Flood Frequency Analysis of the Waimakariri River

Robert Ware[1] and Frank Lad[2]

Abstract

Different approaches to flood frequency analysis are investigated, with particular emphasis on estimating extreme hydrological events for a site, or group of sites. Frequentist approaches to flood estimation are examined. At-site and regional estimates are studied, and their accuracy and precision compared.

Flood exceedance quantiles are assessed using updated mixture mass functions as sequential forecasting distributions. These sequential forecasts are scored using three different scoring rules for distributions: the quadratic, logarithmic and spherical.

Both the frequentist methods and the digital forecasting procedures are applied to data collected from the Waimakariri River in Canterbury, New Zealand. Finally, we compare the appropriateness of the frequentist and digital methods. It is found that the mixture distributions computed via the discrete digital method provide much more uniform forecasts across an array of proposed distribution families than do the frequentist forecasting methods.

Key Words: Extreme floods; Digital mass functions; Sequential forecasting; Scoring Rules; Waimakariri River.

1 Introduction

A common problem in many areas of environmental engineering is that of estimating the return period of rare geophysical events, such as extreme floods, for a site or group of sites. A large volume of work considering the estimation of flood risk has appeared in the last 20 years. Approaches have ranged from arcane mathematical formulations to operational institutional guidelines. Despite the large number of publications there is no consensus on how best to proceed. The problem is complicated by the necessity of evaluating flood risk for return periods that exceed the length of the observed record.

Modern flood frequency theory is typical of much conventional statistical theory, in that most effort is expended in determining an appropriate form to model

[1]Research Fellow, School of Population Health, The University of Queensland

[2]Research Associate, Department of Mathematics and Statistics, University of Canterbury

the "underlying distribution" of floods, and then estimating the parameters of this underlying distribution. Conventional estimates of flood exceedance quantiles are highly dependent on the form of the portion of the underlying flood frequency distribution (the right tail) which is most difficult to estimate from observed data. Currently there is no compelling theory on which to base the distributional form of the right hand tail.

This Report will begin with an introduction to flood frequency estimation. We introduce the process undertaken to measure river flow, concentrating on details specific to the Waimakariri River. In Section 3 we examine different frequentist approaches to estimating flood exceedance quantiles. We study both at-site and regional estimates, and compare their accuracy and precision. In Section 4 we develop a procedure for forecasting exceedance quantiles of floods. Different sets of results are obtained and their scores computed. This work is based on the digital updating procedure introduced and studied in Ware and Lad (2003). In Section 5 the results obtained through the frequentist and digital methods are compared. First, we use subjective methods, scoring the conditional expectation and variance for both the Waimakariri River data and a simulated data set. Then we use an objectivist method to compare the accuracy and precision of the two methods for a simulated data set. Finally, a summary of this Report is presented in Section 6.

Note that although the word "exceedance" is not found in the Oxford English Dictionary, it is commonly used in environmental engineering, particularly when the analysis of extreme events is studied. It is probably based on "excedent", meaning "The portion or quantity in excess", as this the meaning regularly attached to it. Nonetheless it is commonly spelled "exceedance" in the general engineering literature that uses the word. See, for example, the standard work of Metcalfe (1997).

2 River Flow Records

Measurements of instantaneous river flow can be produced by combining measures of flow velocity and cross-sectional area of the river. This Section will detail the process taken to measure river flow, concentrating on procedures used specific to the Waimakariri River. First, we describe the process undertaken in New Zealand

to obtain river flow measurements. Second, we introduce the Waimakariri River in depth, and outline its measurement history. We conclude this Section by detailing the problem of estimating flood exceedance quantiles.

2.1 River Flow Measurement in New Zealand

Throughout New Zealand there are a number of government funded authorities responsible for recording river flow. Prior to amalgamation in 1989, each local catchment board was responsible for keeping records about rivers within its catchment. Records for the Waimakariri River were gathered by the North Canterbury Catchment Board. Since amalgamation, water records have become the responsibility of district councils, regional councils and branches of the National Institute of Water and Atmospheric Research (NIWA). For example, records for the Waimakariri River are held by the Canterbury Regional Council (known for promotional purposes as Environment Canterbury), records for the Rakaia River are held by NIWA-Christchurch and records for the Waihopai River are held by the Marlborough District Council.

River flow is measured by combining measurements of a river's water flow velocity and cross-sectional area. A "stage-discharge rating curve relationship" is used to combine these two measures. Stage-discharge rating curves have been constructed for every New Zealand river whose flow is of interest, and are used to estimate the volume of flow in a given time period. The units used are cubic metres per second, or cumecs. The annual maximum instantaneous flood peak is the name given to the largest of these measures of volume in one calendar year. A sequence of these maximum flow values is called an annual maxima series (AMS).

To measure the cross-sectional area of water, two components are needed: the water level and river bed profile. Today, the water level of many New Zealand rivers, including the Waimakariri, is recorded mechanically. The standard interval between recordings is fifteen minutes. The river bed profile of each river whose flow is of interest is mapped on irregular occasions. This cross-section is re-mapped whenever the river's controlling authority suspects the profile may have changed. Re-mapping usually occurs directly after a notable flood event, or "fresh". There

are approximately 10–15 freshes each year in the Waimakariri River, although the bed profile is not re-mapped after every one.

Water velocity is recorded at a number of points across a river. Water velocity was originally recorded either by standing in the river holding a current meter, or by dropping a current meter over the side of a boat. Now, velocity is measured electronically, from a boat that is being driven back and forth across a river. Relationships between average cross-sectional velocities and water level are developed, to allow for extrapolation to large events. It is known hydraulically and through measurement how velocity increases to an asymptotic limit with increasing depth. Since it is not safe to use a boat to undertake an accurate measure of water velocity while a flood is in progress, a more informal process may be used to record the measurement. During extreme floods, the flow velocity can instead be measured from a bridge or cableway, or flow may be measured indirectly, by using standard water depth-velocity relationships. Consequently the measure of water velocity will be less accurate than it is when the river is flowing at a normal level. This a real example of a problem we commented on in Section 2.4 of Ware and Lad (2003), where a measurement device is unable to measure extreme values to the same level of fineness with which it is able to measure commonly recorded values. In assessing the measurement process for river flows it must be recognised that the precision of the process is higher for normal flows than for extreme flows.

The complete river flow measurement process consists of a series of approximations. Flood flow records for large rivers, including the Waimakariri, are recorded in units of $10m^3/sec$ (that is, "ten cumecs"), reflecting the lack of accuracy resulting from the series of approximations inherent within each of the measures in the series. It is standard for NIWA to assume that their flow measures are within $\pm 8\%$ of the "true flow" 95% of the time. It is conventional statistical practice to presume that the recorded measurement is equal to the true measurement value plus some unknown (and unobservable) measurement error, or, symbolically, $X = \mu + \epsilon$. Although in this Thesis we treat such a viewpoint as, in practice, meaningless, it is nevertheless interesting to know with what precision NIWA regard their measurements, particularly when defining the width between successive values in the realms of X and the characterising parameters.

2.2 The Waimakariri River

The Waimakariri River is located in the Canterbury region of the South Island of New Zealand. It is classified as a "Main Divide" river, meaning that it has a catchment which drains from the ranges east of the Main Divide of the South Island. It has a catchment area of $3210km^2$, the largest of any river in Canterbury. The Waimakariri River flows through the northern outskirts of Christchurch. Flood protection works have been constructed to protect most of urban Christchurch and Kaiapoi.

The daily flow of the Waimakariri River has been recorded since 1930 at the site of the Old State Highway One bridge. The site is $5.4km$ from the mouth of the river, and consequently the water level is effected by the tide. The Waimakariri River is also measured at a site in the Waimakariri Gorge. When records of river levels do not exist, for example if the mechanical recorder is broken or if the flow records have been lost, flow records are estimated based on water levels at the Waimakariri Gorge recording site. Studies of the relationship between recorded flow levels at these two sites show this to be a reasonable resolution to the "missing data problem". Recent Waimakariri River flows can be viewed on the Canterbury Regional Council's internet site[3].

Between 1930 and 1966 water levels at the Old State Highway One bridge site were recorded visually at irregular intervals. The AMS series for this period was calculated retrospectively using slope-area calculations and records from the Waimakariri Gorge site. "Slope-area gaugings" are post-flood event measurements. They take into account the highest water level mark, the associated cross-sectional area of the channel, slope of the channel bed, and likely velocities of the flow as read from a standard water depth-flow velocity relationship table. It is widely recognised that slope-area gaugings are less precise than records obtained via conventionally recorded flow levels. In 1966 a mechanical recorder was installed to record the water level every 15 minutes.

Flow measurements for years 1960 through 1966 were complicated by the constant change of the river-bed profile due to large amounts of shingle being removed

[3] www.crc.govt.nz/Water/Rivers-Rainfall/graphist.asp?site_no=66401

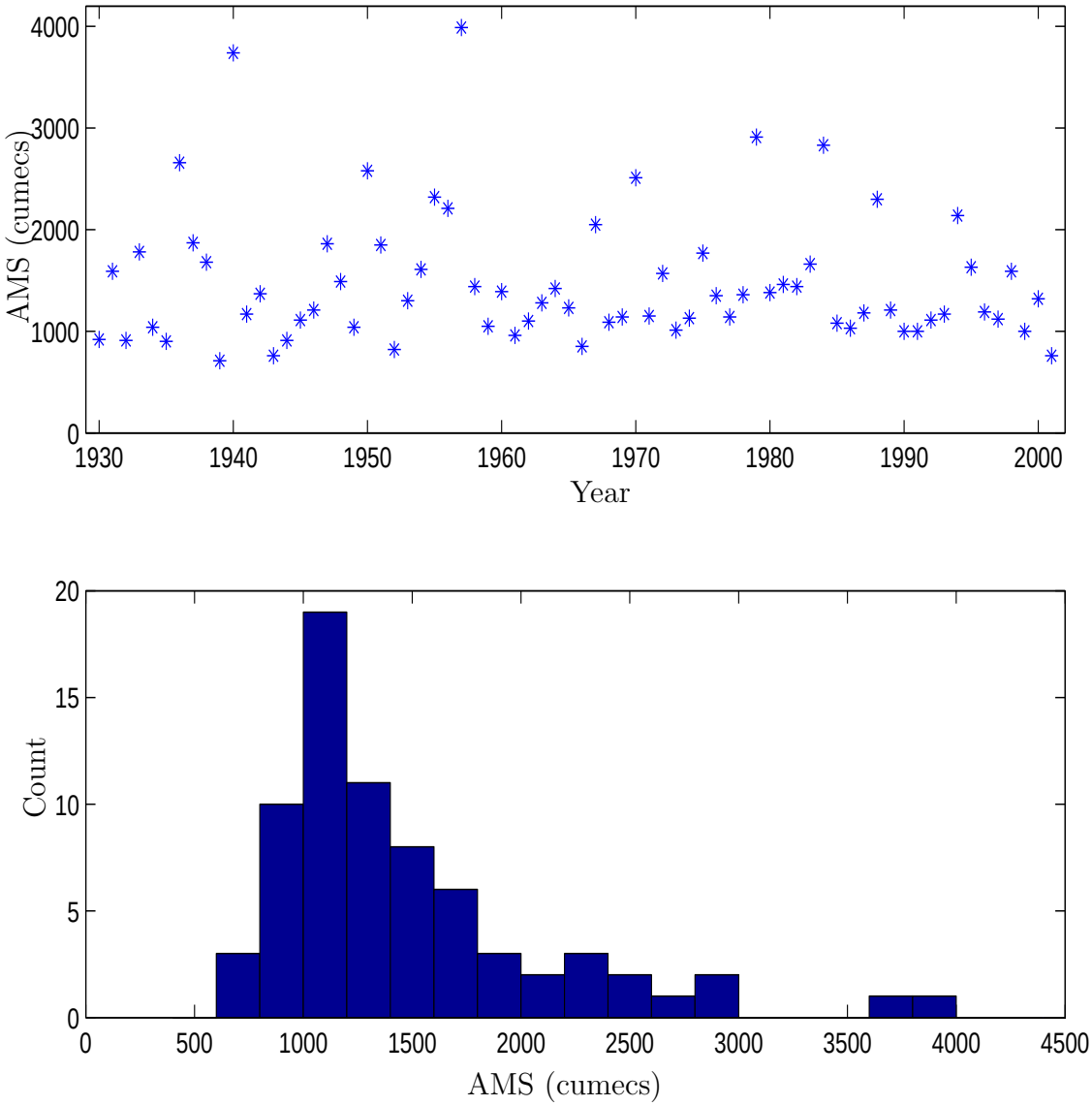

Figure 1: AMS recorded for the Waimakariri River.

from the river bed for use in construction of Christchurch's Northern motorway. During the period on record there has been a small (< 10 cumecs) amount of upstream water diversion. This is an insignificant amount considering the level of measurement precision.

The highest mechanically measured discharge for the Waimakariri River is $3070 m^3/sec$ (1979), a value exceeded twice in the recorded period (1940 and 1957). AMS values range between $710 m^3/sec$ (1939) and $3990 m^3/sec$ (1957) with mean $1485 m^3/sec$, standard deviation $632 m^3/sec$ and skewness 1.77. The upper panel of Figure 1 shows a timeplot of the recorded AMS for the Waimakariri River. The lower panel shows a histogram of the observations, sorted into bins of width $200 m^3/sec$.

2.3 Problem Statement

The problem assessed in this Report concerns the characterisation of extreme floods. At each site i we want to estimate the Fth quantile of non-exceedance probability, denoted $Q_i(F)$, $0 \leq F \leq 1$. Denoting by X_i the annual maximum instantaneous flood peak at site i, the quantile $Q_i(F)$ is the value we expect X_i to exceed with probability $(1-F)$ during the year of interest. That is, $P(X_i \leq Q_i(F)) = F$ for the year of interest.

$Q(F)$ is interpreted to mean that, in any year, we expect that there is a $F\%$ chance that X will be less than $Q(F)$. Conversely there is a $(1-F)\%$ chance that X will exceed $Q(F)$. The "return period" of a flood exceedance is defined to be the reciprocal of the probability of exceedance in one year, $1/(1-F)$. It is the expected time between exceedances of the size of $Q(F)$.

$Q(F)$ is commonly (and misleadingly) referred to as a "one in $1/(1-F)$ year flood" See, for example, Metcalfe (1997). For instance $Q(0.95)$ is the value that we expect X to exceed, during the course of a year, with probability 0.05. The implication of the phrase "one in twenty year flood" is that $Q(0.95)$ will only be exceeded once in the next twenty years. However, a standard binomial calculation shows that the probability of the flow exceeding $Q(0.95)$ in at least one of the next twenty years is almost $\frac{2}{3}$. If we define Y to be the number of years that $Q(0.95)$ is exceeded in a twenty year period, then the probability of observing at least one exceedance in that twenty year time period is

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y = 0) \\ &= 1 - \prod_{i=1}^{20} P(X_i = 0) \\ &= 1 - (0.95)^{20} \\ &= 0.64 \quad (2\text{dp}). \end{aligned} \tag{1}$$

Table 1 displays the actual probabilities that $Q(0.95)$ is exceeded in Y separate years over a twenty year period. Notice that the chance of $Q(0.95)$ being exceeded in one of the twenty years is 0.3774 – which differs considerably from the implication of "one in twenty year flood". Nonetheless the terminology "one in $1/(1-F)$ year flood" is standard within the engineering community, and we shall follow it here.

Y	0	1	2	3	4	5	6+
$P(Y)$	0.3585	0.3774	0.1887	0.0596	0.0133	0.0022	0.0003

Table 1: The probability that $Q(0.95)$ will be exceeded in Y years over a twenty year period.

3 Frequentist Approach to Estimating Flood Exceedance Quantiles

Conventional frequentist flood estimation theory has centred around estimating the parameters of the "underlying distribution" of the floods. This underlying distribution is believed to be a stochastic generating structure that produces a random outcome which is of interest to the researcher. The aim of frequentist estimation is to find the characterising parameter values of this, unobservable, underlying distribution. An opposing view is that there is no "correct" functional form that stochastically generates random outcomes — the observations are what they are, nothing more or less. People are uncertain about what values subsequent observations will be, and probabilities represent their informed knowledge. For the majority of this Thesis this is the paradigm that is followed. However for the remainder of this Section we shall assume that the underlying distribution exists as a meaningful concept, as a frequentist statistician does at all times.

3.1 At-Site Flood Frequency Analysis

The simplest flood estimation methods involve collecting AMS data for a site, and using this data to estimate the characterising parameters of the underlying distribution, the functional form of which is assumed known. Popular estimation techniques include the method of moments and maximum likelihood estimation. The method of moments is notoriously unreliable for fitting extreme value distributions due to the poor sampling properties of second and higher order sample moments. The method of maximum likelihood has been used when dealing with extreme values, however it doesn't work well when the sample size is small to moderate. Moreover its computational aspects are based on iterative procedures which require reasonable starting

values. For the parameter estimation undertaken in this Report we shall use the method of L-moments, a linear extension of the conventional method of moments. L-moments have been widely used in recent studies of extreme phenomenon. For a taste of the breadth of current research see Kjeldsen et al. (2002), Kroll and Vogel (2002) and Park et al. (2001) to see L-moments applied in studies in South Africa, the United States of America and South Korea. The theory of L-moments is introduced in Section 3.3. Three popular candidates for underlying flood distribution are introduced in Section 3.4

3.2 Regional Flood Frequency Analysis

Recent research into flood frequency estimation has focused on developing and evaluating regionally derived flood frequency estimates. In regional flood frequency analysis it is assumed that the data from all gauged sites in a region can be combined in such a way as to produce a single regional flood frequency curve. This curve is applicable, after appropriate rescaling, anywhere in that region. Regionalisation allows us to pool data from m sites. Each site has n_i years of recorded measures, where n_i can be of any length.

Conventional regionalisation techniques identify a fixed set of recording sites which adjoin each other. Each region is identified by considering which sites are 'close' to each other. Proximity can be assessed using statistical measures (e.g. coefficient of variation (CV) or ratio of mean flow to drainage area) or spatial measures (e.g. longitude and latitude of each site).

The biggest advantage of regional estimation is seen to be the increase in record length. A regional approach is necessary when estimating floods at sites with no observed data. Many studies (e.g. Lettenmaier et al., 1987; Hosking, 1990) have shown that flood estimates based on regional information are more accurate (have less absolute error) and are more stable (have less variance) than those based solely on at-site records. The most commonly used regionalisation techniques are based on the index flood approach.

3.2.1 The Index Flood Approach

The index flood approach was first introduced by Dalrymple (1960), and has since been implemented on a regular basis. See the review article by Stedinger and Lu (1995) for examples. It was developed as a way of deriving a regional frequency curve. The underlying flood frequency distribution at each site is assumed to be identical, except for a scale factor. Consequently we are able to use a straightforward pooling approach. First, the data at each site are normalised by the index flood (details of this procedure will be described shortly). Next, the parameters of a dimensionless regional flood frequency curve are estimated. Finally, the parameters are rescaled at the site of interest by a local estimate of the scaling factor, usually the at-site mean.

The key assumption of an index flood procedure is that the region is homogeneous, that is, the frequency distributions of the N sites in a region are identical, apart from a site-specific scaling factor. The distribution common to all sites in the region is called the regional frequency distribution. It is dimensionless and defined by its (regional) quantiles, $q(F)$, $0 \leq F \leq 1$. It is usually assumed that the form of $q(F)$ is known apart from p undetermined parameters $\theta_1, \ldots, \theta_p$. The site-specific scaling factor is called the index flood, denoted μ_i at site i (see Hosking and Wallis, 1993). The index flood is usually taken to be the sample mean of the frequency distribution at site i, although any location parameter of the frequency distribution may be used instead. For example, Smith (1989) uses the quantile $Q(0.9)$. Thus we can write

$$Q_i(F) = \mu_i q(F), \qquad i = 1, \ldots, N, \tag{2}$$

where $Q_i(F)$ is the quantile of non-exceedance probability F at site i.

A standard scaled data approach is the simplest index flood method. This involves dividing each measure by its at-site sample mean, and then treating all the scaled data points as if they were observations from the regional frequency distribution. Parameter estimates are found and the estimated regional flood distribution is then multiplied by the at-site mean of the site under investigation.

A more advanced index flood procedure was outlined by Hosking and Wallis (1993).

1. Estimate the mean at each site, $\hat{\mu}_i$, by the sample mean at site i.

2. Rescale the data, $x'_{ij} = x_{ij}/\hat{\mu}_i$, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, as the basis for estimating $q(F)$. Remember that n_i is the number of years of record at site i and the region consists of n sites.

3. Estimate the parameters separately at each site. Denote the site i estimate of θ_k by $\hat{\theta}_k^{(i)}$.

4. Combine the at-site estimates to give regional estimates:

$$\hat{\theta}_k^{(R)} = \frac{\sum_{i=1}^{n} n_i \hat{\theta}_k^{(i)}}{\sum_{i=1}^{n} n_i}. \tag{3}$$

   Each estimated regional parameter is a weighted average. The site i estimate is given weight proportional to n_i, since for regular statistical models the variance of $\hat{\theta}_k^{(i)}$ is inversely proportional to n_i.

5. Substitute estimates $\hat{\theta}_k^{(1)}, \ldots, \hat{\theta}_k^{(n)}$ into $q(F)$ to give $\hat{q}(F)$, the estimated regional quantile of non-exceedance probability.

6. The site i quantile estimates are obtained by combining the estimates of μ_i and $q(F)$:

$$\hat{Q}_i(F) = \hat{\mu}_i \hat{q}(F). \tag{4}$$

Both the scaled data and the index flood methods are applied to the Waimakariri River AMS data in Section 3.5.

3.2.2 Hierarchical Regional Flood Frequency Approach

Regional flood frequency analysis assumes that all sites in the defined region are homogeneous, that is, all moments (> 1) are assumed to be identical after correction for scale, for each of the n sites in the region. This assumption is highly unlikely to be true, especially when the size of the catchment areas in a region varies. See Stedinger (1983), who showed that CV varies with the size of the drainage area and other basin characteristics.

The more homogeneous a region is, the more accurate the regional approach is. However the reverse is also true: as the heterogeneity among sites increases, the regional approach becomes less accurate. Lettenmaier and Potter (1985) showed that

the performance of index flood methods gets worse as either the regional mean CV, or the site-to-site variation in the CV increases. Homogeneity would be expected to increase as regions are defined to include a smaller number of sites. However the performance of regional estimators also declines as smaller and smaller regions are defined, on account of the increasing variance of parameter estimates. This suggests that a compromise is required. This can be achieved by recognising that different key characteristics of flood behaviour are approximately constant over different spatial scales. By measuring different flood characteristics at different scales we can maximise the benefits of pooling data while minimising the consequences of defining too large a region.

An hierarchical approach to regional flood frequency analysis is one where different key characteristics of flood behaviour are assumed to be approximately constant over different spatial scales. For example, Gabriele and Arnell (1991) developed a hierarchical regional flood frequency estimation procedure which estimates different moments from different, but nested, subsets of data. The higher-order moments are estimated on a regional basis, while the lower-order moments are estimated on a subregional basis. The location of the annual maximum flow is estimated at-site.

The practical value of adopting an hierarchical approach arises because of sampling uncertainties associated with short record lengths. The higher the order of the moment that is to be estimated, the greater the number of observations, and thus the greater the number of sites, we need to record to estimate that moment with the same degree of accuracy. In other words, more samples of a given size are needed to estimate regional skewness to an acceptable level of accuracy than are needed to estimate the regional CV to the same level of accuracy.

Each of the different estimation methods outlined above relies on accurate parameter estimations methods. In flood frequency analysis the current estimation method of choice is the method of L-moments.

3.3 L-moments

L-moments were introduced by Hosking (1990) as expectations of linear combinations of order statistics. L-moments have been widely used in flood frequency

analysis, both overseas (see Stedinger and Lu (1995) for a summary of these investigations) and in New Zealand (Pearson, 1991, 1993; Madsen et al., 1997).

L-moments can be defined for any random variable whose mean exists. They form the basis of a general theory which covers the summarisation and description of theoretical probability distributions and observed data samples, and the estimation of parameters and quantiles of probability distributions. L-moments are analogous to conventional moments. However, a distribution may be specified by its L-moments even if some of its conventional moments do not exist. Such a specification is always unique.

If X is a (real) random variable with cumulative distribution function $F(x)$ and quantile function $x(F)$, and if $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ are the order statistics of a random sample of size n drawn from the distribution of X, then the L-moments of X are defined to be the quantities

$$\lambda_r \equiv r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k:r}, \qquad r = 1, 2, \ldots \tag{5}$$

To standardise the higher L-moments, λ_r, $r \geq 3$, so that they are independent of the units of measurement of X, the L-moment ratios of X are defined as the quantities

$$\tau_r = \lambda_r / \lambda_2, \qquad r = 3, 4, \ldots \tag{6}$$

In particular, λ_1 is the mean of the distribution; λ_2 is a measure of the scale or dispersion; and τ_3 and τ_4 are measures of skewness and kurtosis respectively. The L-CV, $\tau = \lambda_2/\lambda_1$, is analogous to the usual coefficient of variation.

As we have just seen, a common problem in flood frequency analysis is estimating, from a random sample of size n, a probability distribution whose specification involves unknown parameters, $\theta_1, \ldots, \theta_p$. The method of L-moments obtains parameter estimates by equating the first p sample L-moments to the corresponding population quantities, just as the traditional method of moments does. For an ordered sample $x_1 \leq x_2 \leq \ldots \leq x_n$, estimates of the first few L-moments are:

$$l_1 = \sum_{i=1}^{n} x_i / n, \tag{7}$$

$$l_2 = \sum_{i>j} (x_i - x_j) / n(n-1), \tag{8}$$

$$l_3 = \sum_{i>j>k} 2(x_i - 2x_j + x_k) / n(n-1)(n-2). \tag{9}$$

General formulae are given in Hosking (1990). l_1 is the usual sample mean. L-CV and L-skewness are estimated by $t = l_2/l_1$ and $t_3 = l_3/l_2$ respectively. They can be used to judge which distributions are consistent with a given data sample. They can also be used to estimate the parameters when fitting a distribution to the sample, by equating the sample and population L-moments.

L-moments are linear combinations of Probability Weighted Moments (PWMs), which were defined by Greenwood et al. (1979), and used in flood frequency estimation by Landwehr et al. (1979); Greis and Wood (1981); Hosking et al. (1985); Hosking and Wallis (1987); McKerchar and Pearson (1990). Procedures based on PWMs and on L-moments are equivalent. However L-moments are more easily interpretable as measures of distributional shape.

3.4 Candidate Distributions

The choice of the functional form of the underlying flood frequency distribution has a large effect on the flood quantile estimates, especially since the quantiles that interest us are those in the extreme right hand tail of the distribution. Many underlying distributions have been proposed, but none has met with universal approval. The three most common candidates, the Generalised Extreme Value distribution, the Generalised Logistic Distribution and the Lognormal distribution are now introduced.

3.4.1 The Generalised Extreme Value Distribution

The Generalised Extreme Value (GEV) Distribution was introduced by Jenkinson (1955). It combines into a single form the three possible types of limiting distribution for extreme values, as derived by Fisher and Tippett (1928). The GEV is probably the most widely used distribution when measuring AMS of river flow. It has been recommended for this purpose in the UK Flood Studies Report (National Environment Research Council, 1975). A typical application consists of fitting one type of extreme value limiting distribution to the series of annual maxima.

The distribution function is

$$F(x) = \begin{cases} exp\left[-\left\{1 - k\left(\frac{x-\xi}{\alpha}\right)\right\}^{1/k} \right], & k \neq 0, \\ exp\left[-exp\left\{-\left(\frac{x-\xi}{\alpha}\right)\right\} \right], & k = 0, \end{cases} \tag{10}$$

where X is bounded by $(\xi+\alpha/k)$ from above if $k > 0$ and from below if $k < 0$. ξ is the location parameter, α (> 0) is the scale parameter and k is the shape parameter. The shape parameter determines which type of extreme value distribution is represented.

The type I GEV distribution (EV1), also known as the Gumbel distribution, corresponds to $k = 0$. The type II GEV distribution (EV2), also known as the Fréchet distribution, corresponds to $k < 0$. The type III GEV distribution (EV3) corresponds to $k > 0$. Note that if X is assessed to be distributed EV3, then $-X$ has a Weibull distribution. The Weibull distribution is often used in hydrology to analyse extreme low river flows.

The GEV inverse distribution function is

$$x(F) = \begin{cases} \xi + \frac{\alpha}{k}\left\{1 - (-log\ F)^k\right\}, & k \neq 0, \\ \xi - \alpha\ log\ (-log\ F), & k = 0, \end{cases} \tag{11}$$

and the GEV probability density function is

$$f(x) = \begin{cases} \frac{\left(1-\frac{k}{a}(x-\xi)\right)^{1/k} exp\left[-\left(1-\frac{k}{a}(x-\xi)\right)^{1/k} \right]}{\alpha\left(1-\frac{k}{a}(x-\xi)\right)}, & k \neq 0, \\ \alpha^{-1}\ e^{-(x-\xi)/\alpha}\ exp\left[-e^{-(x-\xi)/\alpha}\right], & k = 0. \end{cases} \tag{12}$$

In practice it is usually assessed that k is between -0.5 and 0, so we most often deal with an EV2 distribution. The EV2 distribution has expectation

$$E(X) = \xi + \frac{\alpha}{k}\left(1 - \Gamma(1+k)\right) \tag{13}$$

and variance

$$V(X) = \left(\frac{\alpha}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right). \tag{14}$$

Although the EV2 distribution is bounded below by $\xi + \alpha/k$, of course we cannot actually observe a negative flow. In practice this sub-zero lower bound is rarely a problem. For example, of the ten rivers we analyse in the next Section, the most mass that any of their estimated density functions places on negative values of X is less than 10^{-3}.

Hosking (1990) used L-moments to show that point estimates of the GEV distribution can be obtained using:

$$z = \frac{2}{(3+t_3)} - \frac{log\ 2}{log\ 3}, \tag{15}$$

$$\hat{k} \approx 7.8590z + 2.5994z^2, \tag{16}$$

$$\hat{\alpha} = \frac{l_2\hat{k}}{\left(1-2^{-\hat{k}}\right)\Gamma\left(1+\hat{k}\right)}, \tag{17}$$

$$\hat{\xi} = l_1 + \frac{\hat{\alpha}\left\{\Gamma\left(1+\hat{k}\right)-1\right\}}{\hat{k}}. \tag{18}$$

Remember that l_1 is the sample mean, l_2 is a measure of scale and t_3 is a measure of skewness.

3.4.2 The Generalised Logistic Distribution

The distribution function for the three-parameter Generalised Logistic distribution (GLO) is

$$F(x) = \left[1 + \left(1 - \frac{k}{\alpha}(x-\xi)\right)^{1/k}\right]^{-1}. \tag{19}$$

As with the GEV distribution, ξ is the location parameter, α (> 0) is the scale parameter and k is the shape parameter. When $k = 0$ the GLO reduces to the two-parameter Logistic distribution.

The GLO inverse distribution function is

$$x(F) = \xi + \frac{\alpha}{k}\left(1 - \left((1-F)/F\right)^k\right), \tag{20}$$

and the GLO probability density function is

$$f(x) = \frac{\left(1-\frac{k}{a}(x-\xi)\right)^{1/k-1}}{\alpha\left[1+\left(1-\frac{k}{a}(x-\xi)\right)^{1/k}\right]^2}. \tag{21}$$

Hosking (1990) showed that point estimates of the parameters of the GLO distribution can be obtained via L-moments by:

$$\hat{k} = -t_3, \tag{22}$$

$$\hat{\alpha} = \frac{l_2}{\Gamma(1+\hat{k})\Gamma(1-\hat{k})}, \tag{23}$$

$$\hat{\xi} = l_1 + \frac{l_2 - \hat{\alpha}}{\hat{k}}. \tag{24}$$

3.4.3 The Lognormal Distribution

The distribution of X is said to be Lognormal if $Z = log(X - \xi)$ is Normally distributed. The distribution function for the three-parameter Lognormal distribution (LN3) is

$$F(x) = \Phi\left(\frac{log\,(x-\xi) - \mu}{\sigma}\right), \tag{25}$$

where $x > \xi$ and Φ is the standard Normal distribution function. The expected value and standard deviation of $Z = log(X - \xi)$ are denoted by μ and σ respectively. Any change in the value of ξ affects only the location of the distribution. When $\xi = 0$, Equation 25 reduces to the two-parameter Lognormal distribution.

The LN3 inverse distribution function is

$$x(F) = \xi + exp\left[\mu + \sigma\Phi^{-1}\,(F)\right]. \tag{26}$$

The LN3 probability density function is

$$f(x) = \left((x-\xi)\sqrt{2\pi}\sigma\right)^{-1} exp\left[-\frac{1}{2}\frac{(log\,(x-\xi)-\mu)^2}{\sigma^2}\right]. \tag{27}$$

At-site point estimates of the parameters of the LN3 distribution are given by Hosking (1990) as:

$$z = \sqrt{(8/3)}\Phi^{-1}\left(\frac{1+t_3}{2}\right), \tag{28}$$

$$\hat{\sigma} \approx 0.999281z - 0.006118z^3 + 0.000127z^5, \tag{29}$$

$$\hat{\mu} = log\left(\frac{l_2}{erf(\sigma/2)}\right) - \frac{\hat{\sigma}^2}{2}, \tag{30}$$

$$\hat{\xi} = l_1 - exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right), \tag{31}$$

where *erf* is the error function.

Now that we have been introduced to the theory of L-moments and our three candidate distributions, we are ready to estimate flood exceedance quantile levels for the Waimakariri River.

3.5 Frequentist Estimates of Exceedance Quantiles for the Waimakariri River

The flood distribution of the Waimakariri River has previously been studied by McKerchar and Pearson (1990), Pearson (1993) and Connell and Pearson (2001).

McKerchar and Pearson (1990) used PWMs to test if the shape parameter of the GEV distribution was equal to zero at each of 275 New Zealand river locations. Pearson (1993) re-investigated the same problem using L-moments, and concluded that Canterbury rivers have a parent EV2 distribution. Connell and Pearson (2001) applied the Two-Component Extreme Value distribution (a distribution of the maxima of two independent Gumbel distributions) to AMS data from East Coast rivers. They concluded that the rivers could be split into three homogeneous groups: Main Divide rivers (of which the Waimakariri is one), Northern East Coast rivers (includes rivers from the Ashley to the Rangitata) and Southern East Coast rivers (includes rivers from the Orari to the Hakataramea).

3.5.1 At-site Estimates of Exceedance Quantiles

The simplest method of quantile estimation involves the researchers selecting a distribution they feel adequately represents the underlying flood frequency distribution, and estimating the characterising parameters of the distribution from the AMS recorded at the site of interest. Point estimates of parameters, and estimates of exceedance quantiles, $Q(0.95)$, $Q(0.98)$, $Q(0.99)$, $Q(0.995)$ and $Q(0.999)$, are found for each of the three candidate distributions.

3.5.2 Regional Estimates of Exceedance Quantiles

Regional estimates of exceedance quantiles are calculated by pooling data from a number of different, but related, sites. This has the advantage of increasing the number of recorded AMS values. However, the more heterogeneous a region, the less effective data pooling, meaning that we must make a decision between increasing the sample size and increasing the heterogeneity.

It has been shown (Mosley, 1981; McKerchar and Pearson, 1990; Pearson, 1991) that rivers draining on the East Coast of the South Island form reasonably homogeneous flood frequency regions. Mosley (1981) achieved this through cluster analysis, McKerchar and Pearson (1990) by fitting GEV curves to 275 AMS data records and Pearson (1991) identified homogeneous regions by considering the similarity between the L-skewness and L-kurtosis at different sites. The sites considered to be part of the same region as the Waimakariri River are all rivers from the Canterbury

Site	River	n	area	mean	std dev.	CV	skew
60110	Waihopai	23	764	425	178	0.42	0.69
62103	Acheron	41	973	333	179	0.54	1.92
62105	Clarence	43	440	193	89	0.46	0.67
65104	Hurunui	45	1060	531	218	0.41	0.50
66204	Ashley	30	472	320	222	0.70	1.19
66401	Waimakariri	72	3210	1485	652	0.44	1.77
68001	Selwyn	38	164	77	66	0.85	2.13
68526	Rakaia	44	2560	2419	929	0.38	1.66
68806	South Ashburton	35	539	102	66	0.64	1.55
69302	Rangitata	43	1461	1357	737	0.54	1.09

Table 2: Rivers used in regional analysis. Record length is measured in years. Mean and standard deviation are measured in cumecs. Area is measured in squared kilometres.

region. The rivers are listed in Table 2. See Walter (2000) for more detail on the site records.

We shall consider the three different regionalisation techniques introduced in Section 3.2: a standard scaled data approach, an index flood approach, and an hierarchical analysis.

Scaled Data Regional Estimates of Exceedance Quantiles

First, we shall undertake a simple process which merely involves scaling each observation so that each observation is part of a 'super-site'. Next, we estimate the parameters of the rescaled data. Finally, we construct and rescale the appropriate density curve. The steps involved in this process are:

1. Rescale each observation by its at-site mean, $x^{'} = x_{ij}/\hat{\mu}_i$, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$. The index i is the site number, and n_i is the number of years of AMS recordings at that site. Each site may be measured over any number of years. The $x_{ij}^{'}$ now form a super-site of size $\sum_{i=1}^{N} n_i$.

2. Estimate parameters of super-site using the method of L-moments.

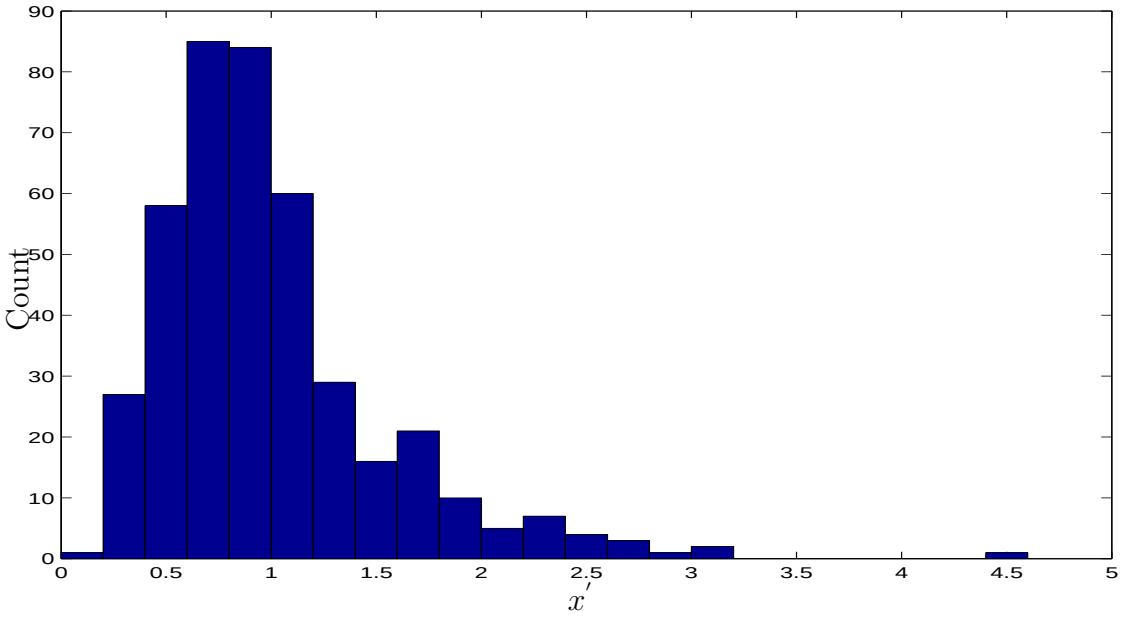

Figure 2: Normalised AMS for Canterbury rivers.

3. Form density function and rescale it by $\hat{\mu}_i$.

This process was undertaken using each of our candidate distributions. A total of 414 normalised observations were obtained, with values ranging from 0.13 to 4.46. Normalised values from the Waimakariri river ranged from 0.48 to 2.69. Figure 2 displays the shape of the normalised observations. Note that the histograms displayed in Figure 2 and in the lower panel of Figure 1 are approximately the same shape.

Index Flood Regional Estimates of Exceedance Quantities

We undertake an index flood procedure as outlined in Section 3.2.1. First, scale each site's data by its mean and estimate the characterising L-moments at each site. Using these, estimate the regional L-moments and thus calculate the regional normalised flood frequency distribution. Finally, rescale the distribution and compute the quantiles of interest. This procedure was undertaken for the Waimakariri data.

Hierarchical Regional Estimates of Exceedance Quantities

Hierarchical regional estimation methods were outlined in Section 3.2.2. A typical hierarchical procedure is to calculate the shape parameter, which controls skewness,

from all the sites in a region while calculating the scale and location parameters using at-site data. In practice this involves normalising and pooling the data from all sites in a region. A regional estimate of the shape parameter is calculated from this pooled data. Using the regional estimate of the shape parameter we estimate the scale and location parameters using the at-site data.

3.6 Results

3.6.1 At-site Results

It is a simple matter to estimate GEV parameters for the Waimakariri River using Equations 12–15. Parameter estimates of the at-site frequency distribution are $\hat{k} = -0.25$, $\hat{\alpha} = 355$ and $\hat{\xi} = 1165$. Using these estimates we can plot the shape of the underlying flood frequency density. This is shown as the blue curve in Figure 3. These parameter estimates can also be used to estimate exceedance quantiles using the inverse GEV distribution function. The parameter and quantile estimates obtained for all three candidate distributions and shown in Table 3 in the row headed "AS". Note that the parameter estimates displayed have been normalised, so they can be easily compared with the parameter estimates obtained through regional procedures. In particular, observe that the estimate of ξ given in Table 3 is 0.78, and the at-site mean, from Table 2, is $1485m^3/sec$. Multiplying these two values, and discounting rounding errors, gives $\hat{\xi} = 1165$, the at-site estimate. $\hat{Q(0.95)}$ is $2729m^3/sec$. This is interpreted to mean that, in any year, we expect there is a 95% chance that the maximum flow will not exceed $2729m^3/sec$.

The at-site GLO parameter estimates for the Waimakariri River are $\hat{k} = -0.34$, $\hat{\alpha} = 215$ and $\hat{\xi} = 1149$. These parameters lead to the density curve plotted in red in Figure 3, and to the exceedance quantiles, measured in cumecs, displayed in Table 3.

The at-site LN3 parameter estimates for the Waimakariri River are $\hat{\sigma} = 0.72$, $\hat{\mu} = 6.49$ and $\hat{\xi} = 636$. The estimated LN3 density curve is shown in green in Figure 3, and estimated exceedance quantiles are displayed in Table 3.

Figure 3 shows the estimated GEV, GLO and LN3 densities. Clearly the GLO density is quite different from the other two. Although the GEV and LN3 densities have different modes, they are very similar in the upper tail, where the quantities we

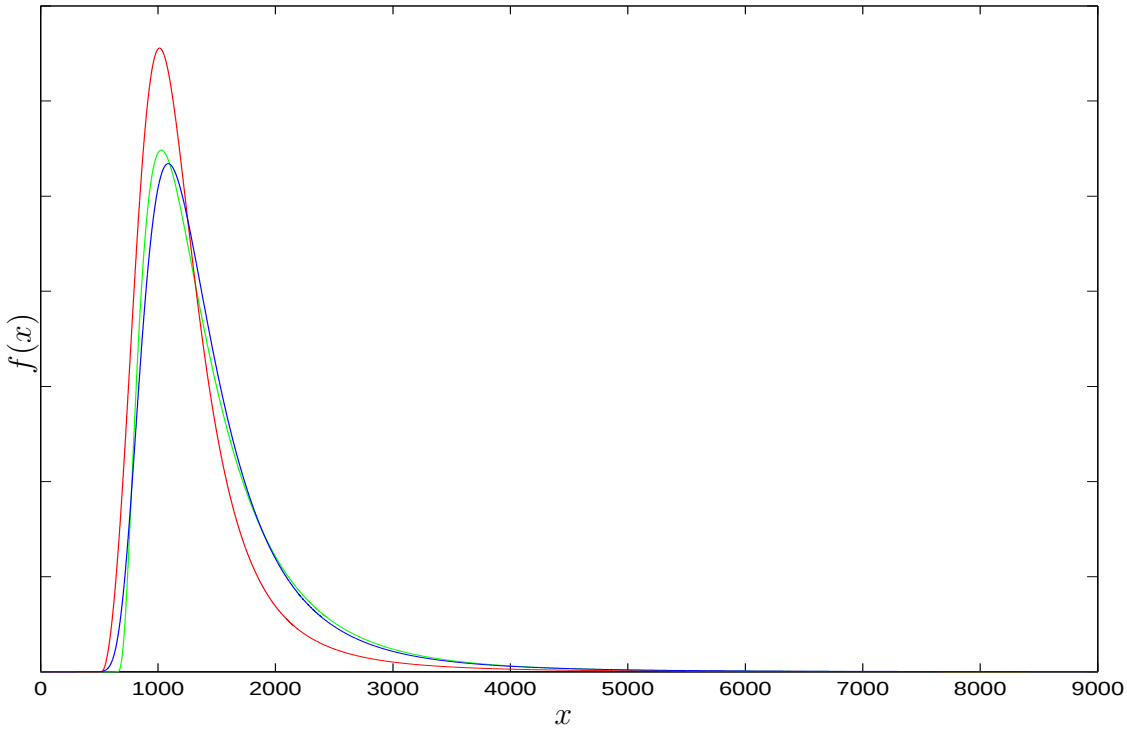

Figure 3: At-site density functions estimated using Waimakariri River AMS. Densities are GEV (blue), GLO (red) and LN3 (green).

are most interested in reside. Figure 4, a plot which concentrates on the upper end of the distribution curve for our three candidate distributions, demonstrates this. For example consider the point $Q(0.95)$. Find the point 0.95 on the $F(x)$ axis and look to the right. The first distribution we encounter, at $x = 2238$, is the GLO, represented by the red line. Continuing to the right we cross the blue GEV line at $x = 2729$, and then almost immediately cross the green LN3 line at $x = 2771$. There is a difference of $533m^3/sec$ between the smallest and largest estimates of $Q(0.95)$ over the three candidate distributions. This illustrates the importance of assuming an appropriate distribution for the AMS. Figure 4 and Table 3 show a similar situation exists for higher quantile estimates. The LN3 and GEV exceedance estimates are more similar than the GLO. For a fixed return period exceedance estimates based on the GLO are lowest for $F(x) < 0.998$. The GEV and LN3 estimates are the same at $F(x) \approx 0.975$, below this quantile the GEV is lower, above it the LN3 estimate is the smaller of the two.

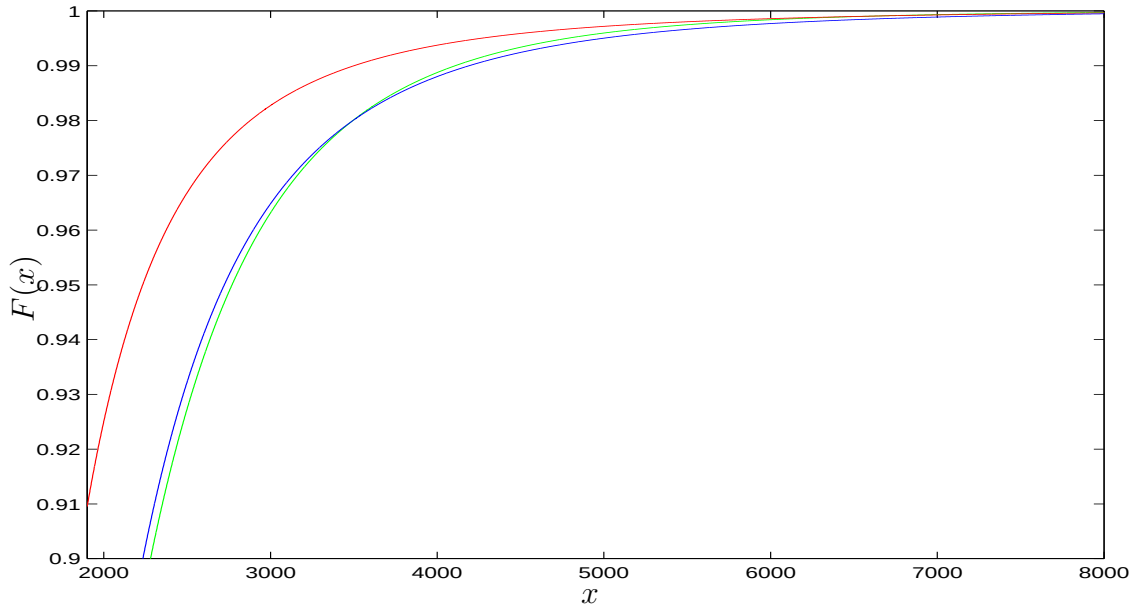

Figure 4: At-site cumulative density functions estimated using Waimakariri River AMS. Distributions are GEV (blue), GLO (red) and LN3 (green).

3.6.2 Regional Results

The parameter and quantile estimates obtained for all three candidate distributions are listed in Table 3. Scaled data estimates are listed in the row headed "SD", index flood estimates are listed in the row headed "IF" and hierarchical estimates are listed in the row headed "Hi". Remember that the parameter estimates are based on the normalised regional frequency curve. For each of the three candidate distributions, the three different regional estimates are closer to each other than any of them are to the at-site estimate. In particular there is a considerable difference between the shape parameter estimated from at-site and regional data. For every estimation method the GEV and LN3 quantile estimates are closer to each other than to the GLO estimates, suggesting that the regional estimation methods produce equivalent differences between the distributions, as the at-site methods do.

3.7 Comparison of Approximation Methods

To this point we have estimated five flood exceedance quantiles for the Waimakariri River, using four estimation methods. How can we judge the worth of these esti-

		$\hat{\xi}$	$\hat{\alpha}$	$\hat{k}$	$\hat{Q(0.95)}$	$\hat{Q(0.98)}$	$\hat{Q(0.99)}$	$\hat{Q(0.995)}$	$\hat{Q(0.999)}$
GEV	AS	0.78	0.24	-0.25	2729	3512	4232	5086	7741
	SD	0.75	0.35	-0.13	2982	3728	4348	5022	6824
	IF	0.75	0.34	-0.16	2993	3789	4465	5214	7278
	Hi	0.79	0.27	-0.16	2699	3340	3885	4490	6167
GLO	AS	0.77	0.14	-0.34	2238	2893	3536	4347	7153
	SD	0.71	0.20	-0.25	2379	3059	3680	4412	6647
	IF	0.72	0.20	-0.27	2387	3095	3750	4534	6957
	Hi	0.77	0.16	-0.27	2216	2787	3318	3954	5953
		$\hat{\xi}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{Q(0.95)}$	$\hat{Q(0.98)}$	$\hat{Q(0.99)}$	$\hat{Q(0.995)}$	$\hat{Q(0.999)}$
LN3	AS	0.43	-0.81	0.72	2771	3499	4117	4800	6659
	SD	0.05	-0.19	0.53	3008	3711	4271	4859	6344
	IF	0.12	-0.29	0.57	3004	3686	4307	4973	6633
	Hi	0.29	-0.50	0.57	2717	3316	3801	4316	5638

Table 3: Parameter and quantile estimates for the Waimakariri River for the three candidate distributions. The letters in column 1 refer to: AS = at-site; SD = scaled data; IF = index flood; Hi = hierarchical. Remember that the parameter estimates are for the (normalised) regional frequency distribution. Exceedance quantiles are estimated in cumecs.

mates? Conventional measures of the adequacy of a specified distribution are the bias and the root-mean-squared error (RMSE). Bias is a measure of how closely the expected value of an estimate is to the parameter that it is supposed to estimate. A statistic, $T = T(X_1, \ldots, X_n)$, is said to be an unbiased estimator of the parameter θ if $E(T) = \theta$ for all θ. If random estimator T is unbiased it possesses a distribution whose mean is the parameter θ being estimated.

Unbiasedness alone is not enough on which to base a choice of method, as more than one statistic can be unbiased. If a number of statistics are unbiased we seek to find the one with the minimum variance — this is called the best unbiased estimator. If T is not an unbiased estimator of parameter θ, we judge its merits on the basis of the mean-squared error, defined as $E\left[(t-\theta)^2\right]$, rather than on $V(t)$. It is well

$M(CV)$	$R^*(CV)$	Site 1			Site 11			Site 21		
		λ_1	λ_2	k	λ_1	λ_2	k	λ_1	λ_2	k
0.5	0.3	2	1.15	-0.17	2	1	-0.14	2	0.85	-0.11
0.5	0.5	2	1.25	-0.17	2	1	-0.14	2	0.75	-0.11
1.0	0.3	1	1.15	-0.17	1	1	-0.14	1	0.85	-0.11
1.0	0.5	1	1.25	-0.17	1	1	-0.14	1	0.75	-0.11

Table 4: Summary of regions used in Monte Carlo experiments.

known that biased estimators can produced lower mean-squared error than unbiased ones.

Since the true form of the underlying distribution of floods is unknown and unobservable, we use a Monte Carlo approach both to generate our own sequence of AMS data and to assess competing estimation procedures. The Monte Carlo procedure assumes that the underlying flood distribution exists and is known. With this Monte Carlo approach we can estimate both the accuracy (variance) and precision (bias) of the quantile estimates.

3.7.1 Data Generation

The Monte Carlo procedure consists of two primary parts. First we generate the data. Then we test the different methods. When generating the data we attempt to produce simulated series that are plausible representations of the real life flood process. Data is simulated from a GEV distribution. A study of the form of Equation 11 and Equation 24 shows that the two most important measures to represent accurately are the measures of spread and skewness. In both cases the spread and skewness determine the shape of the distribution and the location term only acts to translate the distribution along the x-axis.

A region consisting of 21 sites was considered. The region's statistics are summarised in Table 4. Population skewness, record length and CV varied by site. Record lengths ranged from 10 years at site 1, to 30 years at site 21, increasing by 1 year per site. Population skewness ranged linearly from -0.17 at site 1 to -0.11 at site 21. The population skewness was specified to be greatest at the sites with

the shortest record lengths, because small catchment areas, which are associated with high at-site estimates of skewness, tend to have been gauged for a shorter time period than bigger catchments.

The distribution of CV over the sites reflects the degree of heterogeneity of the sites within the region. Remember that one of the assumptions of the index flood method is that sites are homogeneous over a region. Thus, as heterogeneity increases, we expect the estimates to be both increasingly inaccurate and have higher variance. CV is defined in terms of the regional median, denoted $M(CV)$, and the range of the CV within a region, denoted $R(CV)$. The regional range is normalised as $R^*(CV) = R(CV)/M(CV)$.

Two different values of $M(CV)$, 0.5 and 1.0, as well as two different values of $R^*(CV)$, 0.3 and 0.5, are considered. These values were selected to mimic the suspected CV values of the region containing the Waimakariri River. If a researcher asserts a GEV distribution to summarise their uncertainty about a sequence of AMS values, it is inevitable that the GEV distribution will have positive mass for $x < 0$, especially for the $(M(CV), R^*(CV))$ combinations under consideration. For example, site 11 of a region with $M(CV) = 0.5$ has approximately an 8% chance of generating a negative value, while site 11 of a region with $M(CV) = 1$ has approximately a 32% chance. Clearly it is impossible to observe a negative river flow. Thus the distribution our simulated regions were generated from was a truncated GEV distribution.

Simulations are run for each of the four $(M(CV), R^*(CV))$ combinations. For each combination the population CV at site 1 was set to $M(CV)(1 + R^*(CV)/2)$ and parameters α_1 and ξ_1 were determined. The population CV at site 11 was set to $M(CV)$ and parameters α_{11} and ξ_{11} determined. Similarly the population CV at site 21 was set to $M(CV)(1 - R^*(CV)/2)$ and parameters α_{21} and ξ_{21} were determined. The population parameters for the remaining sites were found by linearly interpolating between these three sites. For this experiment 50,000 samples were generated for each $(M(CV), R^*(CV))$ combination. This method of simulating a series of AMS measurements is based on the method implemented by Lettenmaier et al. (1987). The values of $M(CV)$ used in their simulations ranged from 0.5 to 2, while values of $R^*(CV)$ ranged from 0.2 to 0.5. However neither the λ_1 or λ_2 values,

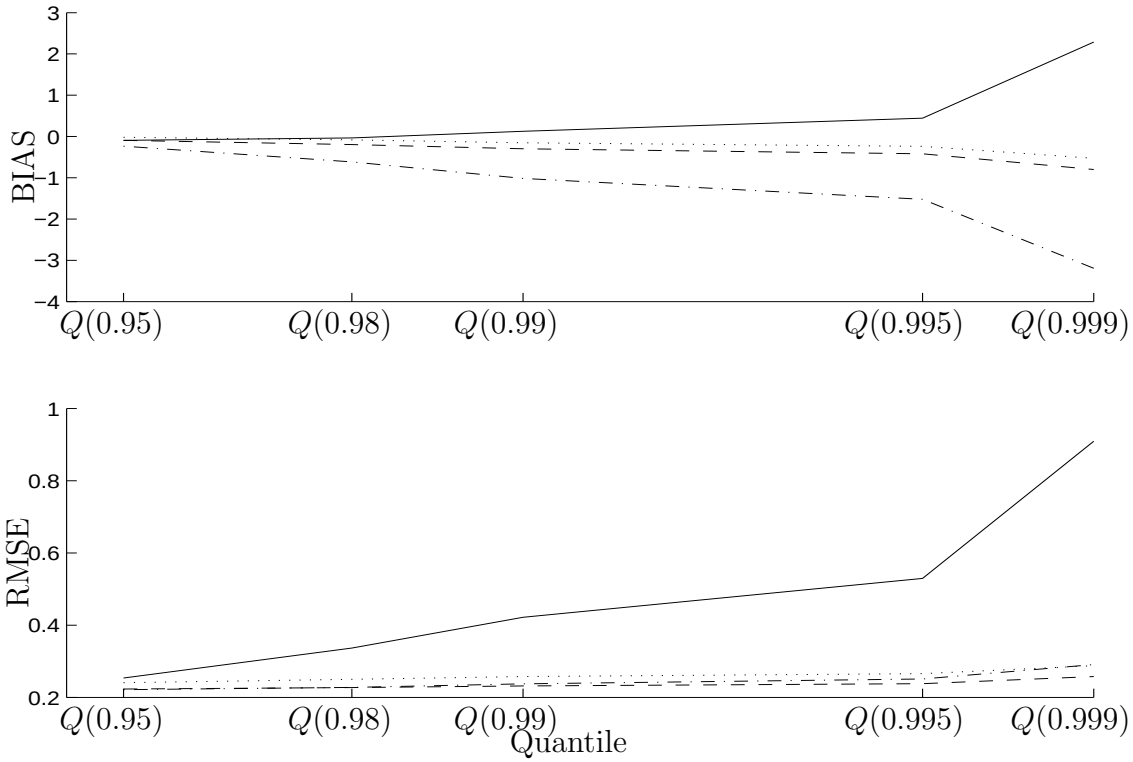

Figure 5: Root-mean-squared error and bias for site 11 of a 21 site region when $M(CV) = 0.5$ and $R^*(CV) = 0.3$. Types of estimation methods are: at-site (—), scaled data (-.-.), index flood (- -) and hierarchical (...).

where $CV = \lambda_2/\lambda_1$, were specified.

3.7.2 Results

Once each of the four data sets has been generated, we use each of the four estimation methods (at-site, regional scaled data, regional index flood, hierarchical) to estimate parameters, and hence return periods, of the (known) underlying flood frequency distribution. The data used for each $(M(CV), R^*(CV))$ combination was generated once, and the different estimation methods were applied to the same data set.

The methods are compared by estimating biases and normalised root-mean-squared errors at each site and for each estimation method. Bias was estimated as

$$\frac{1}{n}\sum_{p=1}^{n}\hat{x}_{pqr} - x_{Tq}, \tag{32}$$

where p is the Monte Carlo simulation index, q is the site number, r is the estimation method and x_{Tq} is the true flood quantile at site q. Normalised root-mean-squared

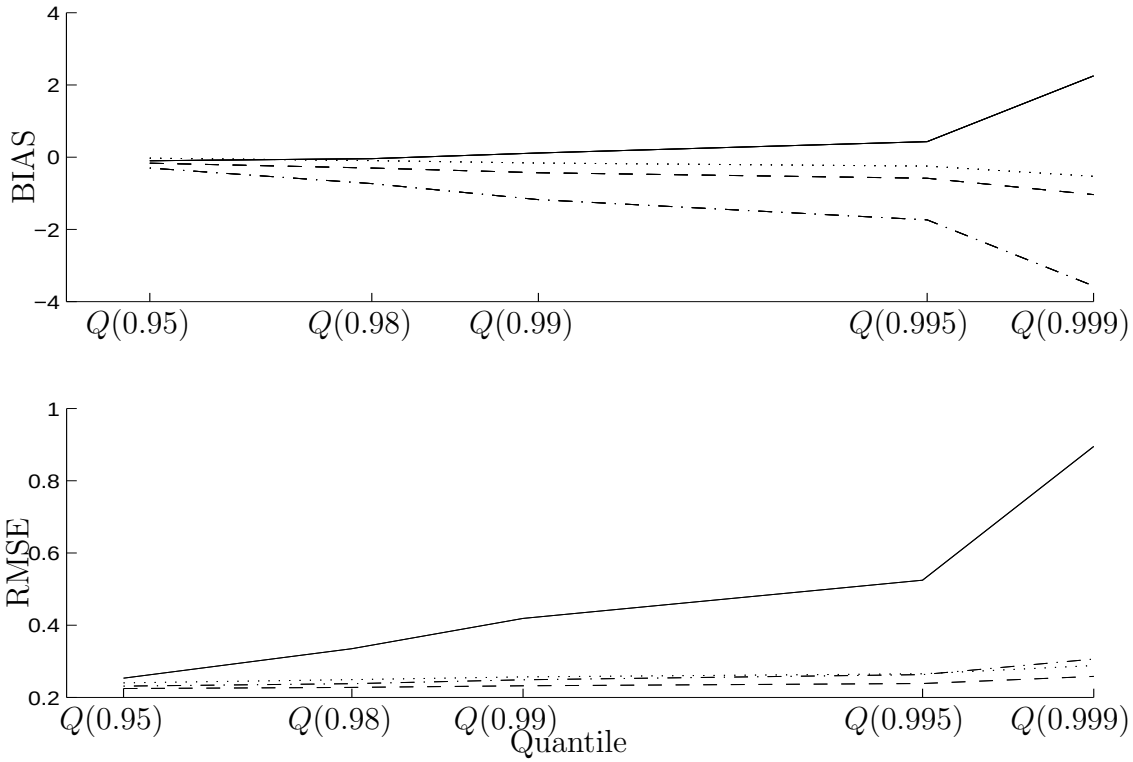

Figure 6: Root-mean-squared error and bias for site 11 of a 21 site region when $M(CV) = 0.5$ and $R^*(CV) = 0.5$. Types of estimation methods are: at-site (—), scaled data (-.-.), index flood (- -) and hierarchical (...).

error was estimated as

$$\frac{\left[n^{-1}\sum_{p=1}^{n}\left(\hat{x}_{Tpqr} - x_{Tq}\right)^2\right]^{1/2}}{x_{Tq}}. \tag{33}$$

Figures 5 to 8 show the estimated bias and RMSE, as a function of quantile level, for the 11th site in our 21 site region. Four fitting methods were used: at-site, scaled data, index flood and hierarchical. The median CV is either 0.5 or 1 and the range of regional CV is 0.3 or 0.5. It is clear from the lower panel of each figure that the RMSE of the at-site estimator is much larger than any of the three regional estimators, which are all relatively similar.

The at-site quantile estimates are biased upwards in every case studied. The three regional estimates are all biased downward. Of the three regional estimates the scaled data estimate is the most biased for every combination except $(1, 0.3)$, when the hierarchical is slightly worse. The index flood estimates are consistently the least biased. The hierarchical estimate is the least biased estimate when the median CV is low, but as CV increases it rapidly becomes biased, especially in the

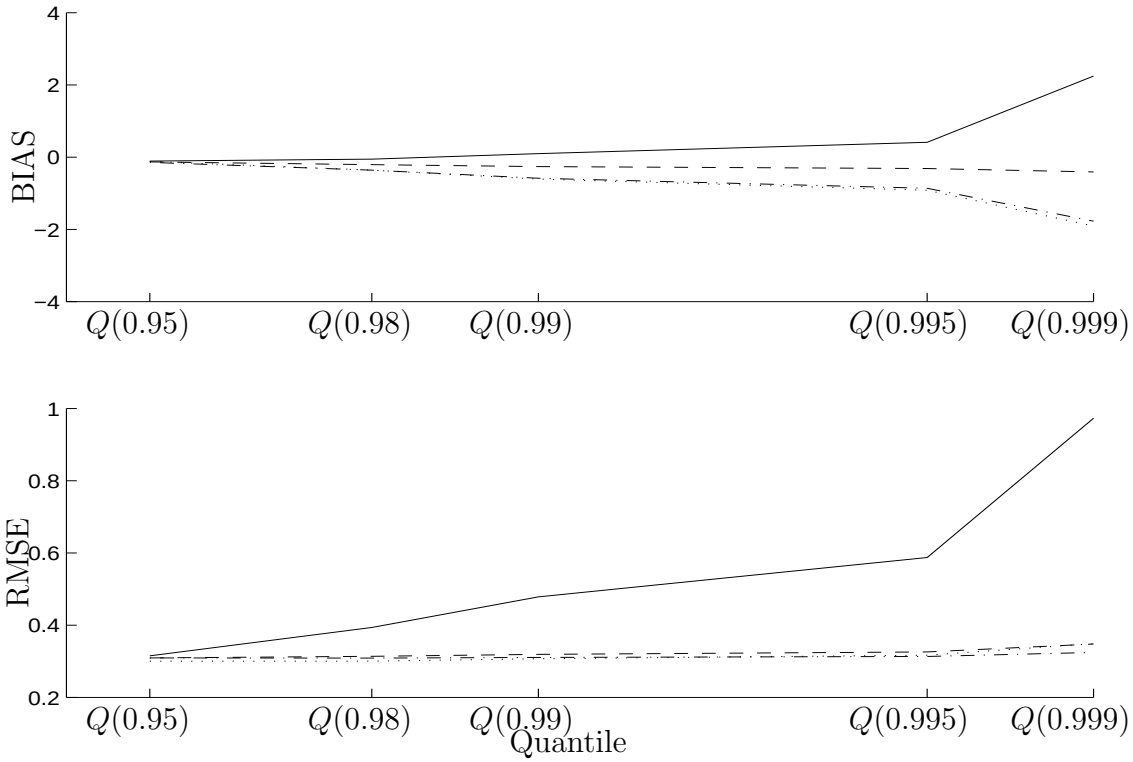

Figure 7: Root-mean-squared error and bias for site 11 of a 21 site region when $M(CV) = 1$ and $R^*(CV) = 0.3$. Types of Estimation methods are: at-site (—), scaled data (-.-.), index flood (- -) and hierarchical (...).

extreme tail (when return period > 500).

These results are equivalent to those found by Lettenmaier et al. (1987), who compared a number of at-site and regional estimators all of which used one of the three types of GEV distribution. They found that the GEV distribution gave excessively variable flood quantile estimates when it was used for evaluating quantiles at-site. However, when it was incorporated into a regional estimation scheme it was relatively insensitive to modest regional heterogeneity in the CV. The higher the value of the regional mean coefficient of variation, $M(CV)$, the more the advantage of methods that assumed regional homogeneity declined.

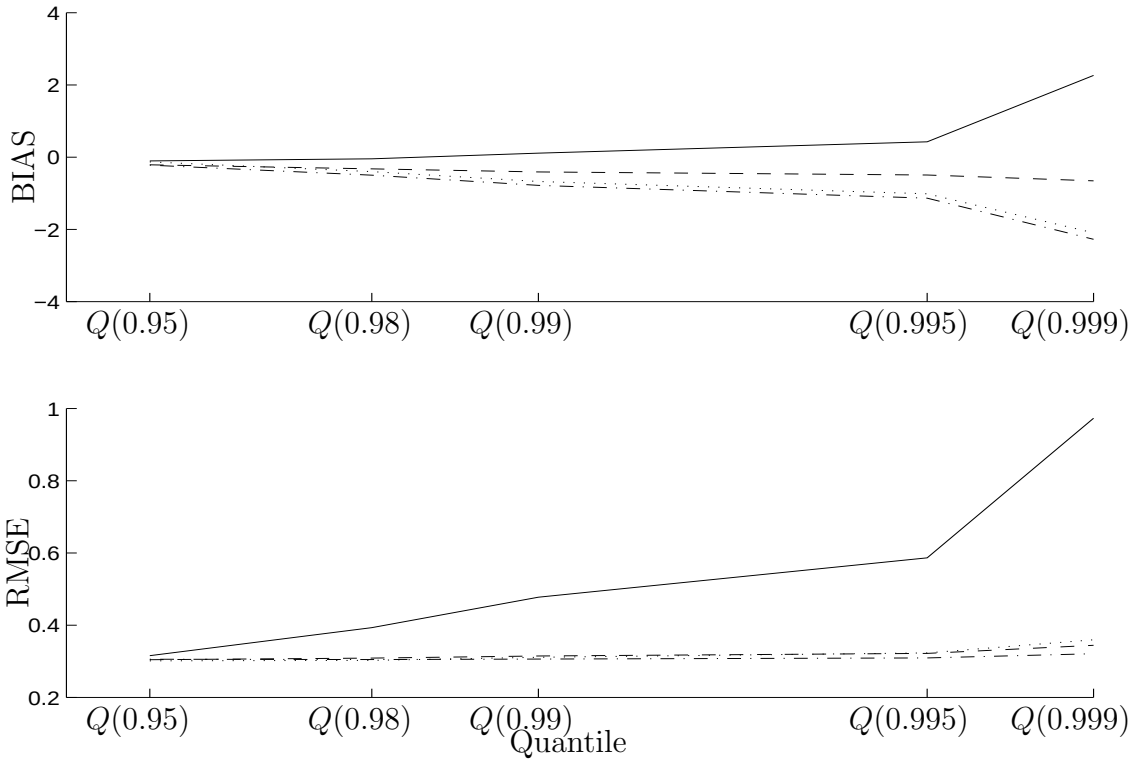

Figure 8: Root-mean-squared error and bias for site 11 of a 21 site region when $M(CV) = 1$ and $R^*(CV) = 0.5$. Types of estimation methods are: at-site (—), scaled data (-.-.), index flood (- -) and hierarchical (...).

4 Assessing Flood Exceedance Quantiles Using Updated Mixture Mass Functions as Sequential Forecasting Distributions

In Ware and Lad (2003) we described in detail a procedure for forecasting the value of various items when the analysis involves sequences of observations that the researcher regards exchangeably, where particular interest centres on the sequence of updated probability mass functions $f(x_{i+1} \mid \mathbf{X_i} = \mathbf{x_i})$. Although the procedure studied dealt specifically with a digitised Normal-Gamma mixture Normal distribution, the updating process can be extended to encompass any number of parameters and to account for any functional form the researcher wishes to specify for the prior distribution and the ITF.

4.1 Introduction

When studying extreme geologic phenomena the items in which we are most interested are exceedance quantiles. This Section concerns the construction and implementation of a procedure for sequentially updating mixture mass functions. These mass functions are used to forecast exceedance quantiles for the Waimakariri River AMS data. We shall specify densities as approximations of our uncertain knowledge, and then construct exact mass functions by digitising these densities. Before moving on to the mechanics of the updating procedure itself, we consider the functional form of the density used to assess uncertain knowledge of AMS.

4.1.1 The Functional Form Of the Information Transfer Function

As discussed previously in Section 3.4, there is no consensus about which distribution, if any, can best be used to represent knowledge about the components of the AMS. In Section 3.4, three common choices of distribution to represent the form of AMS data were introduced: the Generalised Extreme Value distribution, the Generalised Logistic distribution and the Lognormal distribution. Any one of these three distributions could be digitised and specified as our ITF — although of course our choice of ITF is not limited to one of these three distributions.

After deciding on the form of the ITF we implement the digital updating procedure and obtain as many conditional quantile estimates as desired. The question is then: How can we compare results obtained using ITFs of different functional forms? In particular, does one ITF give 'better' results than the other two? One way we can compare the candidate distributions is by using scoring rules to evaluate our assessment of the form of ITF. We now briefly detour from the task at hand to introduce the notion of scoring rules.

4.2 Proper Scoring Rules

Scoring rules are used to evaluate states of uncertain knowledge. Scoring rules are comprehensively covered by Lad (1996, Chapter 6). Scoring rules attach a numerical score to any assertion, $K(X)$, about an uncertain but observable quantity, X, once that quantity is observed formally. A scoring rule is a function that assigns a real

valued number to each possible $(X, K(X))$ combination, where $K(X)$ represents the assertion value. The value of the real number is called the score. A scoring rule is defined so that it achieves its maximum when $K(X) = x$, and is non-increasing as x departs from $K(X)$ for each $K(X)$, and as $K(X)$ departs from x for each x.

One desirable feature of a scoring rule is that it should reward researchers for accurately and honestly assessing $K(X)$. A proper scoring rule is one for which the researcher's prevision for the numerical score $S(X, K(X))$ is greatest when they assert $K(X)$ as their assertion of knowledge about X. Clearly, under a proper scoring rule, it is to the researcher's advantage to honestly specify $K(X)$. There are a number of well-known types of proper scoring rules, which can be divided into two main groups: proper scoring rules for previsions and proper scoring rules for distributions.

4.2.1 Proper Scoring Rules for Distributions

If we assert a probability mass function for a quantity with realm $\mathbf{R}(X) = \{x_1, x_2 \ldots, x_N\}$, then our specification of knowledge about X is denoted by $\mathbf{K}_N(X) = \mathbf{p}_N$, where $\mathbf{p}_N$ represents our assertions of probabilities for the constituents of the partition generated by X. There are a multitude of proper scoring rules for assessing the probability mass function on the basis of observing X. In this subsection we shall consider three of them.

The Quadratic Score of a Distribution

If X is a quantity with realm $\mathbf{R}(X) = \{x_1, x_2 \ldots, x_N\}$, the quadratic score of a distribution is defined as

$$S\left(X, \mathbf{p}_N\right) \equiv -\left[\sum_{i=1}^{N} \left(X = x_i\right)\left(1 - p_i\right)^2 + \sum_{i=1}^{N} \left(X \neq x_i\right) p_i^2\right], \tag{34}$$

where $\mathbf{p}_N = (p_1, p_2, \ldots, p_N)$ is the vector of probabilities defining the asserted mass function $\mathbf{p}_N$ over the realm values.

Note that the quadratic score of a distribution is the sum of the quadratic scores attained by each of the constituents of the distribution, since $(X = x_i) = 1$ for only one element of the realm, and is zero for the other $(N - 1)$ elements. The largest value the score of a quadratic distribution can attain is 0. This is achieved when the

forecaster is sure of the exact value of X, asserting the degenerate distribution that associates probability 1 with the event $(X = x_i)$ that occurs. The worst value of a quadratic score for a distribution is -2, when a distribution specifies $P(X = x_i) = 1$ and x_i does not occur. All other distributions achieve scores between 0 and -2. The quadratic score is sometimes called the Brier Score of the distribution.

Note that quadratic scores are defined so that they are always non-positive. It is always better to have a score closer to 0 than it is to have one farther away from 0. Note also that the quadratic score of prevision K, $S(X, K) = -(X - K)^2$, scores the difference between the asserted and observed values of X. Whereas the quadratic score of a distribution scores the difference between the asserted and observed probabilities. Thus any two distributions that have the same assessed probabilities, in any permutation, will have the same score as long as the assert the same probability for x^*, where $X = x^*$ is observed.

The Logarithmic Score of a Distribution

If X is a quantity with realm $\mathbf{R}(X) = \{x_1, x_2 \ldots, x_N\}$, the logarithmic score of a distribution is defined as

$$S\left(X, \mathbf{p}_N\right) \equiv \sum_{i=1}^{N} \left(X = x_i\right) log\left(p_i\right). \tag{35}$$

The logarithmic score of a distribution is merely the logarithm of your prevision for the event that equals one. Thus two distributions will have the same score as long as they each specify the same probability for the constituent of X that does occur — regardless of how the assessed distributions differ over the remaining $(N - 1)$ possibilities. The logarithmic scoring rule is particularly appealing for researchers who believe that the observation gives no information about the other possible values of X that did not occur — no matter how near or far these other possibilities are from x^*. All logarithmic scores of distribution are non-positive. The closer the score is to 0 the better.

The Spherical Score of a Distribution

If X is a quantity with realm $\mathbf{R}(X) = \{x_1, x_2 \ldots, x_N\}$, the spherical score of a distribution is defined as

$$S\left(X, \mathbf{p}_N\right) \equiv \frac{\sum_{i=1}^{N} \left(X = x_i\right) p_i}{\left(\sum_{i=1}^{N} p_i^2\right)^{1/2}}. \tag{36}$$

The spherical score of a distribution is the expectation placed on x^* divided by the square root of the sum of the squared probabilities for each constituent of X. All spherical scores are non-negative. The minimum achievable score is 0, when a degenerate distribution is specified on a value of X that does not occur. The maximum achievable score is 1, when a degenerate distribution is specified on the value of X that does occur. When considering the spherical score of a distribution, the higher the score the better.

Scoring rules are a measure of the value of information asserted about $X \in \mathbf{R}(X)$ contained in $\mathbf{p}_N$. In general, the scoring rule we choose should be based on how severely we want to penalise distributions that place substantial probabilities on possible values of X that do not occur, depending on their distance from the distribution that is degenerate on x^*. The difference between these three scoring rules is graphically displayed by Lad (1996, pp. 348–349).

4.2.2 Previsions for Scores of Distributions

As well as being able to score any assessed distribution upon observing $X = x^*$, we can also score our assessed distributions to measure how much information is contained in each asserted distribution. The amount of information asserted about $X \in \mathbf{R}(X)$ contained in $\mathbf{p}_N$ depends on the shape of the assessed distribution. The achieved value of X has no bearing on the information contained by the assessment. Information content is measured by the score a distribution expects to achieve. Two differently assessed distributions may contain the same amount of information, but achieve different scores. We should note how this works for the three scoring rules we have discussed.

Prevision for Quadratic Distribution Score

The prevision for a quadratic distribution score is

$$P\left(S[X, \mathbf{p}_N]\right) = \sum_{i=1}^{N} p_i^2 - 1. \tag{37}$$

Simple calculus shows that this prevision has a maximum value of 0, which occurs whenever the forecaster is sure of the exact value of X, asserting a distribution that is degenerate at that value. When this is the case, one element of $\mathbf{p}_N$ equals 1 and the other $(N-1)$ elements all equal 0. Remember that our prevised score does not depend on whether it turns out that the assessment is correct, it only measures how sure the assessor is that a certain outcome will occur.

The minimum value of $P\left(S[X, \mathbf{p}_N]\right)$, $(1/N) - 1$, occurs when all members of the constituent set are specified to have the same probability, that is, when a Uniform distribution is specified. Anyone who specifies $\mathbf{p}_N$ in this way is saying that they are equally (un)sure about each of the possible outcomes. Naturally this is when an assessment is least precise. The prevision for a quadratic distribution score is always non-positive. The closer $P\left(S[X, \mathbf{p}_N]\right)$ is to 0 the more information the assessment contains.

Prevision for Logarithmic Distribution Score

The prevision for a logarithmic distribution score is

$$P\left(S[X, \mathbf{p}_N]\right) = \sum_{i=1}^{N} p_i log(p_i). \tag{38}$$

This equals the well-known entropy of the distribution.

Prevision for Spherical Distribution Score

The prevision for a spherical distribution score is

$$P\left(S[X, \mathbf{p}_N]\right) = \left[\sum_{i=1}^{N} p_i^2\right]^{1/2}. \tag{39}$$

This number is the Euclidean length of $\mathbf{p_N}$, our vector of previsions. The larger the prevision for a spherical distribution score, the more information is contained in the assessment. $P\left(S[X, \mathbf{p}_N]\right)$ has a maximum value of 1 when a degenerate distribution is specified and a minimum value of $N^{-1/2}$ when a Uniform distribution is specified.

4.3 Digital Updating Procedure

Now we shall assess the flood exceedance quantiles using updated mixture mass functions as sequential forecasting distributions. We know from Ware and Lad (2003) that to undertake the computations involved in this assessment we must first construct:

- $\mathbf{R}(X)$, a vector to represent the realm of possible measurement values of X.
- $\mathbf{R}(\theta_1), \ldots, \mathbf{R}(\theta_p)$, vectors representing the realms of possible values for parameters $\theta_1, \ldots, \theta_p$.
- $f(x \mid \theta_1, \ldots, \theta_p)$, an array of mass values of size $s_X \times s_{\theta_1} \times \ldots \times s_{\theta_p}$, where s_{θ_i} is the size of $\mathbf{R}(\theta_i)$. The first dimension of the array corresponds to conditional probability mass functions for X given every different combination of values for $\theta_1, \ldots, \theta_p$.
- $f(\theta_1 \mid \theta_2, \ldots, \theta_p), \ldots, f(\theta_{p-1} \mid \theta_p)$, arrays of conditional probability mass values. Each array consists of vector θ_r, $1 \leq r \leq p-1$, evaluated over every $(\theta_{r+1}, \ldots, \theta_p)$ combination, for each element of θ_r. Thus the number of (unique) dimensions varies according to the number of conditioning parameters. For example $f(\theta_1 \mid \theta_2, \ldots, \theta_p)$ differs on p dimensions. Each of its component vectors corresponds to a different combination of $(\theta_1, \ldots, \theta_p)$ values. $f(\theta_{p-1} \mid \theta_p)$ differs on two dimensions. It has different component vectors for each (θ_{p-1}, θ_p) combination. For computational purposes each array must be the same size as $f(x \mid \theta_1, \ldots, \theta_p)$. Thus each array of conditional probability mass values must be replicated and tiled to form an array with $p+1$ dimensions. Each array has size $s_X \times s_{\theta_1} \times \ldots \times s_{\theta_p}$.
- $f(\theta_p)$, an array of marginal probability mass values. This array is of size $s_X \times s_{\theta_1} \times \ldots \times s_{\theta_p}$ and is identical across all but one dimension.
- $f(\theta_1, \theta_2, \ldots, \theta_p)$, a p-dimensional array representing the joint mass function $(\theta_1, \ldots, \theta_p)$. It is formed by element-wise multiplication of arrays $f(\theta_1 \mid \theta_2, \ldots, \theta_p), \ldots, f(\theta_{p-1} \mid \theta_p) f(\theta_p)$.

Once these arrays have been constructed we can implement our procedure for assessing items of interest, in this case exceedance quantiles, in the manner described in Ware and Lad (2003). The process we follow is:

1. Observe $X_i = x_i$.

2. Extract the array corresponding to X_i from array $f(x \mid \theta_1, \ldots, \theta_p)$. The extracted array will be of dimension p. This is the ITF through $\theta_1, \theta_2, \ldots, \theta_p$ from $X_i = x_i$ to X_{i+1}.

3. Implement Bayes' Theorem to update the mixing function $f(\theta_1, \ldots, \theta_p \mid \mathbf{X}_i = \mathbf{x}_i)$. This involves multiplying p-dimensional arrays $f(x_i \mid \theta_1, \ldots, \theta_p)$ and $f(\theta_1, \ldots, \theta_p \mid \mathbf{X}_{i-1} = \mathbf{x}_{i-1})$ element-wise, and normalising.

4. Replicate and tile array $f(\theta_1, \ldots, \theta_p \mid \mathbf{X}_i = \mathbf{x}_i)$ so it has $p+1$ dimensions and is the same size as $f(x_i \mid \theta_1 \ldots \theta_p)$.

5. Calculate $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$, the updated predictive mass function, by multiplying $f(x \mid \theta_1, \ldots, \theta_p)$ and $f(\theta_1, \ldots, \theta_p \mid \mathbf{X}_i = \mathbf{x}_i)$ element-wise, and then summing over $\theta_1, \ldots, \theta_p$.

6. Calculate any items of interest, e.g. conditional expectation, conditional exceedance quantiles.

Repeat this process as many times as required.

This procedure was undertaken with the ITF having three different functional forms. These were digitised GEV, GLO and LN3 densities. The realm of X was defined to range from 100 cumecs to 9000 cumecs at intervals of 10 cumecs. Any further refinement of AMS measures is not of practical use; in particular, treating X as if it is continuous may be useful for calculation purposes (if we could find a conjugate prior) but it adds nothing to our interpretation. Each of the three parameters for each ITF was specified initially to be Uniformally distributed over the components of its realm. These are reasonable, if conservative, prior distributions. Realms of the parameters are listed in Table 5. Notice that for each of our three examples the size of each of the four realms is the same: for the GEV and GLO

	GEV			GLO				LN3		
	Min	Inc	Max	Min	Inc	Max		Min	Inc	Max
X	100	10	9000	100	10	9000	X	100	10	9000
ξ	1000	12.5	1500	1000	12.5	1500	ξ	250	12.5	740
α	150	12.5	650	150	12.5	650	μ	3.5	0.125	8.5
k	-0.498	0.012	-0.03	-0.498	0.012	-0.03	σ	0.49	0.052	1.01

Table 5: Elements of realms used in digital computations. "Min" and "Max" denote the smallest and largest elements of the realm. "Inc" denotes the increment between successive elements.

ITF's we have $s_X = 891$, $s_\xi = 41$, $s_\alpha = 41$ and $s_k = 40$, and for the LN3 ITF the realm sizes are $s_X = 891$, $s_\xi = 41$, $s_\mu = 41$ and $s_\sigma = 40$. Thus the size of each four dimensional array is $891 \times 41 \times 41 \times 40$. For simplicity, from here on the characterising parameters of the ITF will be labeled as ξ, α, and k, regardless of the form of the ITF. An array representing the ITF is computed by evaluating $f(x \mid \xi, \alpha, k)$ at each of the 59,910,840 possible (X, ξ, α, k) combinations.

4.4 Results of the Digital Forecasting Procedure

At any step of our digital computations we can compute any item that is of interest to us. In this case we are particularly interested in the shape of $f(x_{i+1} \mid \mathbf{X_i} = \mathbf{x_i})$, the updated predictive mass function, from which we can compute the value of any conditional exceedance quantile we wish. Figure 9 demonstrates the shape of the predictive mass function at different stages during the observation of the 72 AMS measurements from the Waimakariri River, when the ITF is specified to have a GEV form. Notice that the variance of $f(x_{i+1} \mid \mathbf{X_i} = \mathbf{x_i})$ decreases as more observations are processed. As we record more observations $f(x_{i+1} \mid \mathbf{X_i} = \mathbf{x_i})$ becomes more 'peaked'.

Conditional exceedance quantile forecasts are shown in Table 6. Note the similarity across the three different ITF's. Remember that for the frequentist at-site method the GEV and LN3 estimates were close to one another, but the estimate based on the GLO was significantly smaller for $Q(F) < 0.995$. Comparing the fore-

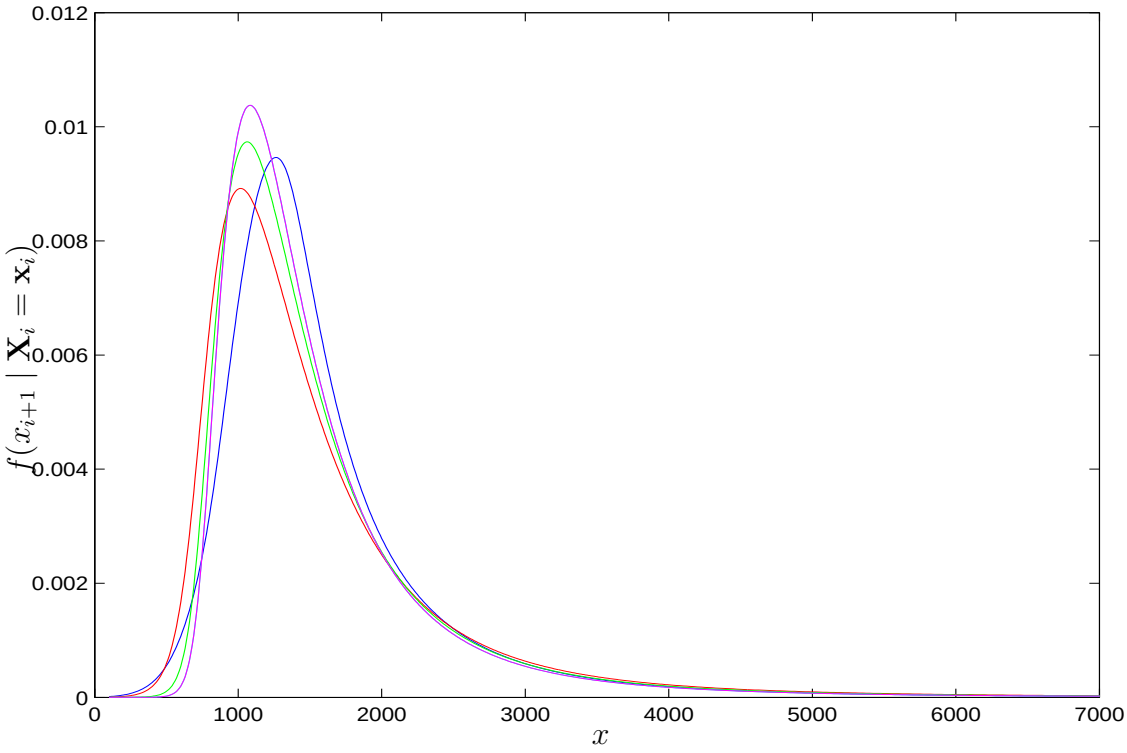

Figure 9: Marginal mass function of X after 0 (blue), 25 (red), 50 (green) and 72 (magenta) recorded observations. The ITF is specified to have a GEV form.

casts to the frequentist quantile estimates we see that the values of $Q(0.95)$ are similar, but as the return period increases the distance between the digital forecasts and the frequentist estimates increases.

In Section 3.5.1 we found point estimates for each parameter in the distribution under investigation. Now we can compute a mass function that represents our updated knowledge about the location of each parameter. Figure 10 represents these mass functions for the GEV ITF.

	$Q(0.95)$	$Q(0.98)$	$Q(0.99)$	$Q(0.995)$	$Q(0.999)$
GEV	2840	3500	3920	4720	7000
GLO	2770	3440	3890	4650	6950
LN3	2800	3390	3770	4200	5950

Table 6: Conditional exceedance quantile forecasts for the Waimakariri River when the ITF is assessed to have a particular functional form.

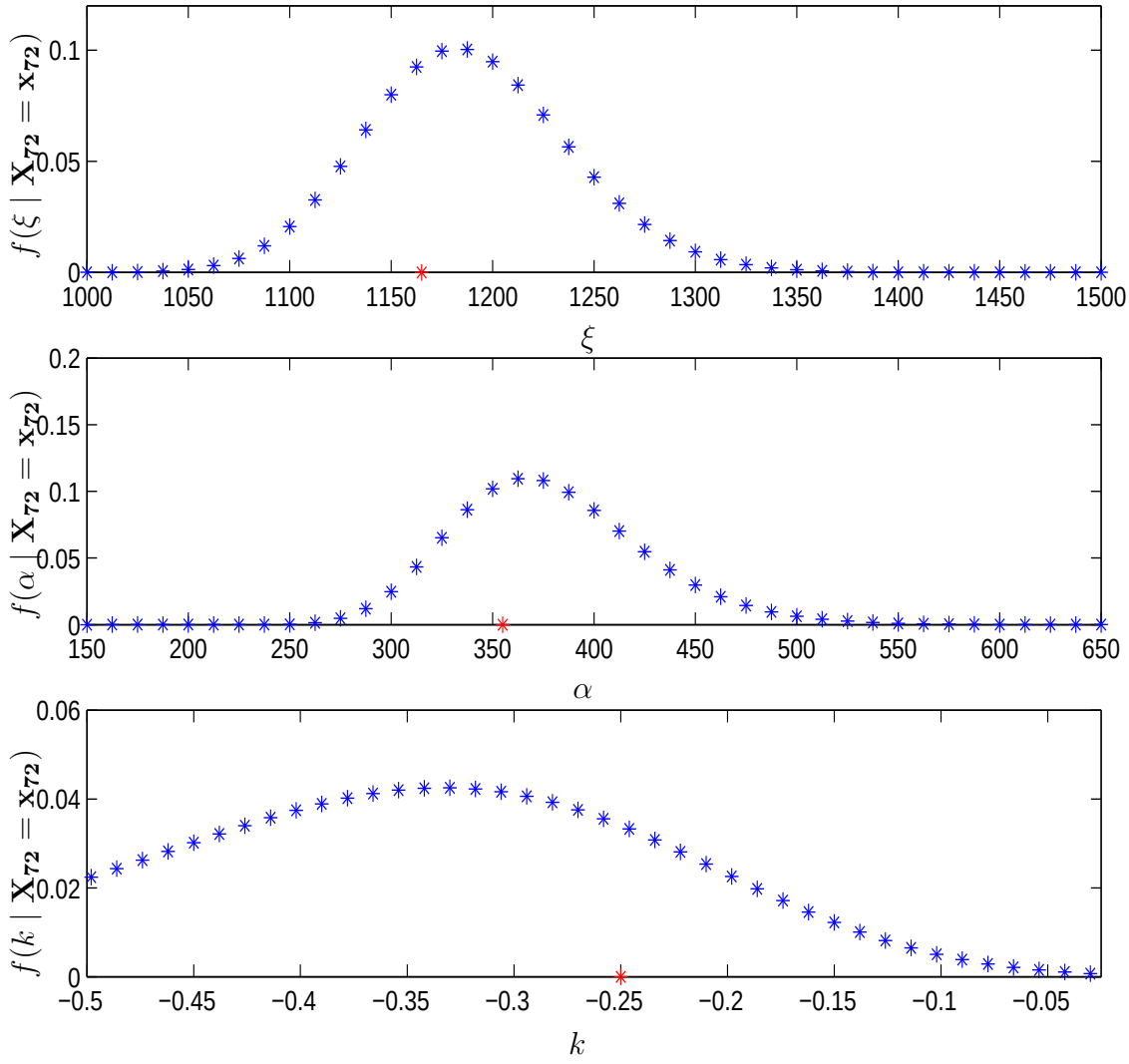

Figure 10: Conditional marginal mass functions for, in descending order, ξ, α and k, when we specify that the ITF has a GEV form. The frequentist at-site point estimates are marked on the x-axis by red "$*$".

Remember that by using a frequentist at-site procedure we calculated point estimates of $\hat{\xi} = 1165$, $\hat{\alpha} = 355$ and $\hat{k} = -0.25$. These estimates are marked with a red "$*$" in Figure 10. We can see that for the location and spread parameters the frequentist at-site parameter point estimate is close to the mode of the predictive mass function. The mode of shape parameter k is approximately -0.33, as opposed to the point estimate of -0.25. The shape of the mass functions tells us how sure we are in mixing forecasts over the size of a parameter. In this case, the broadness of the mixing mass functions displays how uncertain we are about the values of the characterising parameters. Compare this to the frequentist procedure, where all

that is specified about a parameter is a point estimate. Of the three parameters k is defined in the least-precise manner. This is consistent with the paradigm of the frequentist regional hierarchical estimation method — that we will need more observations to achieve an equivalently good estimate of the shape parameters than we will for the spread parameter or location parameter.

4.4.1 The Scores of Distributions

We have obtained exceedance quantile forecasts for each of the three forms of specified ITF. To empirically evaluate these three competing theories we will score each of the observations and compare the results. This is called scored sequential forecasting. As well as investigating the difference between the candidate ITFs, we shall use this as an exercise to examine the differences between the quadratic, logarithmic and spherical scoring rules.

Quadratic Score

The quadratic score is the first scoring rule that we shall consider. Quadratic scores were calculated for each observation, for each of the three forms of ITF. The results are shown in Figure 11. The upper panel displays the score of the distribution as each observation is processed. We score the GEV (blue), GLO (red) and LN3 (green) distributions. The lower panel shows the cumulative scores. For clarity the cumulative scores are plotted as the difference between the achieved score and -1.008, the minimum value achieved by any of the scores. Plotting cumulative scores in this manner means that the larger the score is, the better.

The scores achieved using ITFs of different functional forms are most different from each other when i is small. The bigger discrepancies in scores appear early on in the analysis because this is when the form of the ITF itself has most effect, because the data has not been recorded for a long enough period to dominate the results. Notice that the first major 'jag' in the upper panel of Figure 11 occurs at $i = 7$, which, according to Figure 1, is when the first recorded observation differed significantly from the previous sequence of AMS recordings. Notice that, although the scores produced when the ITF has a GEV or GLO form are more similar to each other than to the scores produced when the ITF has a LN3 distribution (especially

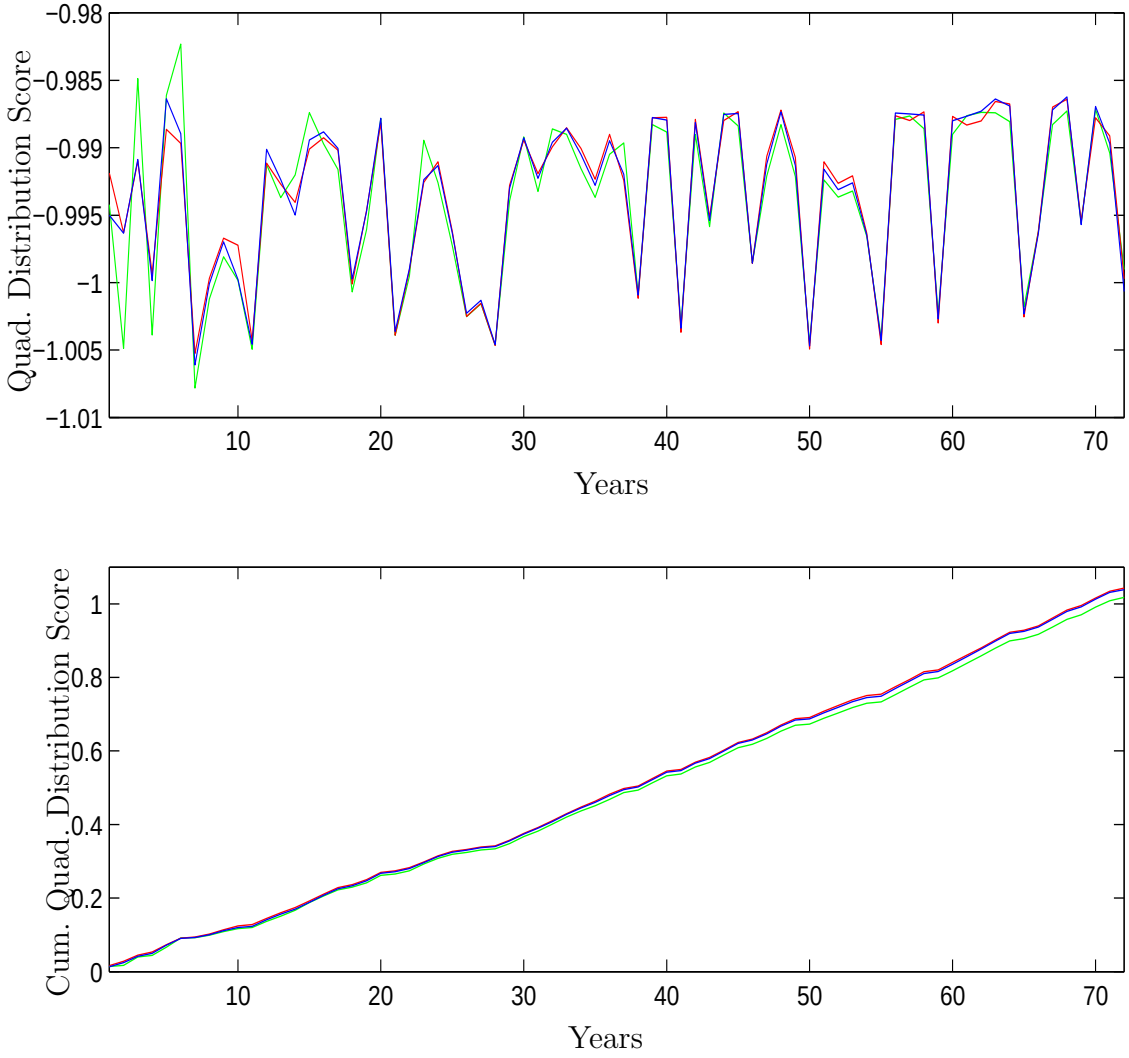

Figure 11: Sequential individual (upper panel) and cumulative (lower panel) quadratic score of a distribution for ITF with functional form: GEV (blue), GLO (red) and LN3 (green).

when $i < 17$), in general the scores associated with the three ITFs are similar. All scores take values within a very narrow range. This is not unexpected when we consider the construction of $S(X, \mathbf{p}_N)$. As Equation 34 shows, the quadratic score consists of two parts. The first part of $S(X, \mathbf{p}_N)$ is $\sum_{i=1}^{N} (X = x_i)(1 - p_i)^2$. Remember that $(X = x_i)$ is a vector that contains one "1" and $(N-1)$ "0"s. There are a large number of elements in the realm of X, so the expectation that any single one of them will occur is small. In fact, as Figure 9 shows, the maximum expectation that any X will occur is less than 0.01. Thus the difference between alternative potential scores is small for different outcome vectors. Similarly the

second part of $S(X, \mathbf{p}_N)$, $\sum_{i=1}^{N}(X \neq x_i)p_i^2$, is the sum of the squares of all the values contained in $\mathbf{p}_N$, excluding the value assigned to the actual outcome. Again, because $\mathbf{R}(X)$ is large and the value assigned to any individual X is small there will only be a small difference over different "unsuccessful" values of X. The lower panel shows the LN3 distribution has the minimum (and therefore worst) score. The GEV and GLO scores are very close, with the GLO score being slightly better.

Logarithmic Score

The logarithmic scores can be interpreted in a similar manner to the quadratic scores. Remember that the logarithmic score is just the logarithm of the probability assigned to the value of X that occurs. Because there is little difference between the various assigned values of $f(x_{i+1} \mid \mathbf{X_i} = \mathbf{x_i})$, the difference of their logarithms will also be small. Thus we can expect the range of scores for individual recorded measurements to be small. For example, the maximum score attained is -4.36, for $i = 6$ when the functional form of the ITF is assessed as LN3. The minimum score is -8.44 for $i = 28$ when the functional form of the ITF is assessed as GLO. The year corresponding to $i = 28$ is 1963, the year of the largest recorded flow.

Achieved scores are displayed in the upper panel of Figure 12. Cumulative scores, which are displayed in the lower panel, are shown as the cumulative difference of the recorded and minimum score. When scores are displayed in this way, the larger the score is the better. The lower plot shows the cumulative sum of Log Scores. The GEV has the maximum score, followed by GLO and LN3. All three scores appear very similar, which is as we expect considering the size of the differences between the different forms of the ITF in the upper panel, and the large scale on the y-axis in the lower panel. .

Spherical Score

The last scoring method we consider is the spherical score. Remember from Equation 36 that the spherical score is the assessment that has been placed on x^* divided by the square root of the sum of each of the asserted probabilities squared. As in the previous two cases we do not expect there to be a large difference between the different scores because of the relative similarity between the mass values attached

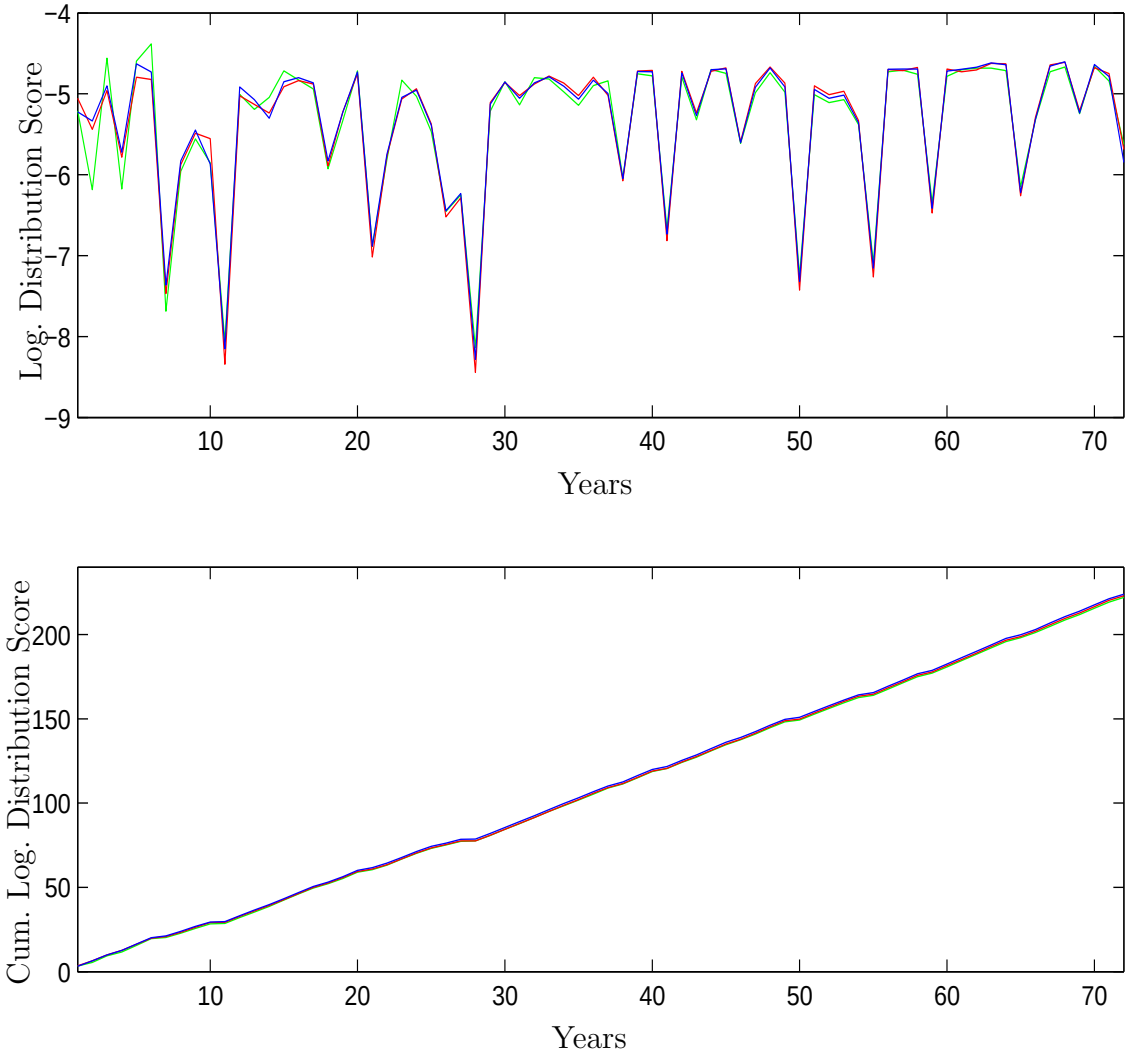

Figure 12: Sequential individual (upper panel) and cumulative (lower panel) logarithmic score of a distribution for ITF with functional form: GEV (blue), GLO (red) and LN3 (green).

to the elements of $\mathbf{R}(X)$

The upper panel of Figure 13 shows that the shape of the individual scores is similar to that obtained from the quadratic and logarithmic scores. However there is a relatively large difference between the three forms of ITF, and in particular between the LN3 and the other two, for $i < 17$. As the number of observations increases, each individual observation has less of an impact on the shape of $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$, and consequently there is less of a difference between the scores attained through the three candidate distributions. The lower panel of Figure 13 shows the cumulative scores. Remember that, when dealing with spherical scores, the larger the score is

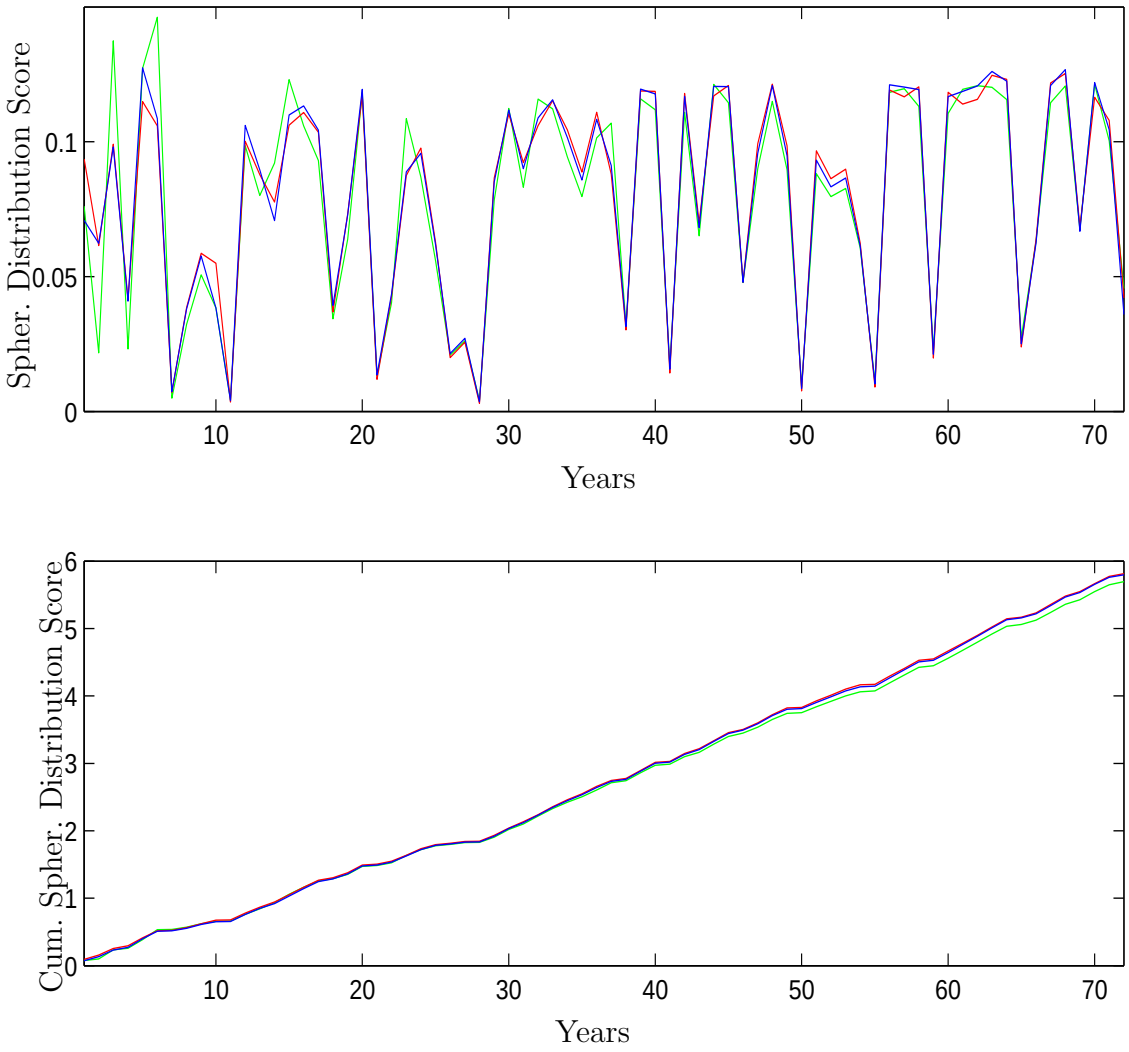

Figure 13: Sequential individual (upper panel) and cumulative (lower panel) spherical score of a distribution for ITF with functional form: GEV (blue), GLO (red) and LN3 (green).

the better. Once again the LN3 score is the worst. The GEV and GLO scores are very similar, with the GLO score being slightly better.

Before we leave scoring rules, we shall briefly examine what information the predictive mass functions claim to provide about our current state of uncertain knowledge about X.

4.4.2 Previsions for Scores of Distributions

Previsions for scores of distributions measure how sure the researcher is about the assertions they have made. The most apparent thing about the Previsions of Scores

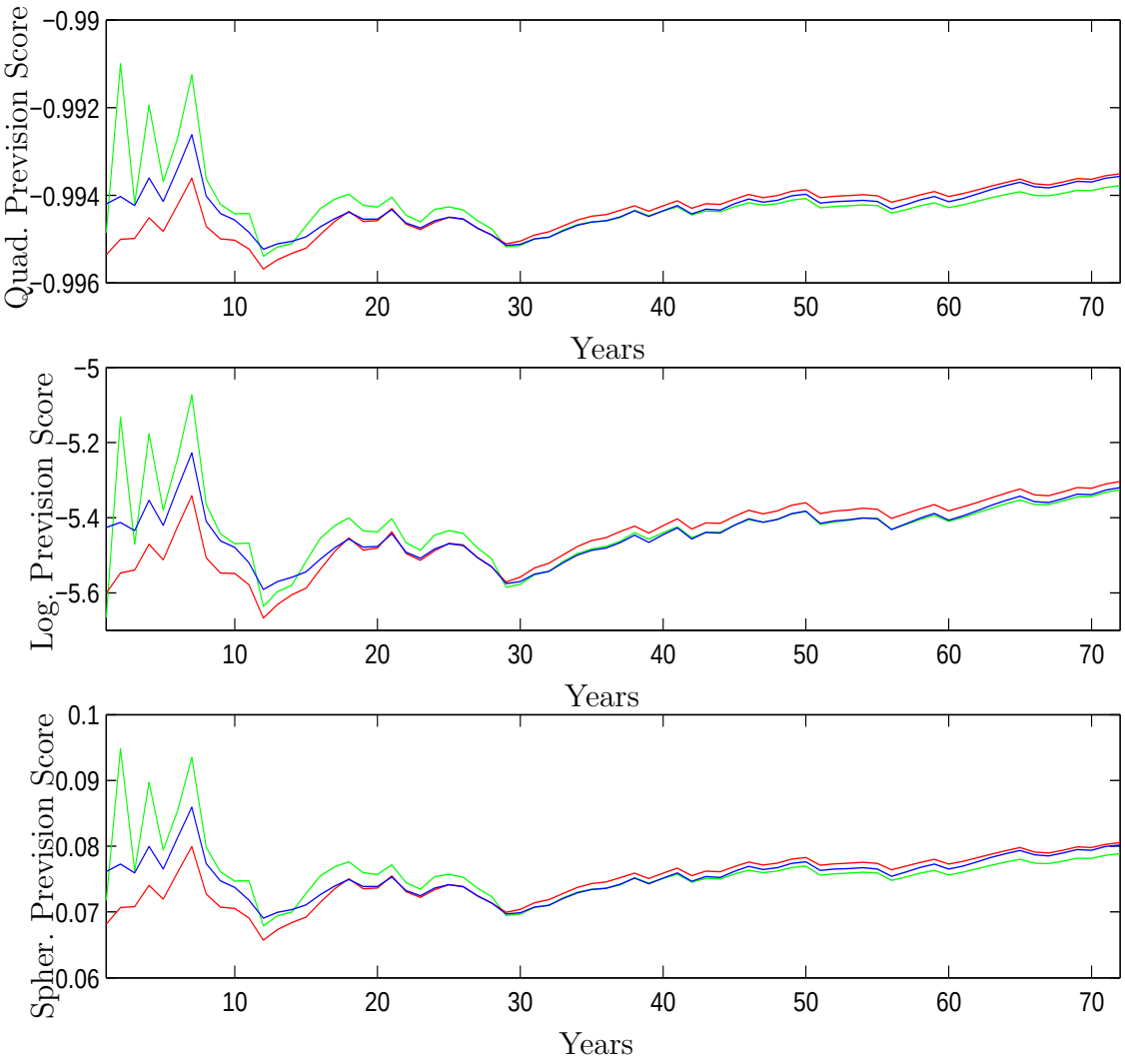

Figure 14: Scores of sequential previsions for ITF with functional form: GEV (blue), GLO (red) and LN3 (green). Scoring rules used are (in descending order) quadratic, logarithmic and spherical.

in Figure 14 is how similar they look over the three different types of score. Notice that the range of differences between any of the previsions for scores is small, as it was with the scores we discussed in the preceding Section. We start being most sure about assessments when the form of the ITF is specified as LN3. At approximately $i = 30$ the previsions related to the LN3 scores drop until they are slightly below previsions relating to the GEV and GLO forms of ITF. From $i = 30$ onward our assessment of previsions appears to improve to some extent, in all cases exceeding the initial expected score of the first period. Now that we have established a procedure for scoring sequential forecasts created by updating mass functions, we move on to

comparing these digital approximations with frequentist estimates.

5 Comparing Frequentist Estimates and Digital Forecasts

In Section 3 we used conventional frequentist techniques to find point estimates of the characterising parameters of distributions commonly used in the study of flood frequency analysis. These point estimates were then used to estimate flood exceedance quantiles. In Section 4 we constructed and implemented a digital updating procedure. This was used to score sequential forecasting methods for an analysis which involved sequences of observations regarded exchangeably. To conclude this Report we shall compare the two approaches. First, we base our comparison on scores obtained with the Waimakariri River AMS data. Second, we shall compare the two procedures using a simulated data set.

Frequentist estimates, calculated by using the method of L-moments, and subjective forecasts, achieved via updated mixture mass functions, are fundamentally different. Thus, no method for comparing the two procedures really makes sense. Subjective theory is not orientated towards estimating parameters (that don't exist) but toward forecasting measurements (which are functions of real historical records). Ultimately, what any statistical method should be able to do is to forecast historical measurements in the context of uncertainty. Thus, we shall score forecasts based on objectivist estimators against real subjectivist sequential forecasts. To conclude this Report we shall cater to the objectivist and abuse the subjectivist outlook to measure subjective Bayesian forecasts using frequentist criteria, namely the bias and root mean-squared-error.

Before studying the Waimakariri River AMS data again, we shall take another brief diversion to describe proper scoring rules for expectations and variances.

5.1 Proper Scoring Rules for Expectations

In Section 4.2 we introduced proper scoring rules as measures used to evaluate states of uncertain knowledge. Although our main focus was on proper scoring

rules for distributions, we briefly mentioned that we can also score assessments of expectations and variances, which we shall denote $E(X)$ and $V(X)$ respectively.

One proper scoring rule we can use is the quadratic scoring rule. The quadratic score of K is defined as

$$S(X, K(X)) = -(X - K(X))^2, \tag{40}$$

where $K(X)$ is some numerical assessment of X. This is an analogue of the quadratic scoring rule for distributions given in Equation 34. The quadratic scoring function is concave. It's maximum occurs when the prevised value of X is achieved. Notice that the achieved score only depends on the difference between X and $K(X)$, and that $S(X, K(X)) = S(K(X), X)$. The quadratic scoring rule is the only proper scoring rule for which either of these properties holds true. Notice that the quadratic score is the negative of the squared difference — this is in keeping with our idea that a larger score is better. Remember that the score for a distribution is the difference between asserted and observed probabilities, whereas the score for an expectation is the difference between the asserted and achieved values of X.

The quadratic score for the expectation and variance of X can be written as

$$S(X, E(X)) = -(X - E(X))^2 \tag{41}$$

$$\text{and} \qquad S(X, V(X)) = -((X - E(X))^2 - V(X))^2. \tag{42}$$

In the subsequent subsections we shall score the expectation and variance of $X_{i+1} \mid (\mathbf{X}_i = \mathbf{x}_i)$, for various values of i.

Our motivation for introducing scoring rules for expectations and variances is because we cannot directly compare the scores of distributions for the frequentist and digital procedures. Remember that the frequentist method of L-moments is used to estimate the characterising parameters of the GEV distribution. These parameters are then used to construct a continuous density function. The continuous density function is an approximation of the predictive mass function of X. Scores of approximating continuous probability density functions are found using continuous analogues of Equations 34, 35 and 36. For example, the logarithmic score of a continuous probability density function is

$$S(X, f_X(\cdot)) \equiv log\left(f_X\left(X\right)\right). \tag{43}$$

See pp. 350 of the text of Lad (1996) for the proper scores of other continuous approximating distributions. This score is not readily comparable to the score of the predictive mass function obtained through updated mixture mass functions. Remember that the predictive mass function, $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$, is a vector whose size is defined by the size of $\mathbf{R}(X)$. Since $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ is normalised to sum to 1, the size of the mass attached to the observed outcome will depend on the size of $\mathbf{R}(X)$. That is, the more finely delimited $\mathbf{R}(X)$ is, the lower the mass will be at any particular point. Although we can't compare the score of an approximating continuous distribution directly with the score of a mass function, one compromise which allows us to compare the two different procedures is to digitise the approximating continuous density over $\mathbf{R}(X)$.

5.2 Comparing L-moment Estimates and Updated Mixture Mass Forecasts using Scoring Rules for the Waimakariri River Annual Maxima Series Data

This Section involves using proper scoring rules to sequentially score the frequentist estimates and digital forecasts of exceedance quantiles.

5.2.1 Estimating the Conditional Mean and Variance Using the Method of L-moments

In Section 3.3 we described how to use the method of L-moments to estimate the parameters of the "underlying distribution" which "generates the series of random outcomes that compose the AMS". We can estimate these parameters at any stage that interests us. The only constraint, when using the method of L-moments to estimate the parameters of the GEV distribution, is that we must have observed at least three values in the sequence, see Equations 9 and 15. Once L-moment estimates of the parameters of the GEV distribution, ξ, α and k, are obtained, they can be used to compute $E(X)$ and $V(X)$ using Equation 13 and Equation 14.

We shall forecast $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ as $E(X_i)$ and $V(X_i)$ respectively, where $E(X_i)$ and $V(X_i)$ are the expectation and variance after i observations. From now on whenever we refer to the "frequentist forecast at stage

$i + 1$", we will be referring to the statistics estimated after i observations.

5.2.2 Forecasting the Conditional Mean and Variance Using Updated Mixture Mass Functions

In Section 4 we described how to forecast the value of various items when interest centres on the sequence of updated mass functions $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$. At each successive step the digitally updated predictive mass function is used to assess our expectation and variance for the next X value to be observed. The computational procedure involves generating a digitised prior mixing mass function and digitised ITF and using them, via Bayes' Theorem, to update the posterior mass function, and hence the predictive mass function. The ITF is specified to have a GEV form.

The realm of possible AMS values, X, as well as the realms of the characterising parameters ξ, α and k, are the same as those defined as in Section 4. They are displayed in the column of Table 5 headed "GEV". The updated predictive mass function is used to calculate the expectation and variance of $X_{i+1} \mid (\mathbf{X}_i = \mathbf{x}_i)$, as detailed in Section 2.2 of Ware and Lad (2003).

5.2.3 Results

Forecasts $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ were computed for both the digital and frequentist cases, for $i \geq 4$. In the upper panel of Figure 15 the sequence of updated conditional expectations are displayed. The subjectivist forecasts are represented by blue "+" and the frequentist forecasts are represented by red "×". The lower panel of Figure 15 displays the sequence of updated conditional variances. Notice that for both items the frequentist estimates are much more varied at the start, when each new observation has a greater influence on the parameter estimates. For example, notice how the frequentist variance increased in the $8th$ year, after being influenced by (what was at that stage) the highest AMS recorded value of $2660m^3/sec$. Forecasts for both procedures stabilise toward the same value, although even after 72 observations they are still noticeably different.

The sequential quadratic scores of the conditional expectation are displayed in Figure 16. The upper panel displays the cumulative sum of individual scores, starting at year 20. The digital forecast is displayed in blue and the frequentist forecast

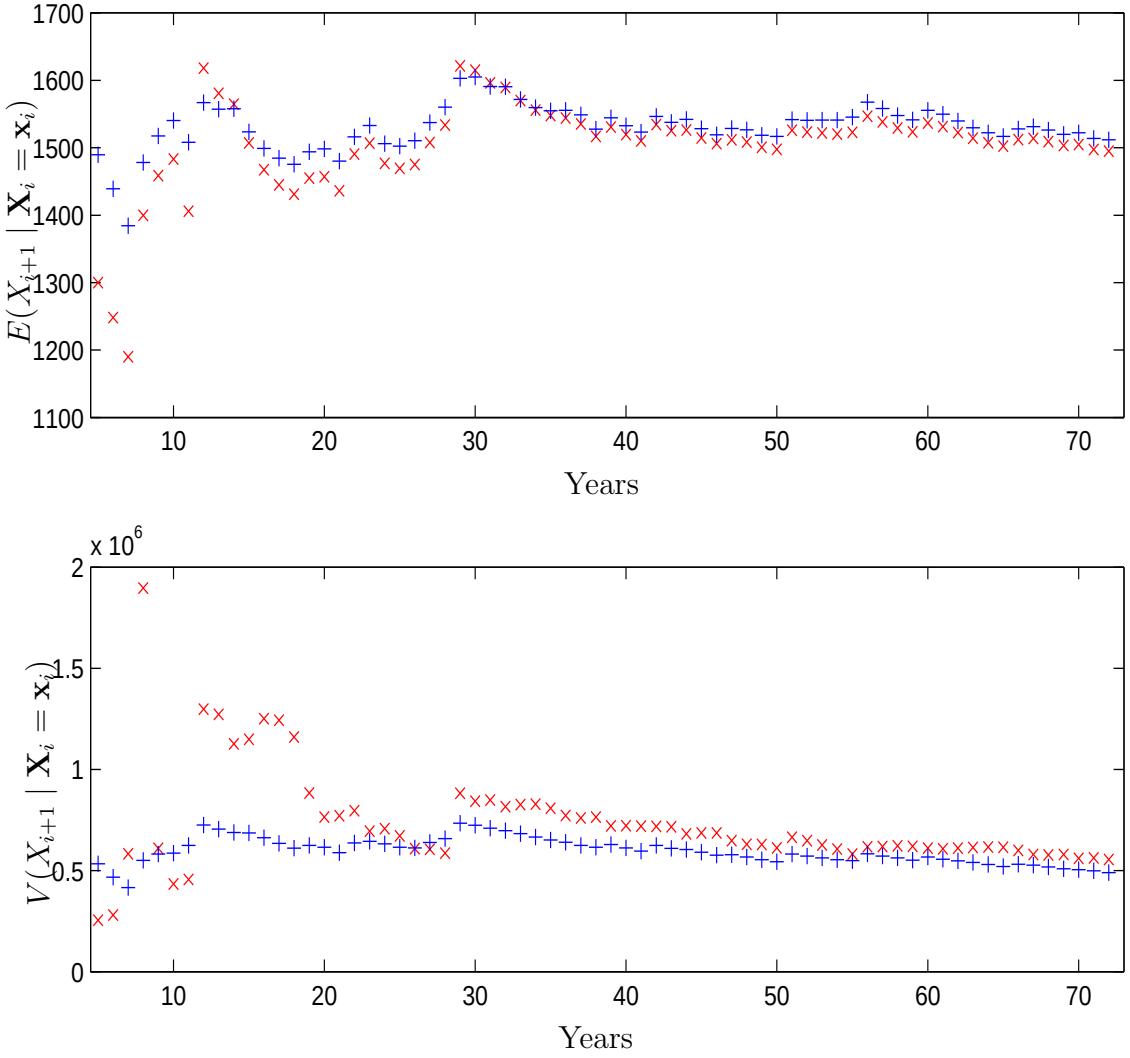

Figure 15: Conditional Means and Variances of the Waimakariri River AMS series. Digital Forecasts are displayed as "+". Estimates obtained via L-moments are displayed as "×".

in red. Remember that each individual score is the difference between the assessed expectation and the observed AMS value. By viewing the upper panel of Figure 1 and the upper panel of Figure 15, and squaring the distance between observation X_{i+1} and forecast $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ we get an idea of the quadratic score for any particular observation. For example, notice there is a big jump in the cumulative score for $E(X_{28} \mid \mathbf{X}_{27} = \mathbf{x}_{27})$. If we look at Figure 1 we can see that the observed maximum flow in 1957 (the 28th year on record) is $3990m^3/sec$. The digital conditional expectation for the instantaneous maximum flow in 1957 was approximately $1560m^3/sec$. Consequently, the score for the $28th$ year is approximately $-6,000,000$

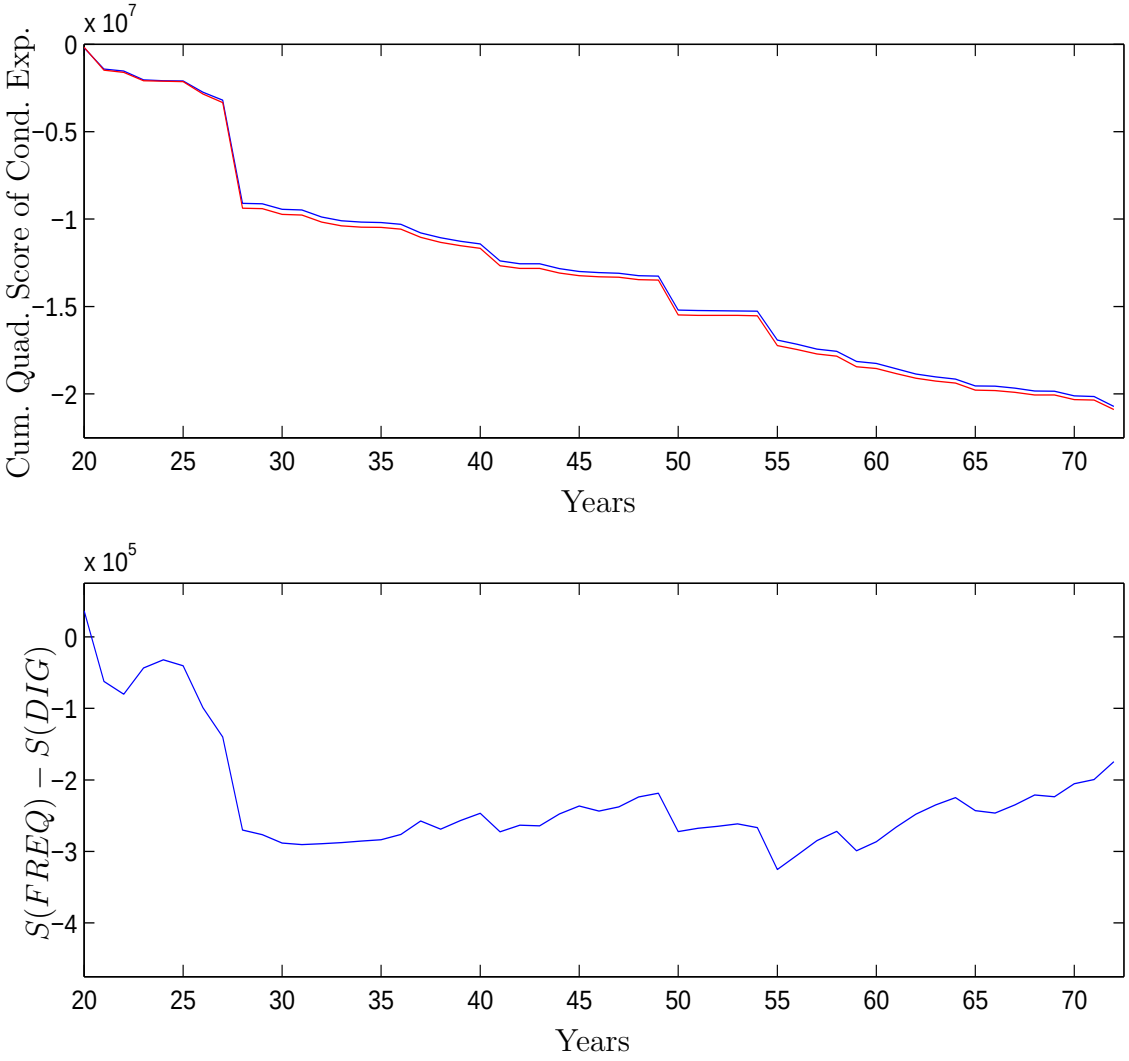

Figure 16: Sequential cumulative quadratic score of the conditional expectation for digital (blue) and frequentist (red) forecasts (upper panel). The lower panel displays the difference between the forecasts.

(from $-(3990-1560)^2$). The scores of the expectations of the digital and frequentist procedures follow the same pattern. At the end of the recording period the digital score is marginally larger, that is, it is closer to 0, and thus slightly better.

The lower panel of Figure 16 displays the difference between the cumulative scores recorded by the two procedures. The scores of the digital and frequentist procedures are denoted "S(DIG)" and "S(FREQ)" respectively. We can see that until the $30th$ value of the AMS sequence is recorded, the digital score has been consistently improving relative to the frequentist score, but after this point the difference between the scores stabilises. There is a suggestion that after the $60th$

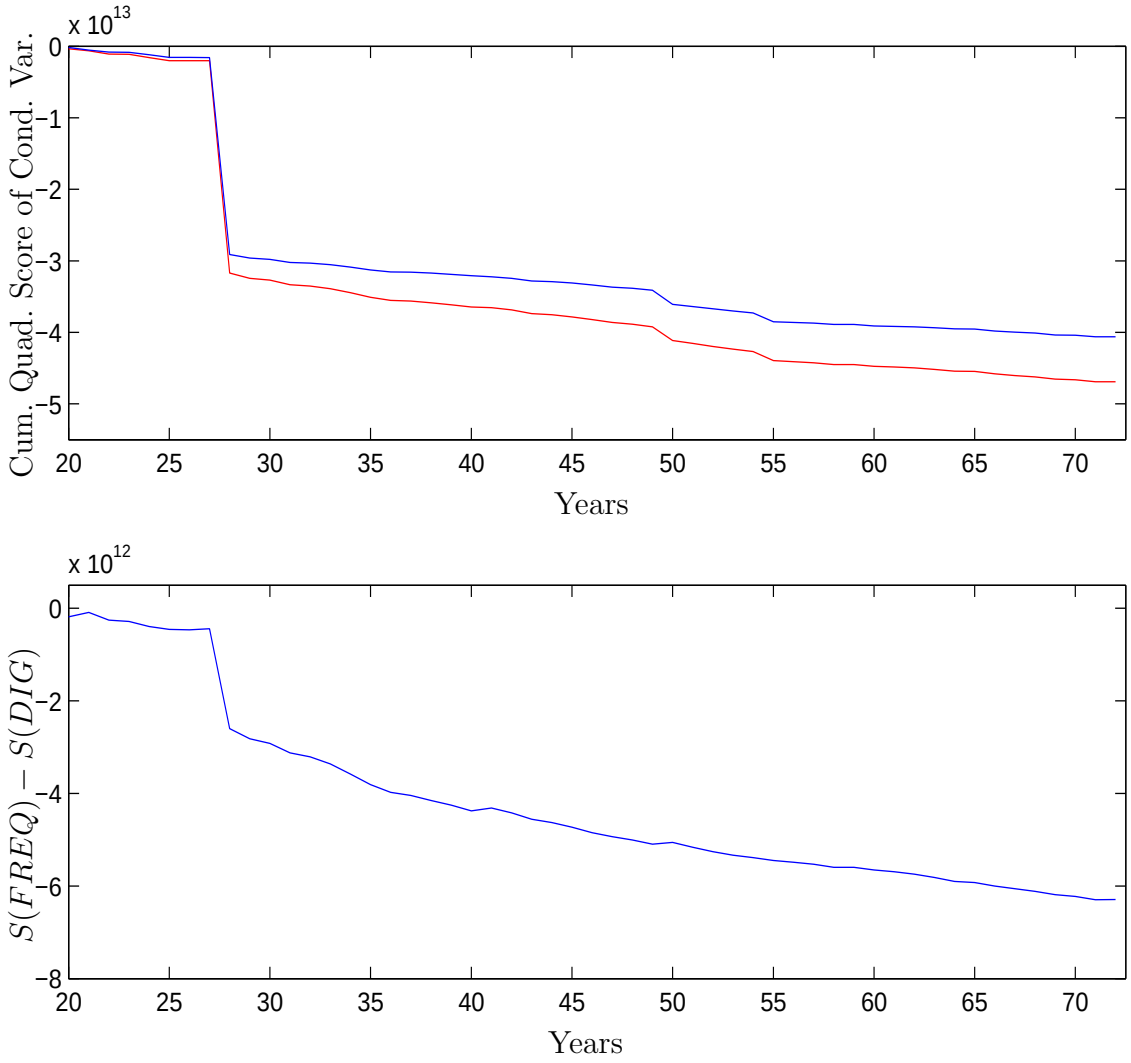

Figure 17: Sequential cumulative quadratic score of the conditional variance for digital (blue) and frequentist (red) forecasts (upper panel). The lower panel displays the difference between the forecasts.

year the frequentist forecast scores slightly better than the digital forecast.

If we had have started scoring earlier than the $20th$ year, then the difference between the cumulative scores would have been greater, as a study of the forecast expectations in Figure 15 and the actual outcomes in Figure 1 shows.

The sequential quadratic scores of the conditional variance are displayed in Figure 17. The upper panel displays the cumulative sum of individual scores, starting in the $20th$ year. The lower panel of Figure 17 displays the difference between the cumulative scores. Observe that the scores have approximately the same shape, but the difference between the scores continues to increase as the number of observa-

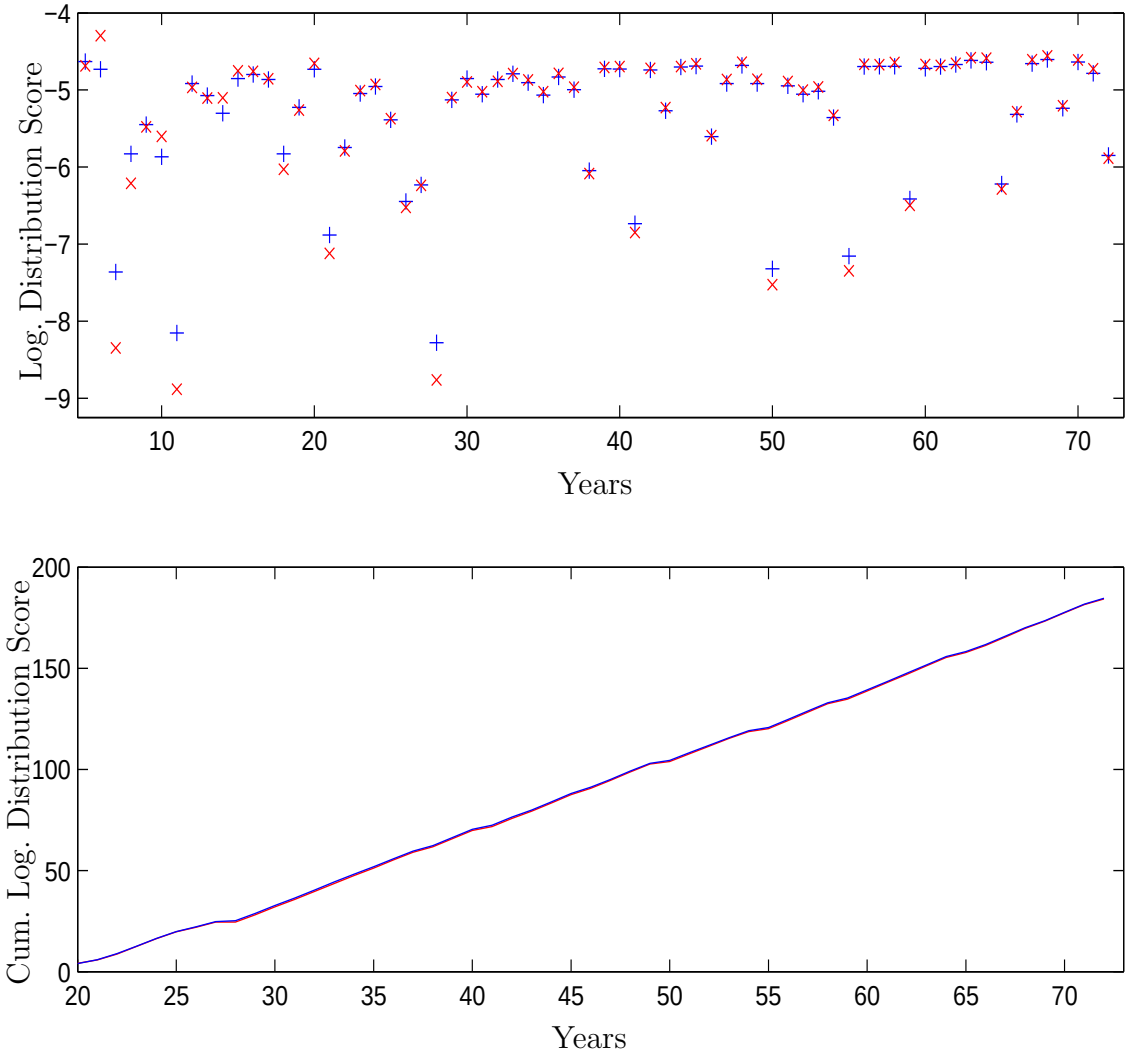

Figure 18: Sequential individual (upper panel) and cumulative (lower panel) logarithmic score of a distribution for digital (blue) and frequentist (red) forecasts. The cumulative scores are almost indistinguishable.

tions increases. Clearly the digital procedure produces a better overall score. The conclusion we can take from Figures 16 and 17 is that the frequentist and digital procedures do a similar job of assessing conditional expectation, but the digital procedure is much better at assessing conditional variance.

Now that we have detailed the scores of the conditional expectation and variance, we shall briefly consider the logarithmic score of the forecasting distributions. Remember that after each new observation is recorded, we can use the method of L-moments to re-estimate the parameters of the GEV distribution and thus construct the probability density function. After digitising the density over $\mathbf{R}(X)$, we

can find the logarithmic score of the forecasting distribution of X using Equation 35.

The logarithmic score for the sequential forecasting distributions are displayed in Figure 18. The upper panel displays the individual scores from the $5th$ year onwards. The lower panel displays the cumulative scores, where scoring starts in the $20th$ year. In the lower panel the cumulative scores are shown as the difference between the recorded score and -8.85, the minimum score achieved by any value after the $20th$ year. Displaying scores in this way means that, as usual, the larger the score is, the better. At this scale the different cumulative scores are indistinguishable.

There is little difference between the cumulative scores, although the updated mixture mass forecasting procedure is slightly better, with a score of 184.5. The frequentist procedure scored 184.2. If we had have started cumulatively scoring earlier, there would be a bigger difference between the scores — in particular see the individual scores for years 7 and 11.

Now that we have compared scores from a real data set, we shall use a Monte Carlo procedure to compare scores from a simulated data set.

5.3 Comparing L-moment Estimates and Updated Mixture Mass Forecasts using Scoring Rules for Data Generated from a Generalised Extreme Value Distribution

We shall use a Monte Carlo procedure to compare forecasts made using L-moment estimates against forecasts achieved via updated mixture mass functions. The Monte Carlo procedure consists of two parts. First we simulate AMS data for a large number of sites. Then we score each site's data and analyse the results.

Data was simulated as in Section 3.7.1. The data was generated from a GEV distribution with $E(X) = 2$, $CV = 0.5$ and $k = -0.14$. These are the same conditions that we used in Section 3.7.1 when simulating data from site 11 of a region with median CV of 0.5. We simulate data from 10,000 sites, each of which has 100 years of recorded AMS values.

Forecasts are made using L-moment estimates in the manner described in Section 5.2.1. Forecasts are achieved via updated mixture mass functions as described

CV		At-site			Regional					
					I			II		
		Min	Inc	Max	Min	Inc	Max	Min	Inc	Max
0.5	X	0	0.05	20	0	0.05	20	0	0.05	20
	ξ	0.975	0.075	1.5	0.5	0.025	0.7	0.975	0.075	1.5
	α	0.75	0.1	1.45	0.4	0.025	0.6	0.75	0.1	1.45
	k	-0.375	0.025	-0.025	-0.375	0.025	-0.025	-0.375	0.025	-0.025
1	X	0	0.05	20	0	0.05	20	0	0.05	20
	ξ	0.6	0.05	1.3	0.3	0.05	0.7	0.6	0.05	1.3
	α	0.6	0.05	1.3	0.45	0.025	0.65	0.6	0.05	1.3
	k	-0.375	0.025	-0.025	-0.375	0.025	-0.025	-0.375	0.025	-0.025

Table 7: Elements of realms used in digital computations. "Min" and "Max" denote the smallest and largest elements of the realm. "Inc" denotes the increment between successive elements. "I" and "II" refer to the two different stages in the regional procedure.

in Section 5.2.2. The digital forecasting procedure requires that we specify the realm of X, and the realms of the characterising parameters of the GEV distribution, ξ, α and k, before any calculations can take place. The realms we use are listed in the column of Table 7 headed "At-site". The conditional mass functions, $f(\xi \mid \alpha, k)$ and $f(\alpha \mid k)$, and the marginal mass function $f(k)$, were defined to be Uniformally distributed over their realm.

Sequential quadratic scores of conditional expectation and variance were calculated for every site, starting at the $20th$ year. The mean of the cumulative scores of the conditional expectation are displayed in the upper panel of Figure 19. As usual, the larger the score the better and because the quadratic score is the negative of the squared difference, the closer the score is to 0 the better. The difference between the cumulative scores is displayed in the lower panel of Figure 19. It is clear that as the number of observations increases, the mixture mass forecast is increasingly better than the frequentist forecast, although the rate of increase is slowing. However, after 100 observations the mean difference between the two scores is only just above 4, out of a total score of approximately -320. Thus, although the score achieved by the

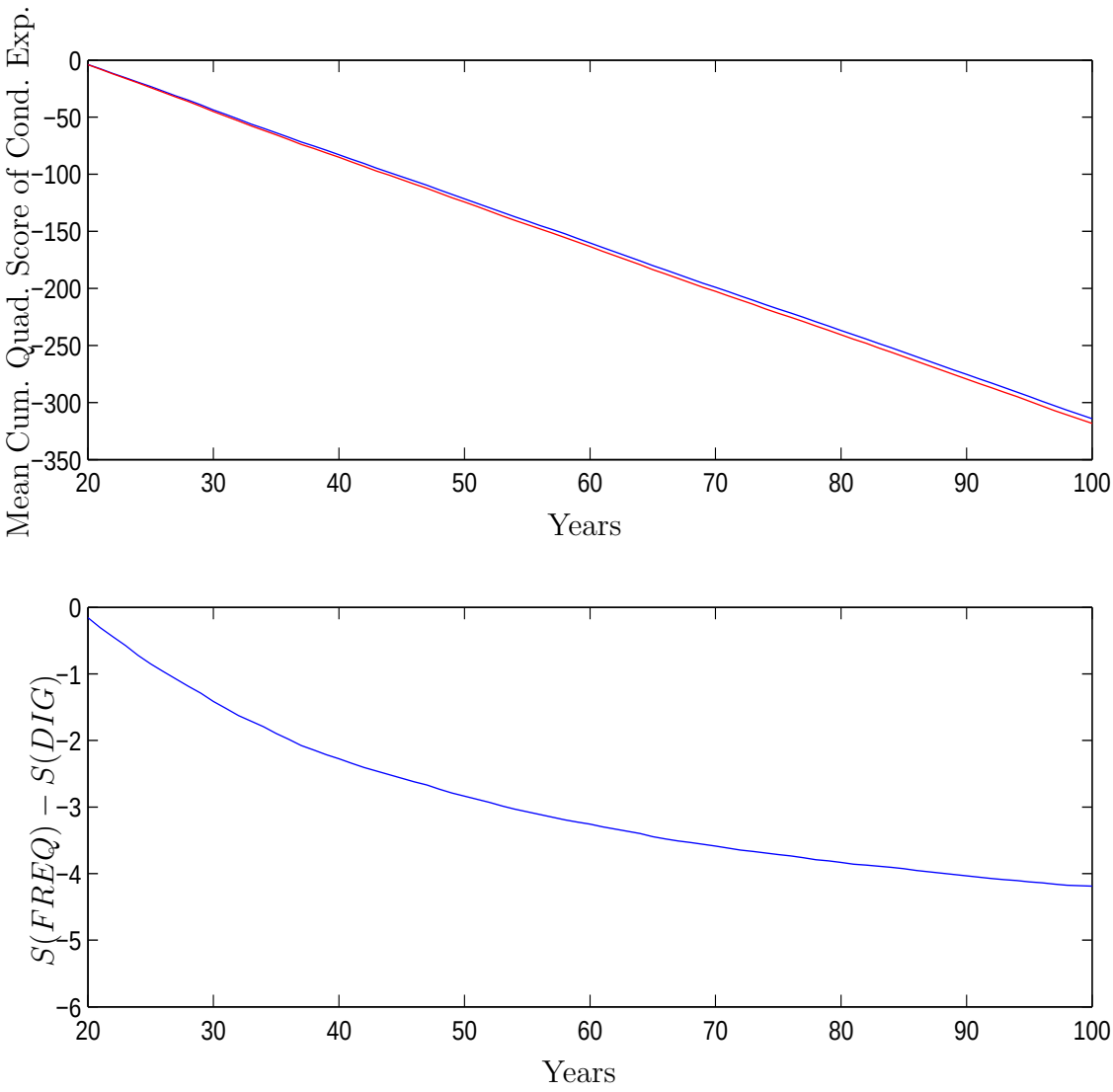

Figure 19: Sequential mean cumulative quadratic score of the conditional expectation for digital (blue) and frequentist (red) forecasts (upper panel). The lower panel displays the difference between the forecasts.

digital procedure is better, both scores are still relatively similar.

Figure 20 demonstrates how the difference between the scores of the conditional expectation for the two forecasting procedures develops as the number of recorded observations increase. The biggest difference between the two scores, at the end of the recording period, is 92.7. For clarity, we have only displayed scores for which $-16 < (S(DIG) - S(FREQ)) < 24$. The top panel displays the difference in cumulative scores after the $20th$ observation. As this is the first observation to be scored, it is hardly surprising that there is little to separate the cumulative scores at this stage. The second panel displays the difference between the cumulative scores after

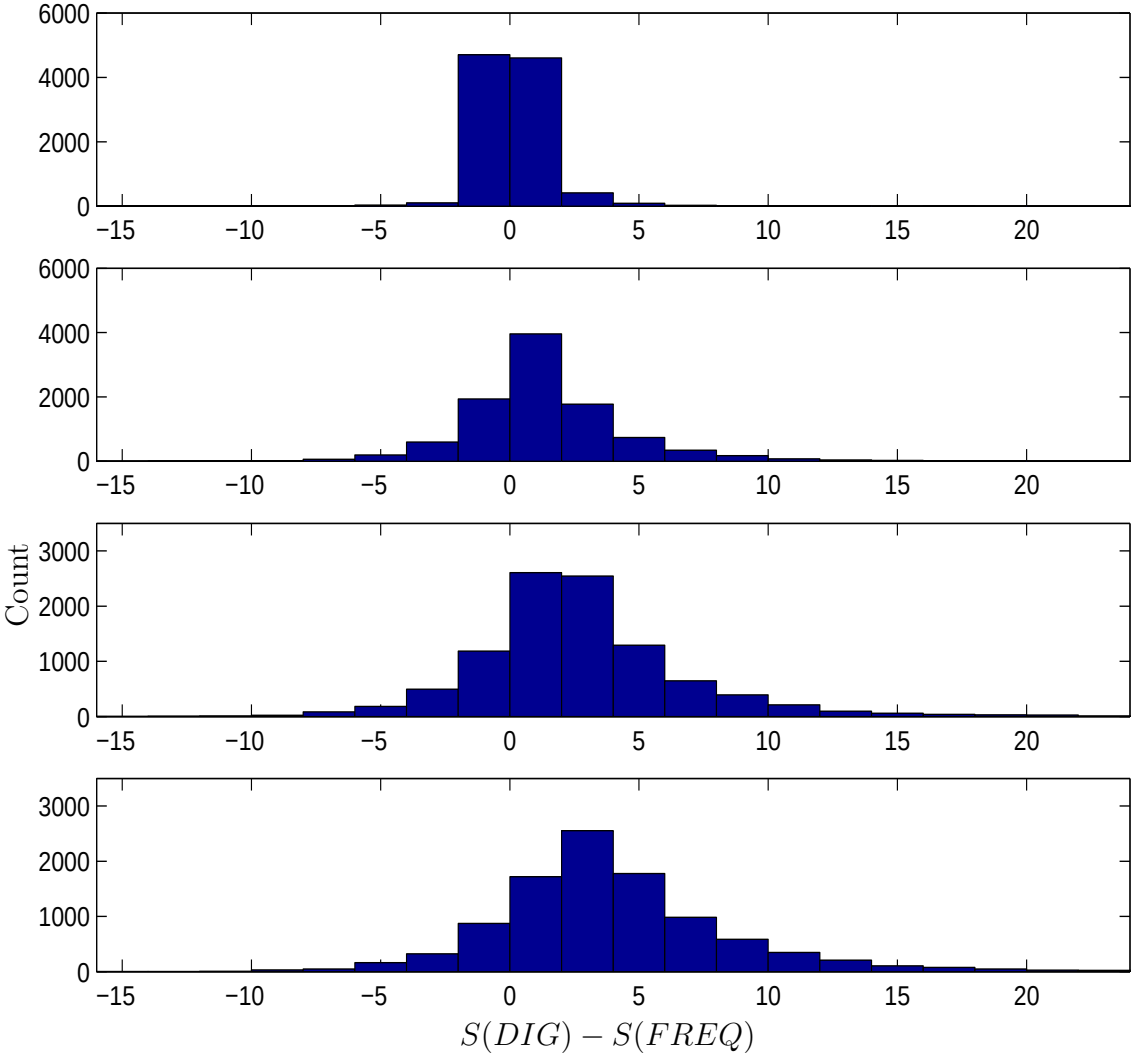

Figure 20: Difference between mean cumulative quadratic scores of the conditional expectation after 20, 30, 50 and 100 observations.

30 observations. We can see that the digital forecasts are beginning to have better scores. This trend continues in the third and forth panels, which correspond to the difference between the cumulative scores after 50 and 100 observations. As more observations are scored, the average difference between the two procedures continues to increase, reinforcing the conclusion the mixture mass forecast is increasingly better than the frequentist forecast

Figure 21 displays the difference between the median cumulative quadratic scores of the conditional variance for the simulated region. The first observation scored was the $20th$. Notice that on this occasion we consider the median cumulative score rather than the mean cumulative score. This is because, as Equation 14 shows, one

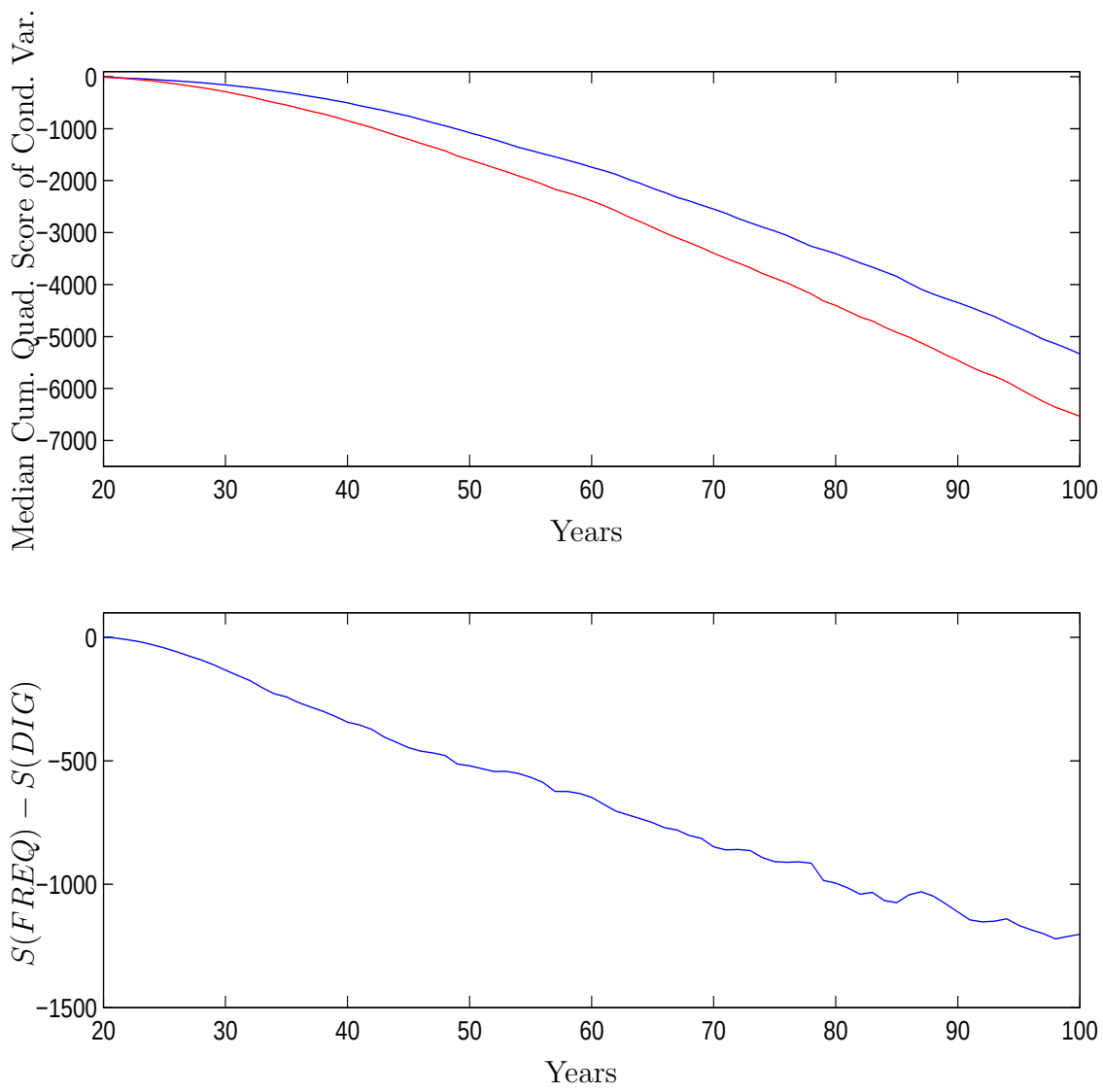

Figure 21: Sequential median cumulative quadratic score of the conditional variance for digital (blue) and frequentist (red) forecasts (upper panel). The lower panel displays the difference between the forecasts.

component of the variance of the GEV distribution is $\Gamma(1+2k)$. On rare occasions the L-moment estimate of the shape parameter, k, is very close to -0.5. When this happens the estimated value of $V(X)$ can become extremely large, since $\Gamma(\theta)$ increases rapidly as θ gets close to 0. Consequently, in this case it is more appropriate to compare the medians than the means.

The upper panel of Figure 21 shows that the digital forecast of the conditional variance scores significantly better than the frequentist forecast. The lower panel shows that the digital forecast continues to improve against the frequentist forecast. As with the Waimakariri River AMS data, both procedures score similarly for their

forecasts of conditional expectation, but the updated mixture mass functions are considerably better at assessing conditional variance.

5.4 Comparing L-moment Estimates and Updated Mixture Mass Forecasts using Frequentist Measures for Data Generated from a Generalised Extreme Value Distribution

In Section 3 we used conventional frequentist techniques to find point estimates of the characterising parameters of distributions commonly used in the study of flood frequency analysis. These point estimates were then used to estimate flood exceedance quantiles. Estimates were calculated for both at-site and regional cases. The appropriateness of the frequentist estimates was judged by comparing the bias (a measure of accuracy) and root mean-squared-error (a measure of precision) for a particular site. The site was the 11th site of a 21 site region. The region was simulated to have sites of varying length, skewness and CV. Data was generated using a GEV distribution. In Section 4 we constructed and implemented a digital updating procedure. This was used to score sequential forecasting methods for an analysis which involved sequences of observations regarded exchangeably. We conclude this Section by comparing frequentist estimates and digitally updated forecasts under objectivist criteria.

We shall compare four quantile approximation methods: two frequentist methods and two digital forecasting methods. A Monte Carlo method was used to compare the four procedures. The comparison consists of two parts: data generation and method testing. The data was generated according to the procedure specified in Section 3.7. The same four regions are used, each comprising 21 sites. Site record lengths range from 10 years, at site 1, to 30 years, at site 21. Data is generated from an EV2 distribution. The skewness of the data ranges linearly from -0.17 to -0.11 depending on the length of record. The median CV of sites in a region is set to be either 0.5 or 1.0. The normalised regional range in CV, $R^*(CV)$, is either 0.3 or 0.5. For this experiment 50,000 simulated regions were generated. See Table 4 for a summary of the regions used in the Monte Carlo experiments.

Forecasts of exceedance quantiles were obtained using four methods. Frequentist and digital measures were calculated using both at-site and regional procedures. The frequentist at-site method used is described in Section 3.5.1. The frequentist regional method used is the index flood method, which was described in Section 3.5.2. Digital forecasts are also computed at-site and regionally. The at-site digital forecast is found by updating mixture mass functions as sequential forecasting distributions, as detailed in Section 4.3. The procedure used to forecast regional digital exceedance forecasts is:

1. Normalise the data from each site in a region by the at-site mean.
2. Pool the normalised data and treat it as a series of observations from a super-site.
3. Calculate the marginal mass function from the observations in the super-site.
4. Forecast the at-site exceedance quantiles, using $f(k)$ as the prior mass function for k.

This is essentially a digital version of the hierarchical regionalisation method proposed in Section 3.5.2.

The digital forecasting procedures require that the realms of X, ξ, α and k, are specified before any computations can commence. The specified realms are listed in Table 7. The regional updating procedure requires that parameters are defined to have different realms at each of the two stages. This is due to the use of scaled data in the approximation of $f(k)$. $\mathbf{R}(k)$ is the same for both stages. As previously mentioned, scaling the data by its mean has no effect of the distribution's shape. The second stage sees a reversion to the realms used in the at-site updating procedure. This is because we are merely repeating the at-site procedure with an updated $f(k)$ mass function. It eventuates that the ranges of $\mathbf{R}(\xi)$ and $\mathbf{R}(\alpha)$ have little effect on the exceedance quantile forecasts. For both the at-site and regional updating procedure, the conditional mass functions, $f(\xi \mid \alpha, k)$ and $f(\alpha \mid k)$, and the marginal mass function $f(k)$, were defined to be Uniformally distributed over their realm.

$min(\mathbf{R}(k))$	-0.01	-0.03	-0.05	-0.07	-0.09	-0.11	-0.13
Q_{20}	5.31	5.34	5.39	5.46	5.54	5.65	5.77

Table 8: Estimates of $Q(0.95)$ obtained via the digital regional forecast, for different $\mathbf{R}(k)$. $(M(CV), R^*(CV)) = (0.5, 0.5)$. In each case $max(\mathbf{R}(k)) = -0.35$ and elements increment in steps of 0.02. The experiment is designed to generate data so $Q(0.95) = 5.67$

The at-site estimates should be the same for any $(M(CV), R^*(CV))$ combination containing the same $M(CV)$ value. For example, at-site estimates of combinations $(0.5, 0.5)$ and $(0.5, 0.3)$ should have the same bias and RMSE. This is because the only data used in an at-site estimate comes from the site itself, in this case Site 11, which is not effected by different $R^*(CV)$ measures. The data at Site 11 is generated from a GEV distribution with parameters $\xi = 1.08$, $\alpha = 1.25$ and $k = -0.14$. The differences in the regional results are due to the effect of heterogeneity in the region.

5.4.1 Results

The bias and RMSE for the 11th site of the 21 site region for the $(M(CV), R^*(CV))$ combinations $(0.5, 0.5)$ and $(0.5, 0.3)$ are displayed in Figure 22. The upper panel displays $(M(CV), R^*(CV)) = (0.5, 0.5)$ and the lower panel displays $(M(CV), R^*(CV)) = (0.5, 0.3)$.

The RMSE and bias estimates appear very similar for both different values of $R^*(CV)$. For both cases the digital at-site forecast appears to be the best of the four methods. Despite the sizeable bias values, the root-mean-squared error is considerably lower than for either of the frequentist estimates. The frequentist regional estimate procedure appears relatively unbiased but inefficient. A surprising result is the large negative bias shown by the regional digital forecast. On closer investigation the quantile forecasts obtained through the use of the regional digital procedure are highly dependent on the range of $\mathbf{R}(k)$. Table 8 demonstrates that the marginal mass function of k computed in the first stage of the regional procedure places more mass on the smaller elements of $\mathbf{R}(k)$ then expected. It would be interesting to see if this happens for all (ξ, α) values, or whether the estimation improves as ξ or α increase.

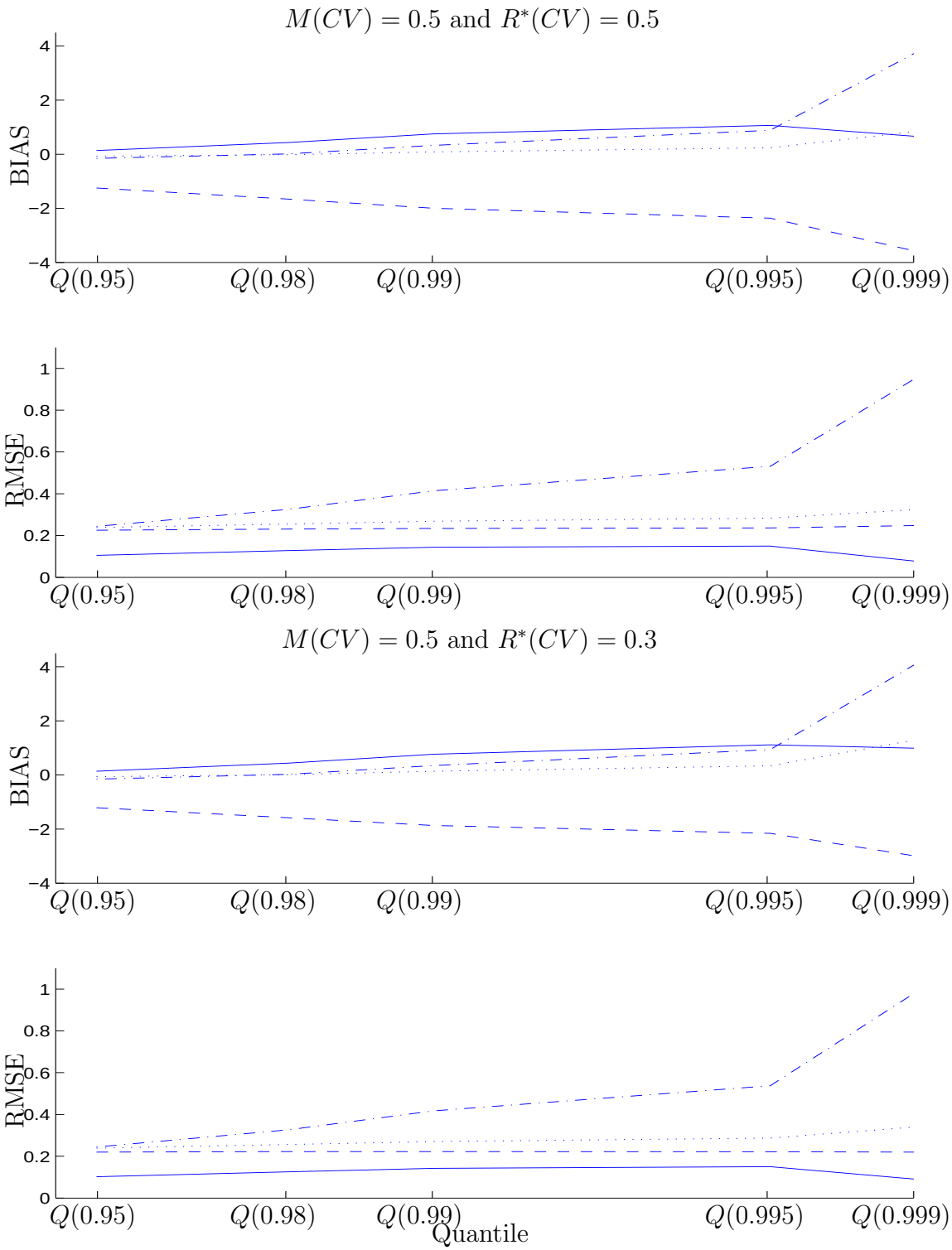

Figure 22: Root-mean-squared error and bias for site 11 of a 21 site region. Methods are digital at-site (—), frequentist at-site (-.-.), digital regional (–) and frequentist regional (...). The upper panel shows $(M(CV), R^*(CV)) = (0.5, 0.5)$, the lower panel shows $(0.5, 0.3)$.

The bias and RMSE for the 11th site of the 21 site region for the $(M(CV), R^*(CV))$ combinations, $(1, 0.5)$ and $(1, 0.3)$, are displayed in Figure 23. The upper panel displays $(M(CV), R^*(CV)) = (1, 0.5)$ and the lower panel displays $(M(CV), R^*(CV)) = (1, 0.3)$. For both combinations the digital at-site forecast has the lowest bias and lowest RMSE. One possible reason for the improvement of the digital forecasts, relative to the frequentist estimates, is that the distribution the data were generated from is more severely truncated when $M(CV) = 1$ then it is when $M(CV) = 0.5$, consequently the frequentist methods are attempting to fit parameters to a distribution using GEV L-moment estimates, when the distribution does not have a GEV shape at all.

The digital regional forecast still has a large negative bias, but the bias is relatively stable for different return periods, especially compared to the frequentist quantile estimates. In all four cases the digital regional forecast has the second lowest RMSE, suggesting that there could be some way of finding a better performing estimate. Possible alternative procedures for implementing a regional digital updating procedure would be:

- Scale all the data by its at-site mean. Treat these observations as coming from a 'super-site'. Calculate quantile forecasts and rescale.

- Calculate mass functions $f(\xi \mid \alpha, k)$, $f(\alpha \mid k)$ and $f(k)$ at sites 1–10 and 12–21. Combine these mass functions using a weighted average. Consider the new combined mass functions as prior mass functions and conduct the at-site procedure at site 11.

A considerable advantage that the digital procedure has over frequentist measures is that the digital procedure can begin in the first period, and advance sequentially as data is recorded. Conversely, the frequentist estimates rely on gathering a sizeable data set before quantile estimates can be calculated, hence the motivation of regionalisation techniques.

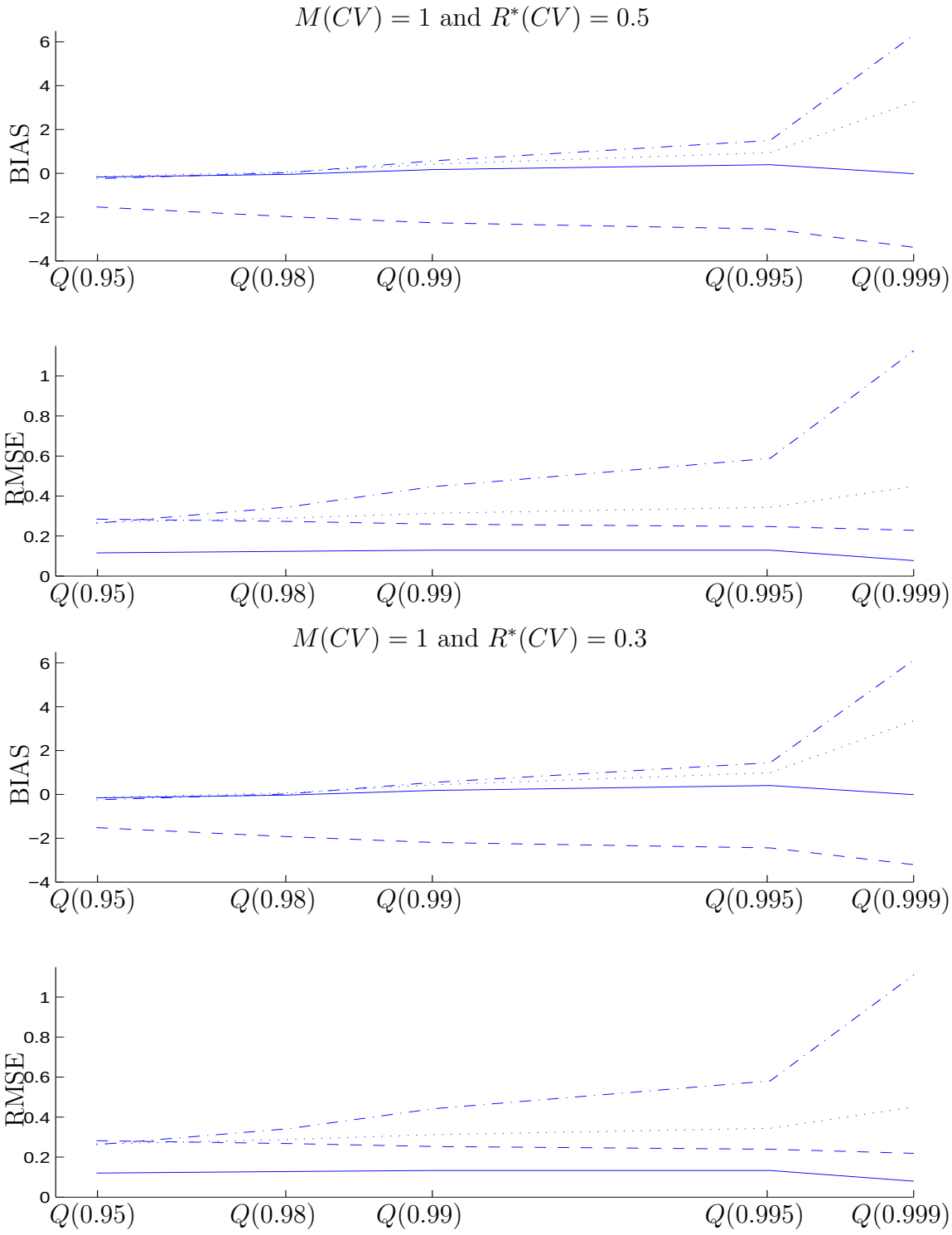

Figure 23: Root-mean-squared error and bias for site 11 of a 21 site region. Methods are digital at-site (—), frequentist at-site (-.-.), digital regional (–) and frequentist regional (...). The upper panel shows $(M(CV), R^*(CV)) = (1, 0.5)$, the lower panel shows $(1, 0.3)$.

6 Summary

This Report has focussed on flood frequency analysis, and in particular on the estimation of flood quantile levels. Our problem of characterising extreme floods was introduced in Section 2. We described river flow measurement procedures, paying particular interest to the Waimakariri River. In Section 3, conventional frequentist estimates were calculated for the Waimakariri River. A Monte Carlo procedure was used to compare different estimation methods in terms of accuracy and precision, for an experimental data set. In Section 4 a procedure for scoring sequential forecasts using digitised mass functions was developed. This procedure is based on the work developed in Ware and Lad (2003). Different scoring rules were examined using a sequence of annual maximum river flows from the Waimakariri River. Finally, we compared the appropriateness of the frequentist and digital procedures using both subjective and objective techniques. The scores of conditional expectations of both procedures were similar, but the score of the conditional variance was much better for the updated mixture forecasts. When objectivist measures were considered, the mixture distributions computed via the discrete digital method provide forecasts with lower root mean-squared-error. This is despite the fact that when the coefficient of variation is small the digital methods are more biased than the frequentist methods. As the coefficient of variation increases, the accuracy of the digital methods improves rapidly, even reducing the bias.

Acknowledgements

This work was undertaken as part of Robert Ware's doctoral thesis. He was partially supported by the University of Canterbury Keith Laugesen Scholarship and a Mathematics and Statistics Departmental Scholarship. Thank you to Doris Barnard, Alistair Smith and Charles Pearson for many helpful comments.

References

Connell, R. J. and Pearson, C. P. (2001). Two-component extreme value distribution applied to Canterbury annual maximum flood peaks. *Journal of Hydrology (NZ)*, 40(2):105–127.

Dalrymple, T. (1960). Flood Frequency Analysis. Technical Report Water Supply Pap., 1543-A, U.S. Geol. Surv.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting Forms of the Frequency Distribution of the Largest r Smallest Member of a Sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190.

Gabriele, S. and Arnell, N. (1991). A Hierarchical Approach to Regional Flood Frequency Analysis. *Water Resour. Res.*, 27(6):1281–1289.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Expressed in Inverse Form. *Water Resour. Res.*, 15:1049–1054.

Greis, N. P. and Wood, E. F. (1981). Regional Flood Frequency Estimation and Network Design. *Water Resour. Res.*, 17:1167–1177.

Hosking, J. R. M. (1990). L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. *J. R. Statist. Soc. B*, 52(1):105–124.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and Quantile Estimation for the Generalised Pareto Distribution. *Technometrics*, 29:339–348.

Hosking, J. R. M. and Wallis, J. R. (1993). Some Statistics Useful in Regional Frequency Analysis. *Water Resour. Res.*, 29(2):271–281.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics*, 27:251–261.

Jenkinson, A. F. (1955). The Frequency Distribution of the Annual Maximum (or Minimum) of Meteorological Elements. *Journal of the Royal Meteorological Society*, 81:158–171.

Kjeldsen, T. R., Smithers, J. C., and Schulze, R. E. (2002). Regional flood frequency analysis in the Kagouls – Natal province, South Africa, using the index flood method. *J. Hydroa.*, 255:194–211.

Kroll, C. N. and Vogel, R. M. (2002). Probability distribution of low streamflow series in the United States. *Journal of Hydrologic Engineering*, 7:137–146.

Lad, F. (1996). *Operational Subjective Statistical Methods. A Mathematical, Philosophical and Historical Introduction*. Wiley-Interscience.

Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability Weighted Moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resour. Res.*, 15:1055–1064.

Lettenmaier, D. and Potter, K. W. (1985). Testing Flood Frequency Estimation Methods using a Regional Flood Generation Model. *Water Resour. Res.*, 21(12):1903–1914.

Lettenmaier, D. P., Wallis, J. R., and Wood, E. F. (1987). Effect of Regional Heterogeneity on Flood Frequency Analysis. *Water Resour. Res.*, 23(2):313–323.

Madsen, H., Pearson, C. P., and Rosbjerg, D. (1997). Comparison of Annual Maximum Series and Partial Duration Series Methods for Modeling Extreme Hydrologic Events. 2 Regional Modeling. *Water Resour. Res.*, 33(4):759–769.

McKerchar, A. I. and Pearson, C. P. (1990). Maps of Flood Statistics for Regional Flood Frequency Analysis in New Zealand. *Hydrological Sciences Journal*, 35(6):609–621.

Metcalfe, A. V. (1997). *Statistics in Civil Engineering*. Arnold.

Mosley, M. P. (1981). Delimitation of New Zealand Hydrologic Regions. *J. Hydrol.*, 49:173–192.

National Environment Research Council (1975). *Flood Studies Report, vol. 1.* London, U.K.

Park, J. S., Jung, H. S., Kim, R. S., and Oh, J. H. (2001). Modelling summer extreme rainfall over the Korean peninsula using Wakeby distribution. *Int. J. Climatol.*, 6(5):1371–1384.

Pearson, C. P. (1991). Regional Flood Frequency Analysis for Small New Zealand Basins. *Journal of Hydrology (N.Z.)*, 30(2):77–92.

Pearson, C. P. (1993). Application of L Moments to Maximum River Flows. *New Zealand Statistician*, 28(1):2–10.

Smith, J. A. (1989). Regional Flood Frequency Analysis Using Extreme Order Statistics of the Annual Peak Record. *Water Resour. Res.*, 25:313–317.

Stedinger, J. R. (1983). Estimating a regional flood frequency distribution. *Water Resour. Res.*, 19(2):503–510.

Stedinger, J. R. and Lu, L. H. (1995). Appraisal of Regional and Index Flood Quantile Estimators. *Stochastic Hydrology and Hydraulics*, 9(1):49–75.

Walter, K. (2000). Index to hydrological recording sites in New Zealand. Technical Report 73, NIWA, Wellington, N.Z.

Ware, R. and Lad, F. (2003). Approximation of Posterior Means and Variances of the Digitised Normal Distribution using Continuous Normal Approximation. Technical Report UCDMS2003/16, Department of Mathematics and Statistics, Univeristy of Canterbury, Christchurch, N.Z.