stable process we used a region S. Generate a random location within this region, $S_j$.

(ii) For the magnitude find $x_j = \sum_{i=1}^{J} E_i$ where the set \( \{ E_i, \ i = 1, 2, \ldots \} \)

are independent unit exponential random variables. Hence we find that
\[ x_j = x_{j-1} - \ln(1 - r) \]  
(5.1)

where \( r \) is a random number in the range \((0, 1)\) and we set \( x_0 = 0 \).

(iii) For each site \( t_i, i = 1,2,\ldots,k \) find:

\[ z_{ij} = x_j [f(x_j, t_i)]^{-1} \]  
(5.2)

where \( f(s, t) \) is the spread function for the process.

(iv) For each site find:

\[ z_i = \min_{1 \leq r \leq j} \{z_{ir}\}, \quad i = 1, 2, \ldots, k \]  
(5.3)

(v) Suppose that for all \( s \in S \) and for all \( t_i, i = 1,2,\ldots,k \) \( f(s, t_i) < f_{\text{max}} \).

Then repeat steps (i) to (iv) while for any \( \bar{z}_i, \quad i = 1, 2, \ldots, k \):

\[ \bar{z}_i > x_j / f_{\text{max}} \]  
(5.4)

(vi) Repeat steps (i) - (v) \( N \) times to generate \( N \) "years" of data for each site. Notice that at each site the data should follow a unit exponential distribution.

Therefore, to generate data from a particular Gaussian Extreme Value process we need to specify the required spread function. For these simulations we choose to use the bi-variate normal density function with mean vector zero and covariance matrix \( \Sigma \), for various choices of \( \Sigma \). Hence we define the region \( S = \mathbb{R}^2 \) and define \( J \) as a set of \( k \) position vectors in \( \mathbb{R}^2 \). Notice that we cannot define a uniform distribution over
\[ \mathbb{R}^2 \]. This means that we cannot generate the location variables \( s_j \). However, since all the sites in the region will be in some sub-region of \( \mathbb{R}^2 \), say \( s_a \), it will be possible to find another sub-region, \( s_{b'} \), such that:

\[
\min_{1 \leq i, j \leq k} \left[ \int_{s_a} f(s, t_i) ds \right] = 1 - \varepsilon \tag{5.5}
\]

where we can choose \( \varepsilon > 0 \). Hence, if we set \( \varepsilon \) to a small value, then we can make the probability of a pair \((x_i, s_i)\) influencing any of the \( z_i \), \( i=1,2,...,k \), as small as we like. Therefore, we simulate the locations, \( s_j \), only over the sub-region \( s_{b'} \). This sub-region needs to be calculated separately for each set of sites, \( J \), and for each choice of the covariance matrix. This problem clearly will not arise if we choose to use a process defined over a finite space - for example the logistic model, see equation (4.22).

We may reduce the number of cases we need to consider in the simulation by appreciating the fact that we may standardise the site location vectors without affecting the dependence function. Notice, from equation (4.44), that the dependence between two sites depends on the function

\[
a_{ij} = (t_i - t_j)^T \sum^{-1} (t_i - t_j) \tag{5.6}
\]

where \( t_i \) and \( t_j \) are the position vectors of the two sites. Hence in another region, if the locations of the new sites are defined by \( t_i^* = d t_i \) then the dependence between the sites \( i \) and \( j \) will be the same if we choose the covariance matrix, \( \Sigma_{i_j} \), of the second region's spread function to be \( \Sigma_1 = d^2 \Sigma \). Therefore, for these simulation studies we have assumed that the sites are distributed in a unit circle about the origin.
In order to further simplify the simulation procedure and the interpretation of the result we make a further assumption about the structure of the region. The sites in the region will not be allocated randomly. Rather we select a limited set of regional site distribution patterns and simulate using these. This step is merely a convenience since if we used random regions, to ensure we have the range of regional distributions that we require we would need to simulate from many more regions. Random site locations may also help disguise other aspects of the inter-site dependence.

Using the method outlined above we may generate data from a given min-stable process. To fit a min-stable process we use a simplified form of the regional estimation procedure defined in section 4.7. Assume that we have a region of K sites each with N observations. We wish to fit a Gaussian Extreme Value process. Suppose that the spread function for this process is the density function of a bi-variate normal distribution with zero mean vector and covariance matrix \( \Sigma \). Define the inverse of this matrix as:

\[
\Sigma^{-1} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}
\]

(5.7)

We will in future refer to \( \Sigma^{-1} \) as the “dispersion matrix”. We see, from equation (4.57), that the function we have to minimise is

\[
\sum_{i=2}^{k} \sum_{j=1}^{i-1} \left[ \hat{\theta}_{ij} - 2\Phi(a_{ij}/2) \right] w_{ij}^{-1}
\]

(5.8)

where \( a_{ij} \) is defined in equation (5.6). Notice that instead of using equation (5.8) we may simplify the estimation procedure if we find \( \Sigma^{-1} \) by minimising the equation
This is now a linear weighted least squares estimation problem if we estimate \( S_{11}, S_{22}, S_{12} \) - the elements of the dispersion matrix, \( \Sigma^{-1} \) (since combining equations (5.6) and (5.7) we see \( \alpha_{ij} \) is linear in the three unknowns). Recall, from section 4.7, that in some cases we may find \( \hat{\theta}_{ij} > 2 \). We can not use such values in equation (5.9). Since we know that \( \theta \) has an upper limit of 2 then we set \( \hat{\theta}_{ij} \) to a value very close to 2 if we find \( \hat{\theta}_{ij} > 2 \). Hence we decide to force all the dependence estimates to fall in the required range. We may also experience problems if the estimates are based on equation (5.8). If there are values of \( \hat{\theta}_{ij} \) well above the time threshold of 2 then we may expect these results to have an undue effect on the parameter estimates - since these values can tend to dominate the summation. In order to ensure that the estimates \( \hat{\theta}_{ij} \) are all greater than the lower limit of the dependence - that is \( \hat{\theta}_{ij} \geq 1 \) - we simply need to standardise the data at each site to have mean 1.

Therefore, the following methodology was used for the simulations:

(i) Select the min-stable process to be used. For these simulations the Gaussian Extreme Value process was used. To specify the process we therefore need to define the dispersion matrix \( \Sigma^{-1} \). The results reported below were generated using one of the set of dispersion matrices:

\[
\begin{pmatrix}
\frac{1}{2} & 0 \\
0 & \frac{1}{2}
\end{pmatrix}, \quad
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}, \quad
\begin{pmatrix}
2 & 0 \\
0 & 2
\end{pmatrix}, \quad
\begin{pmatrix}
3 & 0 \\
0 & 3
\end{pmatrix}, \quad
\begin{pmatrix}
4 & 0 \\
0 & 4
\end{pmatrix}
\]
We also included the case that all the sites in the region were fully independently distributed.

(ii) Select a regional distribution of sites. The distribution of sites mostly used was a regular lattice of 9 sites centered at the origin with the sides of the lattice of length 1.

(iii) Generate the data for each site. For each site we generate records of 20, 40 and 60 years length.

(iv) Standardise the data at each site to have unit mean.

(v) Calculate the raw dependence estimator for each pair of sites.

(vi) Estimate the dispersion matrix by minimising equation (5.9).

(vii) Use the fitted dispersion matrix to estimate the dependence between each pair of sites.

(viii) Repeat steps (i) - (vii) a large number of times. In practice for each separate min-stable process and for each regional distribution 250 replications were performed.

To assess the accuracy of the estimated dependence structure the bias and residual mean square error of both the raw dependence estimator and the fitted dependence were calculated for each pair of sites over the given set of replications.

5.3 Simulation Results

The first step is to consider the properties of the dependence
estimator, \( \hat{\theta}_{ij} \), defined in equation (4.49). We shall call this the "raw dependence estimator". The properties of this estimator are of interest for two reasons: we wish to use the raw dependence estimator to fit a min-stable distribution and we also wish to assess its use as a measure of inter-site dependence.

To assess the accuracy of the raw estimator we first consider its bias. Figure 3 shows the mean bias of the raw estimator plotted against the true dependence. The 95% confidence limits about this mean value are also shown. Figure 4 shows the mean bias as a percentage of the true dependence. These results were obtained using a variety of dispersion matrices and also using regions of fully independent sites, that is with dependence function equal to 2. The results obtained for the bias of the raw estimator did not appear to be much influenced by the choice of region or the dispersion matrix - for a given true dependence the raw estimator produced consistent results for a wide range of regions and dispersion matrices. Any effects of altering the regional structure appear to be swamped by the general variability of raw estimator. The general pattern of the bias shown in Figures 3 and 4 is mostly as expected: the value of the dependence functions is over-estimated near complete dependence and under-estimated nearer to independence. The surprising feature of these results is the near unbiasedness of the raw estimator when the sites in the region are exactly independent. It is not clear if, as the true dependence function increases from 1.9 to 2, the bias of the raw estimator varies continuously from the relatively large negative value at 1.9 to the near zero for complete independence or if there is a discontinuity.

We now need to consider the variability of the raw estimator. Figure 5 shows the range of values that the raw estimator produced taken over all replications and plotted against true dependence. We can also see from this graph at what dependence levels the raw estimator exceeded its theoretical bound. The fact that this occurred for a dependence as low as
1.3 is somewhat surprising. Figure 6 shows the estimated standard deviations of the raw estimator plotted against the true dependence. For moderate dependence values the standard deviation appears almost a linear function of the true dependence, however as the dependence becomes near to 2 the standard deviation rapidly increases.

To combine bias and variability, Figure 7 shows the root mean square error - RMSE - of the raw estimator plotted against the true dependence. Notice how that RMSE decreases at complete independence - the reduction in the bias dominates the increase in variability. This feature is more apparent in Figure 8 where RMSE is given as a percentage of the true dependence. From this figure we also notice the clear minimum around a dependence value of 1.3.

In Section 4.7 we commented that the raw dependence estimator could exceed its theoretical limit of 2. It was also suggested that this was most likely to occur for sites that were nearly independent. We have seen, from Figure 5, when the raw estimator has been observed to be greater than 2 in these simulations. Figure 9 shows the percentage of pairs of sites that produced a raw dependence estimate greater than 2 taken over all replications and plotted against the true dependence. For clarity we omit the result for an exactly independent region. In this case we find that nearly half the pairs of sites produce raw dependence estimates greater than 2. Hence we see that to produce a raw dependence estimate greater than 2 is expected to be a rare event unless the true intersite dependence value is quite high, say over 1.8.

These results for bias, variability and the estimator exceeding 2 tend to be consistent over the range of regions and dispersion matrices used. The results do depend however, on the number of years data used at each site. The raw estimator becomes in general more accurate as longer data records are used - the absolute value of the bias decreases and so does the standard deviation. This increase in accuracy is however, slow and appears at its slowest near complete independence.
In general we may say that the accuracy of the raw estimator is not good. The estimator combines moderately high bias - up to 15% of the true dependence - with quite high variability. Since the raw estimator will tend to over estimate the dependence value near complete dependence and under estimate it near independence in practice if we observe an estimated dependence in the mid-range of value then we cannot have much confidence that this value is an accurate description of the true inter-site dependence. However, the raw estimator may still be of use if we wish to decide if a region is independent. Since for an exactly independent region the raw estimator appears nearly to be nearly unbiased we may wish to construct a test for the estimator equals 2. If we can demonstrate that, in an independent region, the raw estimator takes a certain distribution - or if we are willing to assume a distribution - then we may generate such a test. However, given the high variability of the raw estimator in independent regions, we may expect the power of any such test to be low. In an informal and heuristic way we might also try to use the fact that in truly independent regions it is common to find the raw estimator greater than 2. Hence, we might suspect a region is not independent if no estimates are greater than 2 although it probably is not possible to quantify this "suspicion" in any way.

We now consider the fit of the Gaussian Extreme Value distribution. Clearly, since the model is fitted using the raw dependence estimator discussed above the fit will be rather variable. Unlike the raw estimator we may not simply plot the bias and variability of the modelled dependence against true dependence. This is because the modelled dependence is a function of both the dispersion matrix, \( \Sigma^{-1} \), and the distance between the sites - as we see from equation (4.49). Hence the fit of the modelled dependence will vary between regions in a way that the raw dependence estimator did not appear to. Within a region however, the bias and variability of the fitted dependence, will be qualitatively similar to that shown in Figure 3 (for bias) and Figure 6 (standard deviation). The best modelled fits are obtained in regions with evenly
distributed sites rather than regions with "clusters" of sites. We may improve the general fit of the model over the region if we use a weighted minimisation on equation (5.9). To minimise the variance of the modelled dependence we should choose \( w_{ij} \) proportional to the variance of the raw dependence estimator. As a first step the weights \( w_{ij} = \hat{\theta}_{ij} \) and \( w_{ij} = \hat{\theta}_{ij}^2 \) were used. It was found however that mean sum of squared difference between the modelled and true dependences calculated using these two different weights were almost the same. Figure 10 shows the mean sum of squared difference for the modelled dependence fitted using the weights \( w_{ij} = \hat{\theta}_{ij} \) and for the unweighted case plotted against the mean regional dependence. If we look at the fit of the estimated dispersion matrix this is found to be extremely variable. The estimated standard deviations of the elements of the dispersion matrix can be as high as half the estimated value itself and are typically not less than 10% of the estimated value. The use of a weighted fit reduces this variability only slightly.

We know that the raw dependence estimator may produce invalid estimates of dependence. It was found that in some circumstances the model could fail to find a valid solution in that the estimated dispersion matrix was not positive definite. As with the illegal raw dependence estimates this problem tended to occur in regions in which the sites were all nearly independent. This failure to fit was however fairly rare, occurring in at most 1% of the replications for the most independent regions modelled. Curiously no failure to fit was observed for any of the exactly independent regions.

Clearly the accuracy of the modelled dependence is limited by the results obtained from the raw dependence estimator. To improve the modelled dependence an obvious step would be to find a more accurate form of the raw dependence estimator. If this is not possible then we may experiment with different weighted fits but in general the fit of the
modelled dependence is not good.

This method of fitting a min-stable process over a region is therefore a feasible way of studying the dependence structure of the region. However, we must accept that the dependence estimates produced are liable to be quite imprecise. The raw dependence estimator used by itself may also be seen as a simple, feasible but imprecise method of studying the inter-site dependence structures of real hydrological regions. The raw dependence estimator may be able to identify those regions that almost certainly are, or are not, independent. We may readily extend the raw dependence estimator to deal with any number of sites, rather than simply looking at pairs of sites. However, given the imprecise nature of the estimates even in the simple bivariate case this may not prove to be of much value. To evaluate the raw dependence estimator used in higher dimensions remains to be done.

The main conclusion that we may draw from this simulation study is the usefulness of min-stable processes as a simulation tool. We may use a min-stable process in an elegant and simple way to generate dependent data exhibiting a wide variety of dependence structures. Furthermore, the "site" data generated by a min-stable distribution - which should be drawn from unit exponential distributions - may be transformed individually into more "flood like" distributions. Hence we may use min-stable processes to simulate data from a region containing sites that are neither identically nor independently distributed. This should prove to be an extremely useful tool for studying the effects of non-homogeneity on the regional estimation techniques.

5.4 Analysis of Real Data

To examine the performance of the raw dependence estimator
and the regional min-stable model using real annual maxima rainfall data we select the southern rainfall region of the U.K. - as defined by the Flood Studies Report, N.E.R.C. (1975). This region has as its core Hampshire, Sussex, Surrey and London and it contains 405 rain gauges. Clearly to examine the full region is not practical - we would need to examine 81810 pairs of gauges for inter-site dependence. To reduce the problem to more manageable proportions we examined various small sub-regions of rain gauges. Firstly, we looked at sub-regions created by selecting a site at random and then finding its nine nearest neighbours. These sub-regions should be in general quite strongly dependent. Other sub-regions were selected by finding ten gauges with a minimum inter-site distance above some, large, limit. These sub-regions should be much closer to independence.

For all the sub-regions analysed the estimation technique finds a valid min-stable model. The fit of the model to the raw dependence estimator appears to be conditioned by two main factors: the average estimated dependence in the sub-region and the distribution of the sites in the sub-region. The fit generally improves as the average estimated dependence tends towards 1 and also improves if the gauges are evenly distributed within the sub-region. If the sub-region contains separate groups of gauges then the fit of the model can be noticeably degraded. From the previous section we know that the raw dependence estimator can be rather inaccurate and so we can only examine the modelled dependence results in a very general way. One interesting result that requires comment is the relative lack of raw dependence estimates greater than 2 even for sub-regions containing very widely dispersed sites that should be nearly independent. In the sub-regions studied we seldom observe more than one pair of sites (out of 45 possible pairs) with an estimated raw dependence greater than 2. Compare this with the percentage of raw estimates exceeding 2 generated by the simulation study, see Figure 9.

As an example of the kind of results obtained we now present the
results for one particular sub-region. This sub-region was defined by choosing a rain gauge and then finding the nine nearest gauges. Hence the sub-region may be expected to be reasonably dependent. Figure 11 shows the locations of the gauges and Table 1 shows the results for the raw and modelled dependence for this sub-region. In general the agreement between the modelled and raw dependence is good. Notice that gauge 7, the central gauge, has the worst fit of all the gauges. We find that the mean squared difference between the raw and modelled dependence for this sub-region is 0.0042, and of this no less than 0.0032 is accounted for by gauge 7. Hence we might attempt to improve the fit of the model by excluding gauge 7. However, perhaps a more reasonable approach is to notice that gauges 2, 3, 4, 5, 6 and 7 form a sub-region within the main sub-region. To improve the modelled fit we attempt to exploit the fact that the fit tends to improve as the average level of dependence increases and the distribution of the sites becomes more even. Table 2 shows the comparison between the raw and the modelled dependence estimates for this new sub-region. Obviously the raw dependence estimates are the same as before. We find that the mean squared difference between raw and modelled dependence is reduced to 0.0017. In this new sub-region the raw estimated dependence between gauges 6 and 7 stands out as strange. The unexpectedly high value may well be due to some real factor making the two gauges less dependent. Alternatively it may be an illustration of how variable the raw dependence estimator can be.

The real data analysis has been successfully performed on a variety of data sets. However, the knowledge that the raw dependence estimator tends to be rather inaccurate and the presence of some raw estimates that appear to be outliers means that we must be cautious when using these results to assess the true level of inter-site dependence within a region.

In both the real data analysis and the simulation study we used a model that relates inter-site dependence to the distance between the
gauges. This is a reasonable model to use to study the dependence in rainfall regions but it is probably inappropriate to use such a model to analyse flood flow data. However, the use of the raw dependence estimator and some suitable min-stable process would be a perfectly feasible way of studying inter-site dependence in the flood flow data.

**CONCLUSIONS**

We conclude this work with a brief summary of what has been achieved and suggest some ways in which these results may be used and perhaps improved. A new method for studying dependence structures in hydrological regions has been developed using the properties of min-stable processes. The uses of these min-stable processes have been examined and we have drawn some conclusions about their merits and limitations.

We consider first a limitation of min-stable processes as a hydrological model. In this work all we have attempted is to model the dependence structures in hydrological regions. Is it not proposed that we should use min-stable processes as a full regional estimation technique. This is in part due to the fact that despite knowing what form a min-stable survivor function must have - see equation 4.9 - it has not yet proved possible to generate min-stable distributions that have continuous density functions for more than two variables. Indeed the number of bi-variate distributions found is not great. However, we may attempt to circumvent this problem by assuming that some min-stable process governs the behaviour of the region's data and then studying the inter-site dependence between all pairs of sites. Using the inter-site dependence in this way we need a method for fitting a min-stable process. A fitting method has been developed and its properties assessed using
simulation studies. The raw estimator on which the estimation technique was based was found to be rather inaccurate - both variance and bias tended to be large. Hence in practice if we wish to make any statements about the level of inter-site dependence in a region the confidence interval around any estimate that we give will be very wide. Using a min-stable model for the dependence structure does not allow us to find precise estimates of the true dependence. However, the use of a min-stable process does allow us to model a very large range of dependence structures. These dependence structures also have the advantage that they arise from the extreme value nature of the data rather than using some form of linear correlation as the basis for the dependence structure.

There are two main advantages of using min-stable processes as models for inter-site dependence. The first is that although the estimates may be imprecise the estimation technique outlined in Section 4.7 is both simple and feasible. This, coupled with the very wide range of dependence structures that a min-stable process can model, means that we may use this method as a quick and simple rough guide to the level of dependence in a region. The second advantage is probably the most important reason for being interested in min-stable processes. Using a min-stable process is an extremely easy method for simulating data for regions containing dependent sites - and, if necessary, we may also ensure that the sites are non-identically distributed. Hence a min-stable process is an extremely useful tool if we wish to generate data from a non-homogeneous region to study the effect of heterogeneity on the regional estimation techniques currently in use.

We should now ask how these min-stable processes may be used in the future. The first way in which they may be used is to exploit their usefulness as a tool for generating heterogeneous data. The wide range of dependence structures that may be used would allow an extensive investigation of the effects of inter-site dependence. There are also certain things that may be tried to improve the usefulness of min-stable
models in measuring the level of inter-site dependence in real data. The best way of improving the accuracy of the modelled dependence would be to find a better raw dependence estimator. Hence it may be profitable to search for some new dependence estimator. The fact that we have only managed to produce continuous min-stable distributions in two dimensions is a serious weakness and the search for distributions in higher dimensions should continue. If this search is successful it may at some stage prove possible to develop a full min-stable regional estimation technique.

At present very little is known about the effects of inter-site dependence on the regional estimation techniques now in use. The use of min-stable processes provides a simple and elegant method for assessing these effects by simulation. These simulation studies could use the very wide range of dependence structures that min-stable processes can model to assess the effects of inter-site dependence over a range of dependence structures that may occur in the real data. This work could make a major contribution to the confidence we have in the regional estimates and may be regarded as work of high priority.
REFERENCES


Comparing the raw dependence estimator and the modelled dependence for a real rainfall sub-region.

<table>
<thead>
<tr>
<th>Gauge Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Raw</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Raw</td>
<td>1.50</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1.47</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Raw</td>
<td>1.47</td>
<td>1.01</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1.48</td>
<td>1.00</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Raw</td>
<td>1.49</td>
<td>1.25</td>
<td>1.14</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1.62</td>
<td>1.06</td>
<td>1.04</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Raw</td>
<td>1.45</td>
<td>1.21</td>
<td>1.19</td>
<td>1.21</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1.56</td>
<td>1.06</td>
<td>1.04</td>
<td>1.00</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Raw</td>
<td>1.46</td>
<td>1.33</td>
<td>1.34</td>
<td>1.40</td>
<td>1.32</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1.42</td>
<td>1.08</td>
<td>1.06</td>
<td>1.04</td>
<td>1.02</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Raw</td>
<td>1.60</td>
<td>1.47</td>
<td>1.44</td>
<td>1.50</td>
<td>1.44</td>
<td>1.45</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1.31</td>
<td>1.13</td>
<td>1.11</td>
<td>1.10</td>
<td>1.07</td>
<td>1.01</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Raw</td>
<td>1.63</td>
<td>1.55</td>
<td>1.58</td>
<td>1.39</td>
<td>1.37</td>
<td>1.20</td>
<td>1.60</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1.73</td>
<td>1.63</td>
<td>1.58</td>
<td>1.42</td>
<td>1.39</td>
<td>1.30</td>
<td>1.27</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Raw</td>
<td>1.29</td>
<td>1.52</td>
<td>1.48</td>
<td>1.41</td>
<td>1.46</td>
<td>1.40</td>
<td>1.58</td>
<td>1.38</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>1.33</td>
<td>1.60</td>
<td>1.57</td>
<td>1.53</td>
<td>1.47</td>
<td>1.31</td>
<td>1.21</td>
<td>1.18</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10 Raw</td>
<td>1.22</td>
<td>1.46</td>
<td>1.44</td>
<td>1.38</td>
<td>1.43</td>
<td>1.46</td>
<td>1.50</td>
<td>1.42</td>
<td>1.17</td>
<td>1</td>
</tr>
<tr>
<td>Mod</td>
<td>1.19</td>
<td>1.51</td>
<td>1.48</td>
<td>1.49</td>
<td>1.43</td>
<td>1.27</td>
<td>1.16</td>
<td>1.27</td>
<td>1.02</td>
<td>1</td>
</tr>
</tbody>
</table>

For the locations of the gauges see Figure 11.
Mean sum of squared differences = 0.0042.
TABLE 2

Comparing the raw dependence estimator and the modelled dependence for the small rainfall sub-region.

<table>
<thead>
<tr>
<th>Gauge Number</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Raw</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Raw</td>
<td>1.01</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>1.00</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Raw</td>
<td>1.25</td>
<td>1.14</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>1.26</td>
<td>1.19</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Raw</td>
<td>1.21</td>
<td>1.19</td>
<td>1.21</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>1.22</td>
<td>1.15</td>
<td>1.02</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Raw</td>
<td>1.33</td>
<td>1.34</td>
<td>1.40</td>
<td>1.32</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>1.29</td>
<td>1.23</td>
<td>1.26</td>
<td>1.14</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Raw</td>
<td>1.47</td>
<td>1.44</td>
<td>1.50</td>
<td>1.44</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>1.52</td>
<td>1.47</td>
<td>1.59</td>
<td>1.44</td>
<td>1.09</td>
</tr>
</tbody>
</table>

For the locations of the gauges see Figure 11.
Mean sum of squared differences 0.0017.
FIGURE 1

Showing the relationship between the "regional" and "typical" growth curves.
FIGURE 2

Range of the dependence function, $A(w)$. 

[Diagram showing range of dependence function]
FIGURE 3

Mean bias of the raw dependence estimator (with 95% confidence interval).
FIGURE 4

Mean bias of the raw estimator as a percentage of the true dependence.
FIGURE 5

Maximum, mean and minimum values for the raw dependence estimator.
FIGURE 6

Standard deviation of the raw dependence estimator.
FIGURE 7

RMSE of the raw dependence estimator with 95% confidence limits.
RMSE of raw dependence estimator as a percentage of the true dependence.
Percentage of raw estimated values that exceed 2.

FIGURE 9
FIGURE 10

Mean sum of squared differences between fitted and true dependence.
FIGURE 11

Location of rainfall gauges for the real data analysis.