Journal of Hydrology (NZ) 42 (1): 75-95, 2003

Reliability of linear regression for estimation of mean annual low flow: a Monte Carlo approach

R. D. Henderson, R. P. Ibbitt and A. I. McKerchar

National Institute of Water and Atmospheric Research Ltd (NIWA), P O Box 8602, Christchurch, New Zealand

Abstract

To estimate the mean annual low flow, or other statistics on more extreme low flows, a commonly used technique is ordinary least-squares linear regression using small samples of concurrent flows, between a flow recorder and secondary sites with occasional gaugings. The reliability of this technique is seldom tested or even discussed. Furthermore, the inadvisability of using ordinary least-squares when both variables in the regression are subject to natural variability and measurement errors is seldom recognised. Monte Carlo simulation using daily mean flow pairs from recorder sites has been used to assess the uncertainty of such regression procedures. Trials were done with a range of sample sizes typically used in practice. The sampling scheme is described, including sample size, temporal separation, recession length and flow filtering. For each of many samples, two linear regression methods were applied: ordinary least-squares and geometric mean regression. There are sound statistical arguments for the use of geometric mean regression when there are errors in both variables. The predicted value of the mean annual low flow from both regression methods was compared with the "known" value from analysis of annual low flows. For the site pairs sampled, the standard error of prediction of the mean annual low flow is ± 40% (95% confidence).

Keywords

Ordinary-least-squares regression, geometric mean regression, mean annual low flow.

Introduction

Estimation of low flow statistics is vital for water resource management in New Zealand. In general, for larger rivers, there is an adequate network of flow recorders with historical data that allows estimation of average annual and more extreme low flow statistics, and thus provides information for evaluating consent applications for large water resource schemes. However, for the myriad applications for extraction from small rivers and streams, usually in rural areas, the flow recording network will be at best "nearby". Methods of estimating flows within and between monitored catchments are needed, and knowledge of the reliability of estimation is critical when results are used to restrict water use and hence economic activity.

New Zealand's rivers have been gauged for the last one hundred years. Early gauging efforts concentrated on rivers with a potential for hydro-power development, irrigation and water supply. Flow recorder sites were established on many of the rivers where the earliest gaugings occurred, and these provide the long flow series commonly used to investigate long-term trends and other types of variability. With these long series of flow data, reliable design estimates of flood

magnitude, water yield and low flow can be made, as long as gauging is maintained for rating curve definition.

Beginning in the late 1950s and continuing to the present day is another, parallel, gauging effort—to measure the flow of many smaller streams during times of normal to low flow. The reasons for these gaugings have varied from a desire to characterise the water resources of a region, to the need to understand drought impacts and the drought risk to consumptive users and the environment. The legacy of these efforts is a huge collection of gauging cards held in offices and archive stores around New Zealand, a large number of water resource assessment reports written by Ministry of Works, Catchment Board and Regional Council staff, and a handful of papers describing the low flow characteristics of particular areas, e.g. Grant (1968), Waugh (1970a and b), Grant (1971), Whitehouse *et al.* (1983), Harrison (1988), and Caruso (2000).

National studies using flow records and catchment physical characteristics provide one means of estimating low flow yields at unmeasured sites. Hutchinson (1990) used regression equations based on different regions, and Pearson (1995) used both a regression equation and contours of specific yield. Tests with new flow data show these techniques give results that can vary by up to an order of magnitude from estimates made from analysis of flow data (Fig. 1). This large range of uncertainty has led to the development of new ideas for a low flow estimation model, and also to investigations of the reliability of other techniques for estimating low flows. A common alternative to either regression on catchment physical characteristics or mapping techniques is the use of linear regression between low flow gaugings collected at sites without flow recorders, and the recorded flow at a nearby flow recorder. If these many "miscellaneous" flow gaugings can be used to estimate low flow statistics, then many more data points will be available for calibrating and validating low flow models. The uncertainty of estimates of a statistic such as the mean annual low flow derived from regression with flow measured at a primary site is the subject of this paper.

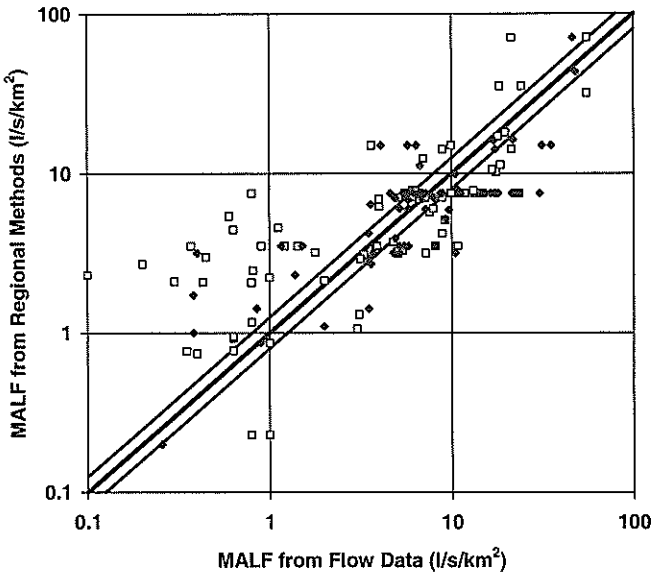

Figure 1 – Comparison of specific estimates of mean annual low flow (MALF) from flow data and from the methods of Hutchinson (1990) and Pearson (1995). Filled diamonds are estimates of mean annual low flow; open squares are estimates of 5-year low flow. A 1:1 line and bounds of ±25% are shown, as this may be considered a desirable level of uncertainty by practising hydrologists. Sites are from comparisons by Hutchinson (1990), McKerchar and Henderson (1999), Caruso (2000) and assessments made by staff of three councils: Wellington Region, Tasman District, and Waikato Region. Flows estimated by regional methods are sometimes banded due to the methods' coarse resolution.

In a recent review paper Smakhtin (2001) discusses many aspects of low flow hydrology, including methods of estimating low flows in ungauged catchments, but neither of the two arbitrary classifications of techniques include use of concurrent gaugings regressed with flow at a primary site. In the United States many workers report on estimation of low flow statistics by regression equations using catchment physical characteristics, e.g. Parrett and Cartier (1990), Hayes (1991), and Giese and Mason (1993).

Stedinger and Thomas (1985) and Thomas and Stedinger (1991) detail problems of bias when applying ordinary least-squares to correlation of concurrent flow measurements. They evaluate several techniques for estimating regression parameters to arrive at their recommended method, which involves the relative standard deviations of the annual low flow series and the one-day sample from the primary site used. The focus of the method is the derivation of estimates for the ungauged site over a range of return periods, and properties of the assumed frequency distribution are assigned to the secondary site. Wilson (2000) evaluates the Stedinger and Thomas method for Indiana streams and finds that standard errors of both two-year and ten-year low flows are predicted within 15%. Critical features of the method's application are the "similarity" of the two sites, and whether or not the two sites were on the same stream (nested catchments) or on different streams.

Potter (2001) presents a simple method for estimating low flows using the product of the primary site parameter and the geometric mean of the ratios of a small number of concurrent observations. The results using two site pairs are better for mean flow than for low flow. A major difficulty in applying these techniques is the impossibility of testing assumptions about similarities of flow distribution and frequency statistics between sites.

Between-site correlation with linear regression is also a common technique in New Zealand, e.g. Bowden (1974), Roke (1979), de Joux (1980), Rae (1988) and Horrell (2001). That ordinary least-squares is used may be assumed but is rarely explicitly stated.

An assumption when using ordinary least-squares is that there is no uncertainty in the independent variable, as the differences are minimised in the dependent variable only. For the regression of one measured flow against another (concurrent gaugings) or a rated flow against a measured flow, there is clearly measurement error in both variables.

Statistics on estimates of low flow have the largest relative errors of any hydrological statistic in common use (Thomas and Benson, 1970). Sources of measurement error include the following:

- Both continuously recorded flows and gaugings may be at sites where flow has been abstracted upstream or enhanced by upstream activities. In many parts of New Zealand, these activities have resource consents, but there is little monitoring. Natural stream flows must then be estimated, commonly by applying rules of thumb to the maximum consented abstraction rate, and the impact of these assumptions on bias or error magnitude is unknown. At low flow the adjustments to calculate natural flow can be half of the total flow or more.
- Flow gaugings are subject to error (ISO, 1978). For gaugings to establish and maintain ratings, there are procedures to ensure that the 95% confidence interval half-width is kept below ±10% (typically ±8%). When water resources assessment gaugings are done at sites selected not for gauging standards but for quick access, errors may be larger. However, the likely magnitude of these can be calculated.
- Continuous flow records are subject to level errors and, more importantly, to rating errors. At low flow, rating errors at

sites with natural controls can be large because of infrequent low flow gauging.

- While the mean annual low flow may be calculated without recourse to a distribution, it is nonetheless the mean of the lowest and most error-prone values from the continuous series. To derive statistics on more extreme (rarer) low flows, a low-flow frequency distribution is chosen and fitted, and both these steps are subject to error.

Riggs *et al.* (1978) explain the reasons for not using ordinary least-squares when there are errors in both variables. A consequence of violating the least-squares requirement that there be no errors in the independent variable is that the estimated slope of the regression is biased downward (closer to zero). When the mean annual low flow is below the mean of the data sample, as is commonly the case, the result of using least-squares regression is an overestimate of the mean annual low flow at the secondary site. They show that the geometric mean regression estimate of regression slope is essentially unbiased when the variances of the measurement errors in each variable are similar. Even when this condition does not hold, the slope bias is less than that for ordinary least-squares. They also found geometric mean regression to be superior when the coefficient of determination (r^2) is small. Other authors have also recommended geometric mean regression for water resource applications (Helsel and Hirsch, 1992; Reckhow and Chapra, 1983).

Whereas ordinary least-squares regression minimises the sum of squares of the vertical (Y on X) or horizontal (X on Y) distances of each point from the regression line, geometric mean regression minimises the sum of the products of the vertical and horizontal distances of each point from the regression line. Let the equation of the ordinary least-squares lines be $Y = a + bX$ (Y on X) and $X = c + dY$ (X on Y) and the equation of the geometric mean regression line be $Y = u + vX$. Then the slope of the geometric mean regression line is:

$$v = \pm\sqrt{\frac{\sum y^2}{\sum x^2}} = \pm\sqrt{\frac{b}{d}} = \pm\frac{s_y}{s_x} = \pm\frac{b}{r}$$

where s_y and s_x are the standard deviations of Y and X respectively, and r is the linear correlation coefficient between X and Y. From these formulae it can be seen that there are no cross products of standard deviation involved in the geometric mean regression.

To estimate mean annual low flow by regression of concurrent flow measurements, we decided to assess both ordinary least-squares and geometric mean regression by using Monte Carlo sampling of flow data from long continuous flow records. Firstly, this will provide an analysis of the errors inherent in either technique, and secondly, allow them to be compared so that the effects of choosing one or the other may be assessed.

Methods

Individual estimates of mean annual low flow at ungauged locations, or maps of a region showing the expected yield of catchments at the mean annual low flow, are commonly required by regional council offices. Typically there are collections of gaugings for a number of widely distributed locations. For a single location there may be of the order of 5 to 10 gaugings, but sometimes for an important site, more than 20. To estimate mean annual low flow at one location (secondary site), or produce a map, the transfer of information from long-term flow records (primary sites) to sites with gaugings is the first step. Regression is a common tool, but needs to be applied with care. Typically a primary site is selected by an experienced hydrologist, on the basis of data availability and hydrological similarity to the secondary site. Flows recorded at the primary

site are regressed against the gauged flows at the secondary site, usually using ordinary least-squares regression. The regression is then used to estimate a flow statistic, for instance mean annual low flow, at the secondary site. Errors are not generally calculated, but an r^2 value is often quoted as evidence of goodness of fit.

Issues in assessing the "goodness" of the estimated flows at the secondary site include:

- What range of flows should be included in the regression analysis? Clearly if too wide a range of flows is used, the regression analysis may be influenced by factors that have nothing to do with the control of low flows.
- How many pairs of concurrent gaugings are needed to produce estimates of the design low flow statistic with a given standard error of prediction, or conversely what will be the standard error of prediction for a given number of gaugings?
- Should zero flows be included if they occur frequently at one site but not the other, and if so, how? Conversely what would be the impact of omitting zero flows from the analysis?
- Assuming hydrological regions can be identified, what is the impact of using a primary site that is in a different hydrological region than the secondary site?

Monte Carlo sampling

We have used Monte Carlo sampling to sample flow pairs from continuous records, in order to simulate the result of concurrent low flow gaugings, so that we could assess the performance of both ordinary least-squares and geometric mean regression techniques. Monte Carlo sampling is the choice of random samples from a population in order to establish the behaviour of a particular statistic. When the population is assumed known, the effect of varying chosen parameters can be studied.

The sampling program selected from pairs of sites having continuous time series of river flow, and stored in the Water Resources Archive (WRA, New Zealand's national archive of river flow data; Walter, 2000). At least ten years of overlapping time series were necessary so that the low flow range would be reasonably well sampled. During periods when both site records had data, daily mean flows were calculated from the raw 15-minute data. Flow pairs were selected from the daily average time series only if they satisfied the following criteria:

- Flow at both sites fell below a threshold set at the lower quartile of recorded flows.
- Flow at both sites was greater than or equal to zero (zero flows were omitted in some trials).
- Flow at both sites each day was not larger than flow the previous day.
- Flow at both sites was from a recession of four days or longer.

The selected flow pairs thus contained all common recessions from the two sites, and formed the basis for the Monte Carlo sampling routine. Figure 2 shows a set of flow pairs from two sites in the same hydrological region, as selected from continuous time series data, and Figure 3 shows flow pairs from two sites in different hydrological regions (Toebes and Palmer, 1969).

Samples were taken from the selected set at random, but subject to the constraint that there should be at least five days between flow pairs. This requirement reduces temporal correlation of samples, and is similar to the field practice of gauging runs, which are often separated by a week or two. Where possible, 10,000 samples were taken from the selected set for all values of N where $3 \leq N \leq 20$ flow pairs. These limits on N were set because three flow pairs are necessary to perform a linear regression, and more than 90% of secondary flow sites have 20 gaugings or less. One site was chosen as the primary site, where the mean annual low flow is known,

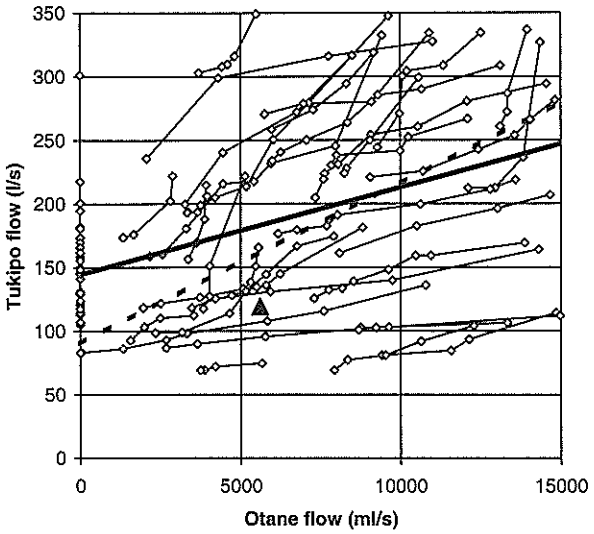

Figure 2 – Flow pairs between Otane and Tukipo (both Napier hydrological region). Lines join the flow pairs on individual recessions. Their mean annual low flow from flow data analysis is shown as the solid triangle. Ordinary least-squares regression line on all data is the solid line, and geometric mean regression on all data is the dashed line.

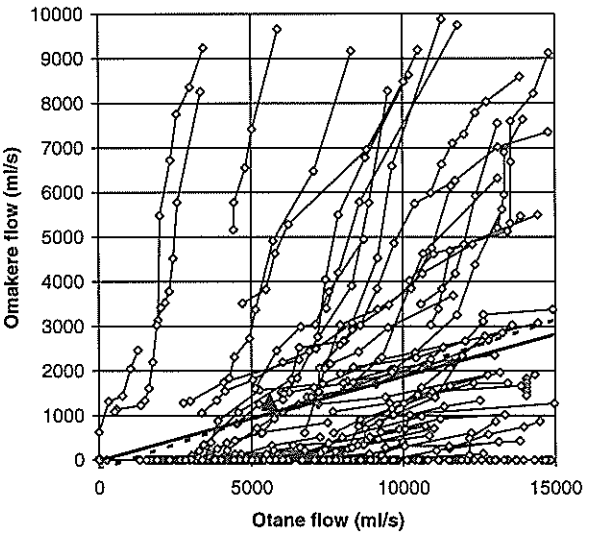

Figure 3 – Flow pairs between Otane and Omakere (Southern Hawke's Bay hydrological region). Lines join the flow pairs on individual recessions. Their mean annual low flow from flow data analysis is shown as the solid triangle. Ordinary least-squares regression line on all data is the solid line, and geometric mean regression on all data is the dashed line.

and the other was nominated as the secondary site, where the mean annual low flow is to be estimated by regression. For the test sites, the value of the mean annual low flow at the secondary site is actually known from analysis of annual low flows, which enables us to calculate the bias of the two linear regression estimates of the mean annual low flow.

Ordinary least-square and geometric mean regressions

For each flow pair sample:

- Two regression lines of the form $Y = aX + b$ were fitted, one using ordinary least-squares and the other using geometric mean regression. X indicates flows at the primary site, and Y indicates flows at the secondary site.
- The primary site mean annual low flow was calculated by averaging the series of lowest daily mean flow in each year of record.
- The secondary site mean annual low flow (MALF) was estimated by substituting the primary site mean annual low flow into each regression line $\hat{Y}_{MALF} = \hat{a}X_{MALF} + \hat{b}$, where $\hat{}$ indicates an estimate from a flow pair sample.
- Estimates were rejected if the slope of the regression line $\hat{a}$ was negative.
- Estimates were rejected if the estimated value $\hat{Y}_{MALF}$ was greater than the flow threshold for flow pair selection. This condition was necessary to reject very large values produced by regression lines with slopes close to infinity.
- If the mean annual low flow estimate was a negative flow, zero was substituted. This is a practical response to the field situation, where a negative answer implies that the

secondary site dries up more quickly than the primary site. This is common when secondary sites are on smaller streams than the primary sites.

Samples rejected for negative slope were on average 12%, 4%, 2.5% and 0.7% of the total samples for N = 5, 10, 15 and 20 respectively. An insignificant number of pairs were rejected for excessive size, and 1% or less had mean annual low flow estimates that were below zero reset to zero.

The variance of the estimated $\hat{Y}_{MALF}$ was calculated for ordinary least-squares using:

$$s^2_{\hat{Y}_{MALF}} = s^2_{y.x}\left(1+\frac{1}{N}+\frac{(X_{MALF}-\overline{X})^2}{(N-1)s^2_x}\right) \qquad (1)$$

where $\overline{X}$ is the mean of the sample X values used in the regression, N is the number of sample pairs, s^2_x is the variance of the sample X values and $s_{y.x}$ is the standard error of regression. See Dixon and Massey (1957) for full details.

The variance of the estimated $\hat{Y}_{MALF}$ was calculated for the geometric mean regression using:

$$s^2_{\hat{Y}_{MALF}} = \frac{\Sigma(Y-\overline{Y})^2(1-r^2)}{N-1} + \hat{\lambda}^2(1-r)^2(X_{MALF}-\overline{X})^2 \quad (2)$$

where r is the correlation coefficient, and $\hat{\lambda}^2$ is the ratio $\dfrac{\Sigma(Y-\overline{Y})^2}{\Sigma(X-\overline{X})^2}$.

Reckhow and Chapra (1983) give this version of an equation in Ricker (1973), who in turn modified the expression of Teissier (1948).

Equations 1 and 2 were used to calculate the prediction limits (±1 standard error) of the mean annual low flow at the secondary site for each sample regression line. One standard error was chosen because of problems involving the sampling distribution; these will be discussed below. Figure 4 illustrates these limits. For all sample sizes $3 \leq N \leq 20$, the mean values of the mean annual low flow estimate and the prediction limits from all samples, for each number of flow pairs, were recorded.

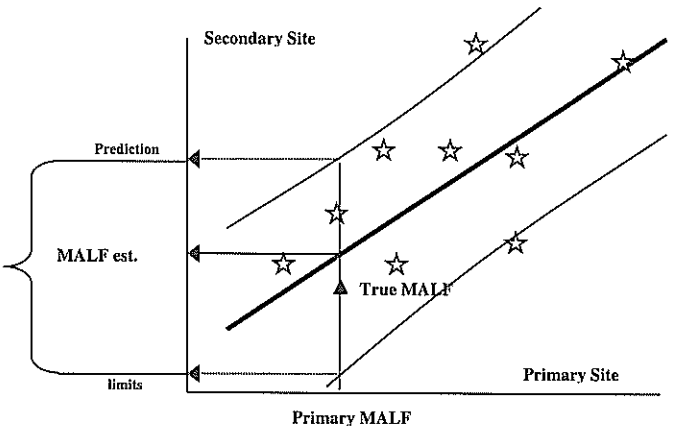

Figure 4 – Diagram illustrating regression estimation of mean annual low flow (MALF) at a secondary site from a primary site, and associated errors. The stars represent a single sample of nine gaugings from the set of selected flow pairs. The filled triangle (▲) represents the true mean annual low flow of each site. The central arrow is the mean annual low flow estimate from the regression line. The ±1 s.e. prediction limits are represented by the bracketed arrows.

Correction factor to geometric mean error term

In early sampling runs we found that the estimated geometric mean regression standard error calculated using equation 2 was smaller at small sample values of N than at larger sample values. We may be using a smaller number of samples than is desirable, but we believe there is a need to evaluate regression techniques for low flow gauging data transfer where $N < 10$ because this is a typical number of gaugings found in practice.

This reduction in standard error with small N is counter-intuitive, and indeed standard error increases in general if N is reduced by removing data points from a single sample. We speculate that this phenomenon is induced by our random sampling strategy, but we do not currently have a satisfactory explanation.

Assuming that the geometric mean regression and ordinary least-squares equations for variance should give similar results at the sample mean ($X = \overline{X}$), we find that the equations are identical, except for a scale factor of $\frac{(N-1)(N+1)}{(N-2)N}$. When we apply this factor to equation 2 we get errors of estimate that increase with smaller sample size N as expected (as will be shown on Figure 7). See the appendix for a derivation of this factor. Further aspects of the prediction variances will be discussed below.

Site pair selection by cluster analysis

To apply regression techniques in the field with some success, it is normal to select catchments with similar physical characteristics that are subject to the same weather patterns. To mimic this requirement in the Monte Carlo simulations, we conducted a cluster analysis of catchments with flow records. Catchments were taken from the dataset analysed by Pearson (1995), for which physical characteristics were readily available.

Variables selected were:

- East and North co-ordinates in the New Zealand Map Grid system, to avoid coincidental correspondences over large distances where the pairs would not be subject to the same weather events.
- The following parameters from the dataset of Pearson (1995): basin area, basin average rainfall, depth-weighted soil porosity, elevation, slope, and indices of erosion, hydrogeology and vegetation.

For each island of New Zealand, each variable was normalised by subtracting the mean value and dividing by the standard deviation. Site pairs with the smallest Euclidean distance were selected. Small catchments were preferred because regression techniques are commonly used to estimate mean annual low flow for small catchments without flow data. Flow data availability, and amount of flow data overlap in time between pairs were also checked. These requirements resulted in a very small set of site pairs, so the criteria for catchment size were relaxed, and additional pairs selected by expert opinion. The final set of 21 primary-secondary pairs (Table 1) showed a reasonable degree of correlation during regressions. We believe that the selection process gave a better chance of small scatter in relationships between flows than if site pairs had been selected only on their proximity or at random.

Results

Application to three sites

The method is first illustrated with data from 3 sites with continuous records (Walter, 2000):

- Otane at Glendon, site 23209, area 24.3 km^2
- Omakere at Fordale, site 23210, area 54.4 km^2
- Tukipo at SH50 (Punawai), site 23220, area 86.5 km^2

Table 1 – Site pairs selected using cluster analysis.

Case	Primary site		Secondary site		Records used yymmdd		7 day MALF from record (l/s)	
	number	name	number	name	from	to	Primary	Secondary
1	5513	Glenbervie at Quarry	5515	Glenbervie at Pines	800101	861231	3.89	0.711
2	5513	Glenbervie at Quarry	5516	Glenbervie at Log Bridge	800101	861231	3.89	1.01
3	11605	Mahakirau at E309 Rd	12509	Wharekawa at Adams Farm	920101	961231	232	307
4	30802	Pauatahanui at Gorge	30516	Mill Ck at Papanui	760101	931231	90.8	12.3
5	30802	Pauatahanui at Gorge	30701	Porirua at Town Centre	760101	931231	90.8	141
6	38401	Timaru at SH45	38501	Oakura at Surrey Hill Road	810101	861231	388	571
7	38401	Timaru at SH45	41301	Marokopa at Falls	810101	861231	388	1490
8	38401	Timaru at SH45	41302	Tawarau at Te Anga	810101	861231	388	2010
9	1143427	Te Tahi at Puketotara	1643461	Kaniwhaniwha at Limeworks Road	870101	941031	60	345
10	1143427	Te Tahi at Puketotara	1943481	Waitomo at Ruakuri Caves Bridge	870101	941031	60	350
11	1643457	Whakapipi at SH22-Tuakau	1643456	Whakapipi at Harrisville Road	860101	941231	97.8	33.7
12	1643457	Whakapipi at SH22-Tuakau	1843412	Pokeno at McDonalds Road	860101	941231	97.8	27.7
13	1643457	Whakapipi at SH22-Tuakau	1643460	Clarkes Rd Stm at Clarkes Rd	860101	941231	97.8	12.3
14	57008	Motueka at Gorge	57022	Hunters at Weir	780101	911231	1610	2.33
15	93211	Matakitaki at Mud Lake	93212	Mangles at Gorge	640101	891231	21600	2130
16	66210	Ashley at Lees Valley	65901	Waipara at White Gorge	900101	971231	855	111
17	66210	Ashley at Lees Valley	66213	Okuku at Fox Ck	900101	971231	855	592
18	66210	Ashley at Lees Valley	68001	Selwyn at Whitecliffs	900101	971231	855	814
19	74368	Elbow Ck at Muster Huts	74367	Deep Ck at Muster Huts	800101	931231	7.62	11.2
20	74368	Elbow Ck at Muster Huts	74369	Poisonous Ck at Lammermoor	800101	851231	7.62	3.88
21	74368	Elbow Ck at Muster Huts	74318	Taieri at Canadian Flat	830101	951231	7.62	1080
Az (with zeroes)	23209	Otane at Glendon	23220	Tukipo at SH50	600101	1000701	5.6	119
Bz (with zeroes)	23209	Otane at Glendon	23210	Omakere at Fordale	600101	1000701	5.6	1.296
Anz (without zeroes)	23209	Otane at Glendon	23220	Tukipo at SH50	600101	1000701	5.6	119
Bnz (without zeroes)	23209	Otane at Glendon	23210	Omakere at Fordale	600101	1000701	5.6	1.296

The Otane and Tukipo basins lie in the Napier hydrological region, while Omakere lies in the adjacent Southern Hawkes Bay region. Toebes and Palmer (1969) define the hydrological regions of New Zealand, while Toebes and Morrissey (1970) describe the characteristics of the Otane and Omakere basins. The main difference between the Napier and Southern Hawke's Bay regions is in their geology. In the Napier region sands, silts and gravels predominate. In the Southern Hawke's Bay region the geology comprises siltstone, sandstone, mudstone, limestone and conglomerates. For the Omakere site there is zero flow for more than 10% of the duration of the record (1963-2000). However, for the Otane and Tukipo sites low flows are more sustained and have few zero values. On the basis of the shapes of the flow duration curves (Fig. 5) one might expect better estimates of the design flow statistic for Tukipo than for Omakere, when Otane is used as the primary site.

Figures 6 (ordinary least-squares) and 7 (geometric mean regression) show the results of applying the sampling and regression technique to pairs from Otane and Tukipo (case Az in Table 1). The biases and errors of the predicted mean annual low flow at Tukipo for geometric mean regression and ordinary least-squares, for sample sizes of 5 and 10 flow pairs, are listed in Tables 2 and 3. Both error and bias are approximately 30% to 50%, with a slight reduction in error at ten samples. However when Otane is used as the primary site to predict the mean annual low flow at Omakere (case Bz in Table 1), the biases become

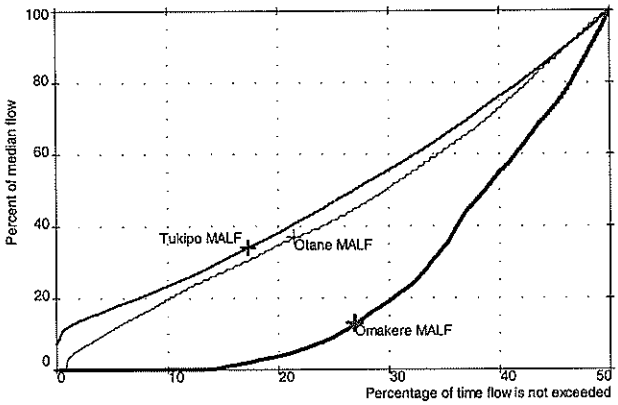

Figure 5 – Flow distribution curves for the flows below median in the Otane, Tukipo and Omakere catchments, scaled to equal 100 at their respective medians (50, 734 and 106 l/s). Otane and Tukipo curves show similar scaling with mean annual low flow (MALF) near 35% of the median flow compared to the Omakere curve, where mean annual low flow is 13% of the median. For the Omakere site flow is zero for more than 10% of the time.

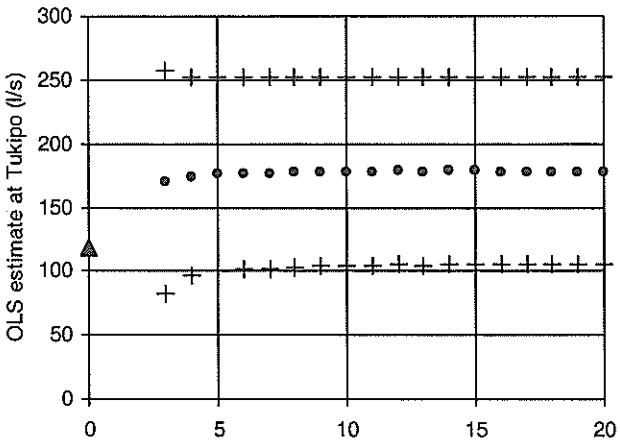

Figure 6 – Ordinary least-squares (OLS) estimates of mean annual low flow at Tukipo derived from flows at Otane, indicated by filled circles. Pluses (+) show ±1se prediction limits of individual estimates. Zero flows were included in the sampling. The filled triangle (▲) shows the secondary site mean annual low flow derived from flow data.

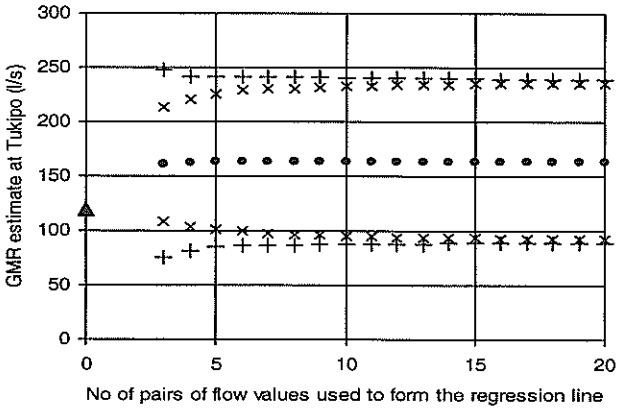

Figure 7 – Geometric mean regression (GMR) estimates of mean annual low flow at Tukipo derived from flows at Otane, indicated by filled circles. Crosses (×) show ±1se prediction limits of individual estimates according to the original equation of Ricker (1973), and pluses (+) show ±1se prediction limits modified as described in the text. Zero flows were included in the sampling. The filled triangle (▲) shows the secondary site mean annual low flow derived from flow data.

negative, and errors are greater than 100%. Thus hydrological similarity is important when choosing a primary site.

When zero flows are excluded from the sampling procedure (cases Anz and Bnz), most biases are reduced and errors increased. An exception is the Bnz case, where the ordinary least-squares bias becomes larger and changes sign, and the ordinary least-squares error reduces. A significant difference between case A and B is that in case A the primary site has all the zero flows, whereas in B the secondary site has most of the zero flows.

Application to 21 site pairs

Table 1 describes and numbers the pairs selected by the cluster analysis, as being similar in the catchment physical parameters listed above; it also gives the values of mean annual low flow derived from time series analysis (the average of the lowest 7-day flow in each calendar year of record). Tables 2 and 3 show the bias and standard errors for samples of five and ten gaugings, for both ordinary least-squares and geometric mean regression methods. The Otane-Tukipo and Otane-Omakere pairs are included at the bottom of each table as cases A and B respectively.

Column headings in Tables 2 and 3 are interpreted as follows:

- "GMR est." is the average estimated mean annual low flow at the secondary site using geometric mean regression on many flow pairs sampled from the time series data.
- "GMR bias" is the difference between the time series value of the secondary site mean annual low flow and the "GMR est." value, divided by the time series value.
- "|GMR error|" is the absolute value of average standard error of the mean annual low flow at the secondary site calculated by equation 2, expressed as a percentage of the time series value.

- "|Revd GMR error|" is the geometric mean regression error adjusted by the scale factor described above.
- Ordinary least-squares (OLS) columns are as for GMR columns except there is no revised standard error.

Some pairs, notably cases 14, 16 and 20 have larger biases and errors than the other pairs. The presence of these few large values suggests median, rather than mean, bias and error should be adopted. For the 21 pairs, with samples of 5 or 10 gaugings, the median biases by geometric mean regression are –4% and –3% respectively, while by ordinary least-squares regression the biases are 5% and 6% respectively. Adopting the modified geometric mean regression error estimates, with samples of 5 or 10 gaugings, the median standard errors of estimate of the predicted mean annual low flow are 25% and 25% respectively, and for ordinary least-squares the equivalent values are 21% and 20% respectively. These values are summarised in Tables 2 and 3. There is little difference in the

Table 2 – Results of regression analysis with 5 flow pairs per sample.

Case	GMR est. (l/s)	GMR bias (%)	|GMR error| (%+/–)	|Revd GM error| (%+/–)	OLS est. (l/s)	OLS bias (%)	|OLS error| (%+/–)
1	0.803	13	17	22	0.827	16	21
2	0.954	–6	7	9	0.968	–4	9
3	289	–6	22	28	331	8	18
4	12.6	2	22	27	13.7	11	24
5	130	–8	25	32	150	6	25
6	553	–3	17	22	596	4	20
7	1447	–3	9	11	1509	1	10
8	1682	–16	26	32	1952	–3	25
9	331	–4	20	25	365	6	21
10	380	9	31	39	449	28	30
11	29	–14	29	36	33	–2	28
12	26.6	–4	17	22	29	5	19
13	10.8	–12	16	20	12	–2	16
14	2.57	10	95	120	3.5	50	81
15	2290	8	18	23	2368	11	21
16	126	14	66	84	168	51	58
17	645	9	20	22	692	17	21
18	795	–2	18	22	841	3	20
19	10.2	–9	16	20	10.7	–4	15
20	1.23	–68	141	178	2.26	–42	86
21	806	–25	31	39	904	–16	38
Mean of 21		–6	32	40		7	29
Median of 21		–4	20	25		5	21
Az	163	37	38	48	176	48	43
Bz	0.969	–25	165	208	1.121	–14	175
Anz	142	19	48	61	167	40	49
Bnz	1.058	–18	211	267	1.775	37	148

Table 3 – Results of regression analysis with 10 flow pairs per sample.

Case	GMR est.	GMR bias	|GMR error|	|Revd GM error|	OLS est.	OLS bias	|OLS error|
	(l/s)	(%)	(%+/–)	(%+/–)	(l/s)	(%)	(%+/–)
1	0.805	13	19	21	0.822	16	20
2	0.955	–5	8	9	0.937	–7	9
3	298	–3	24	27	346	13	18
4	12.6	2	25	27	13.9	13	23
5	132	–6	29	32	156	11	22
6	560	–2	18	21	602	5	18
7	1450	–3	10	11	1460	–2	9
8	1710	–15	27	30	2000	0	21
9	333	–3	22	25	372	8	19
10	387	11	33	36	461	32	26
11	29	–14	31	35	33	–2	27
12	27.3	–1	17	19	29.5	6	16
13	11	–11	17	19	12.3	0	13
14	2.57	10	101	112	3.71	59	70
15	2310	8	19	21	2390	12	20
16	135	22	73	82	190	71	49
17	648	9	20	23	697	18	19
18	795	–2	20	22	860	6	19
19	10.5	–6	17	19	10.7	–4	17
20	1.16	–70	160	178	2.32	–40	72
21	827	–23	31	35	930	–14	34
Mean of 21		–4	34	38		9	26
Median of 21		–3	22	25		6	20
Az	164	38	42	46	178	50	42
Bz	0.938	–28	195	217	1.149	–11	173
Anz	142	19	55	61	174	46	45
Bnz	1.010	–22	252	280	2.054	58	121

error term whether five, ten or twenty pairs are sampled in the Monte Carlo procedure.

When flow pairs with zero flow are excluded from the sampling procedure, results alter. When zeroes are excluded from the Otane-Tukipo pairing (rows Az, Anz in Tables 2 and 3) the mean annual low flow estimate is lowered, and bias is reduced, both for five and ten gauging samples. The standard error increases for both ordinary least-squares and geometric mean regression. For the Otane-Omakere pairing (rows Bz and Bnz) the mean annual low flow estimate is raised, bias is reduced and errors increase for geometric mean regression, but for ordinary least-squares there is an increased absolute bias and a change of sign. The different shifts in mean annual low flow estimate are because for Otane-Tukipo it is the primary site that has zero flow. For Otane-Omakere, zero flows are more common at the secondary site (see Figs. 2 and 3).

Discussion

Previous nation-wide methods for estimating low flows (Pearson, 1995; Hutchinson, 1990) give results with a large uncertainty (see Fig. 1). This is partly a consequence of the wide range of hydrological responses within New Zealand, coupled with a limited number of suitable flow recorder sites (495 in Pearson, 1995), and limited numeric assessments of catchment properties. For example, hydrogeology is only available as an ordinal classification based on rock type (Hutchinson, 1990), but with no consistent or quantified change in properties between adjacent classes.

A database of "miscellaneous" gaugings collected over many years, but only grouped together recently, has allowed the development of new models for low flow estimation. Of over 4000 sites with gaugings but no continuous flow record, 34% have 5 or more gaugings, 19% have 10 or more gaugings, and only 8% have more than 20 gaugings. Low flow statistics from these secondary sites will be useful in new models for estimating low flow. However, the statistics estimated for secondary sites are less reliable than those estimated for primary sites. The need for estimates of uncertainty for these secondary site estimates motivates the current work.

The tests on Otane-Tukipo and Otane-Omakere illustrate the method. The flow duration curves (Fig. 5) show that there is greater similarity between the two catchments from the same Toebes and Palmer (1969) hydrological region (Otane and Tukipo) than between either of these and Omakere. There is also less scatter in the daily flow pairings of Otane and Tukipo (Fig. 3) than between Otane and Omakere (Fig. 4). However, as was noted after the cluster analysis, these sites would not be the first choice for each other. Indeed, by the criteria applied, Omakere and Tukipo are more similar than other pairings among these three catchments, in spite of their differences in flow duration. This indicates that pair selection based only on physical characteristics must be approached with caution.

The choice of including or excluding zero flows when these occur at primary or secondary sites is also not clear from the evidence. For both geometric mean regression examples, and for ordinary least-squares for Otane-Tukipo, bias is reduced when zeroes are excluded, but standard errors increase. For ordinary least-squares for Otane-Omakere, bias is increased, and standard errors decreased. The reduced bias effect is partly explained by the fact that, for these flow pairs, both geometric mean regression and ordinary least-squares biases are in the same direction, which is systematically altered by the exclusion of zero flow at one site. Removal of zero flows affects the standard error term by increasing the mean of the sample. For ordinary least-squares in case Anz, there is a shift in the mean annual low flow estimate as the number of flow pairs increases from 3 to 10. For geometric mean regression there is not such a pronounced change, once again indicating that geometric mean regression is a more stable method. We recommend that zero flows be included in analyses because they are a legitimate low flow event.

When cluster analysis was used to choose the other 21 cases, more consistent results ensued. Figure 8 (case 6 in Tables 1-3) illustrates the generally less variable nature of the relationships between these site pairs. The connected recessions are more generally parallel, and the bias and errors are smaller than for either of the pairings discussed above. The general principle of geometric mean regression versus ordinary least-squares is well illustrated in Figure 8. The overall geometric mean regression line has a greater slope than the ordinary least-squares line, and since the mean annual low flow is smaller than the mean of the data (where the regression lines

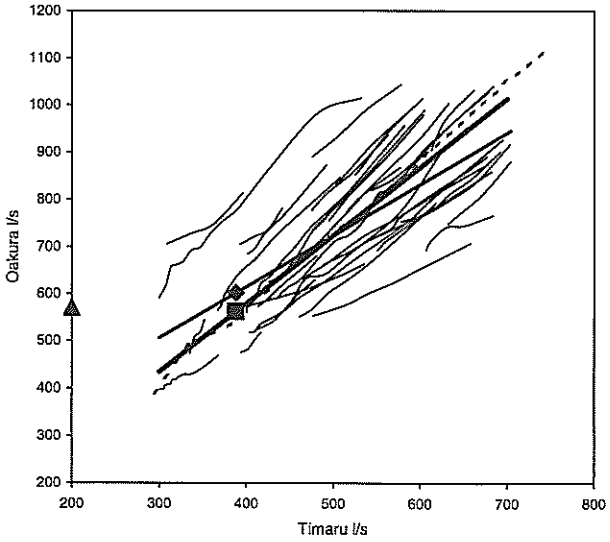

Figure 8 – Flow pairs between Timaru and Oakura, with each recession shown as a fine connected line. The flattest regression line is the ordinary least-squares regression, and the filled diamond (◆) is the ordinary least-squares estimate of mean annual low flow at Oakura. The steeper regression line is the geometric mean regression, and the filled square (■) is the geometric mean regression estimate of mean annual low flow at Oakura. The dashed line from top right is based on matched flow distribution percentiles. The filled triangle (▲) shows the Oakura mean annual low flow derived from flow data.

cross), the bias is reduced accordingly. The other line on the Figure 8 plot shows a different correlation method that has been suggested as a record-extension and gap-filling technique (Hughes and Smakhtin, 1996; Ibbitt and Henderson, 1998). The percentiles of the flow distribution are matched together from two different sites as a non-linear flow-flow transformation. The similarity in slope and location between the geometric mean regression and this method is evident, and lends further support to the use of geometric mean regression.

Figures 9 and 10 show the effect of sample size with ordinary least-squares and geometric mean regression when estimating mean annual low flow at Oakura from flows at Timaru. As expected from the scatter plot comparison (Fig. 8 versus Figs. 2 and 3), there is a smaller standard error of prediction for the mean annual low flow than in the earlier examples. If all cases are examined, there is a general tendency for bias to increase with sample size (i.e. the mean annual low flow estimate gets larger with larger N). Prediction errors do get smaller with larger N, but this effect is very slow, especially when N is larger than 10. This does not mean that ten gauging pairs allow adequate definition of the relationship between two sites at low flow. The goodness of a particular set of simultaneous gauging pairs depends on the extent to which they define the range of the relationship between the two sites, and this depends on the natural variability between two sites.

Testing on many site pairs allowed an appreciation of the natural variability in recession behaviour, even when sites are similar and in proximity. This is the major influence on the overall uncertainty of the mean annual low flow estimates. Scatter such as this, between adjacent or even nested catchments, might be explained by variations in relative catchment wetness as a result of variations in rain distribution from storm to storm. It indicates that a small sample of gauging pairs that show a good correlation is likely to be fortuitous. High values of r^2 can also be achieved in practice by including a few large flow pairs, which have undue influence on this parameter. However, this has the effect of moving the sample mean to larger values, and means that the error on the estimate of the mean annual low flow is larger still. We have avoided this tendency by our sample selection routine and suggest that gaugings at

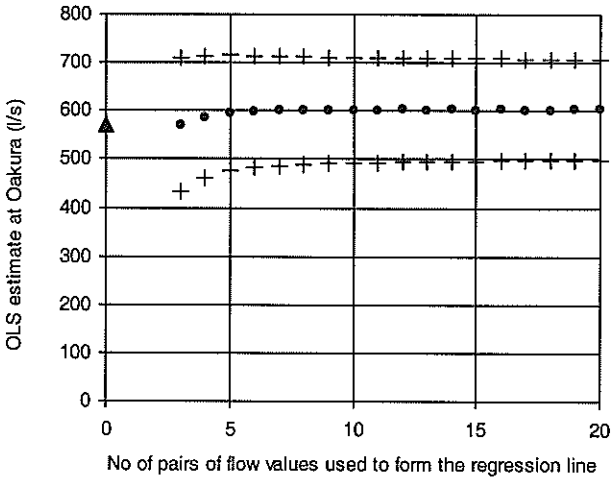

Figure 9 – Ordinary least-squares (OLS) estimates of mean annual low flow at Oakura derived from flows at Timaru (case 6), indicated by filled circles. Pluses (+) show ±1se prediction limits of individual estimates. Zero flows were included in the sampling. The filled triangle (▲) shows the secondary site mean annual low flow derived from flow data.

flows larger than the median flow (or a smaller flow if sufficient data are available), be omitted from analysis of gauging pairs when estimating mean annual low flow.

Consideration of the prediction variance terms in equations 1 and 2 shows that the two factors within control of a field sampling programme are the number of gauging pairs, and the difference between these gaugings and the mean annual low flow at either site. Thus sampling to estimate mean annual low flow at secondary sites needs to concentrate on flow values around the primary site mean annual low flow, on the assumption that the secondary site will be close to that state also. With sufficient sampling an adequate measure of the overall spread of behaviour will be obtained. More than ten gauging pairs over a number of years and dry periods are required to provide an adequate measure of the relationship between two sites.

Zero flows should be included in these analyses since they are often a legitimate occurrence at one site or the other. The effect of removing them varies, depending on whether it is the primary or secondary site that has the zero flows. However, as for frequency analysis, there may be scope for alternative forms of analysis when zero flows are included—i.e. one part of the analysis deals with the positive values and another part with the probability of zero flow.

One aspect of the sampling procedure that caused concern was

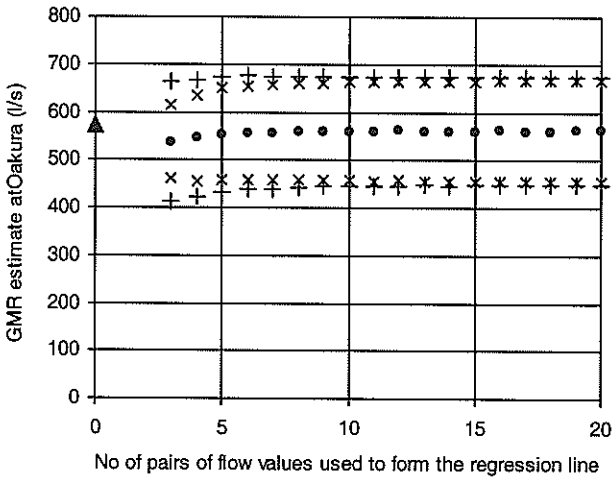

Figure 10 – Geometric mean regression (GMR) estimates of mean annual low flow at Oakura derived from flows at Timaru (case 6), indicated by filled circles. Crosses (×) show ±1se prediction limits of individual estimates according to the original equation of Ricker (1973), and pluses (+) show ±1se prediction limits modified as described in the text. Zero flows were included in the sampling. The filled triangle (▲) shows the secondary site mean annual low flow derived from flow data.

the smaller than expected number of recessions selected for sampling. This was caused by the strict requirement for a recession to be defined by a sequence of monotonically decreasing flows. In some long recessions flows were excluded from the analysis because small freshes caused small rises. If the subsequent recession were long enough, selection from it would recommence. However, if a succession of small freshes occurred, then there are chances of missing some significant recessions. For the sites selected here the records were long enough for this problem to have minimal effect, but it could be important where records are short. To overcome this problem, one option could be to use moving means or accept rises, provided they are less than some threshold. These options have not been considered for the present work as they introduce one more "degree" of freedom into the analysis, and one that may be specific to a given site.

We encountered problems while applying the variance equations (equations 1 and 2). Firstly, Ricker (1973, para. 12, p.414) claims that "Variances calculated from [equation 2] are greater than those from [equation 1] because the GM line is a poorer fit than *b* [ordinary least-squares regression slope] is to the observed points *in the vertical direction* (his emphasis)"; this claim needs examination. In individual cases and in some of the average results presented in Tables 2 and 3, ordinary least-squares standard errors are larger than the geometric mean regression standard error. Secondly, equation 2 gives smaller variances on average at smaller *N*. Our scaling parameter attempts to address both these problems by assuming that variances are equal at the sample mean, and thus calculating an adjustment factor. This does address the small *N* problem, but does not solve the overall problem, especially for values far from the mean of the sample. A more rigorous statistical analysis may be needed to solve these problems, but is beyond the scope of this paper. In the interim, we adopt a statistician's recommendation, based on comments made by Ricker (1973), that the ordinary least-squares variance should be used in statements of significance, since there is a sampling statistic (Student's *t*) available (G. B. McBride pers. comm.).

Conclusions

From the results given here we conclude:

- The standard error of estimate of mean annual low flow at a secondary site derived by geometric mean regression from a primary site with a flow recorder is at least ±20% with five or more gauging pairs. Using ordinary least-squares prediction limits, and assuming normally distributed errors, implies 95% confidence that answers are within ±40%.
- To reduce the uncertainty of mean annual low flow estimates at the secondary site, more than ten flow gaugings should be collected close to the mean annual low flow at the primary site, preferably in a number of different dry periods and years.
- Gaugings at flows greater than the median (or ideally the lower quartile) at the primary site should not be used in regressions for estimating mean annual low flow.
- If zero flows occur at either the primary or secondary sites they should be included in the analysis.
- Sites that are paired should be similar in hydrological character and geographical location, but we do not yet have a robust measure of this similarity.
- A preference for geometric mean regression or ordinary least-squares is not clearly demonstrated by the analysis of bias in the data presented here, but there are sound statistical reasons for preferring geometric mean regression to ordinary least-squares, and support for this from other studies.

Acknowledgements

Regional and District Council partners Martin Doyle (Tasman District Council), Doug Stewart (Environment Waikato), Graeme Horrell (Environment Canterbury) and Mike Harkness (Wellington Regional Council) provided miscellaneous gauging data and useful discussions. FRST Contract CO1X0014 "Floods and Droughts" provided funding for this work. The manuscript has benefited from reviews by Ross Woods, Graham McBride and Paul Mosley.

References

Bowden, M.J. 1974: The Water Resources of the Waiau Catchment. North Canterbury Catchment Board and Regional Water Board, Christchurch, New Zealand. 65 p.

Caruso, B.S. 2000: Evaluation of low-flow-frequency analysis methods. *Journal of Hydrology (NZ) 39(1)*: 19-47.

de Joux, R.T. 1980: The Water Resources of the Orari River. Publication No. 24. South Canterbury Catchment Board and Regional Water Board, Timaru, New Zealand.

Dixon, W.J.; Massey, F.J. 1957: *Introduction to Statistical Analysis.* Tokyo, McGraw Hill Book Company.

Giese, G.L.; Mason, R.R.J. 1993: Low-Flow Characteristics of Streams in North Carolina. USGS Water-Supply Paper 2403, United States Geological Survey, Denver. 29 p.

Grant, P.J. 1968: Variations of Rainfall Frequency in Relation to Drought on the East Coast. *Journal of Hydrology (NZ) 7(2)*: 124-135.

Grant, P.J. 1971: Low Flow Characteristics on Three Rock Types of the East Coast, and the Translation of Some Representative Basin Data. *Journal of Hydrology (NZ) 10(1)*: 22-35.

Harrison, W. 1988: The Influence of the 1982-83 Drought on River Flows in Hawke's Bay. *Journal of Hydrology (NZ) 27(2)*: 1-25.

Hayes, D.C. 1991: Low-flow Characteristics of Streams in Virginia. USGS Water-Supply Paper 2374. United States Geological Survey, Denver, 69 p.

Helsel, D.R.; Hirsch, R.M. 1992: *Statistical Methods in Water Resources*, Elsevier.

Horrell, G.A. 2001: Ashburton River Low Flow Regime. Report No. U01/26. Canterbury Regional Council, Christchurch, New Zealand

Hughes, D.A.; Smakhtin, V. 1996: Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. *Hydrological Sciences Journal 41(6)*: 851-871.

Hutchinson, P.D. 1990: Regression Estimation of Low Flow in New Zealand. Publication of the Hydrology Centre No. 22. DSIR Marine and Freshwater, Christchurch, New Zealand. 51 p.

Ibbitt, R.P.; Henderson, R.D. 1998: Filling in Missing Data in Flow Records. International Symposium on Hydrology, Water Resources and Environmental Development and Management in Southeast Asia, Taegu, Republic of Korea.

ISO 1978: Measurement of fluid flow - Estimation of uncertainty in the flow-rate measurement. Standard ISO 5168, International Standards Organisation. 26 p.

McKerchar, A.I.; Henderson, R.D. 1999: Estimating Low Flows Using Map Information on a CD-ROM, NIWA Ltd, Christchurch, New Zealand.

Parrett, C.; Cartier, K.D. 1990: Methods for Estimating Monthly Streamflow Characteristics at Ungaged Sites in Western Montana. USGS Water-Supply Paper 2365, United States Geological Survey, Denver, 30 p.

Pearson, C.P. 1995: Regional Frequency Analysis of Low Flows in New Zealand Rivers. *Journal of Hydrology (NZ) 33(2)*: 94-122.

Potter, K.W. 2001: A Simple Method for Estimating Baseflow at Ungaged Locations. *Journal of the American Water Resources Association 37(1)*: 177-183.

Rae, S.N. (ed.) 1988: Water and Soil Resources of the Wairau. Blenheim, New Zealand, Marlborough Catchment and Regional Water Board.

Reckhow, K.H.; Chapra, S.C. 1983: *Engineering Approaches for Lake Management*, vol 1, Butterworths.

Ricker, W.E. 1973: Linear Regressions in Fishery Research. *Journal of the Fisheries Research Board of Canada 30(3)*: 409-434.

Riggs, D.S.; Guarnieri, J.A. *et al.* 1978: Fitting Straight Lines When Both Variables are Subject to Error. *Life Sciences 22*: 1305-1360.

Roke, D.L. 1979: The Water Resources of the Kerikeri Inlet Catchments. Water Resources Report 3, Northland Catchment Commission, Whangarei, New Zealand.

Smakhtin, V.U. 2001: Low Flow Hydrology: a Review. *Journal of Hydrology 240*: 147-186.

Stedinger, J.R.; Thomas, W.O.J. 1985: Low-Flow Frequency Estimation Using Base-Flow Measurements. Open-File Report 85-95, United States Geological Survey, Denver, 22 p.

Teissier, G. 1948: La Relation d'Allometrie sa Signification Statistique et Biologique. *Biometrics 4*: 14-48.

Thomas, D.M.; Benson, M.A. 1970: Generalization of Streamflow Characteristics from Drainage-Basin Characteristics. USGS Water-Supply Paper 1975, United States Geological Survey, Denver.

Thomas, W.O.J.; Stedinger, J.R. 1991: Estimating Low-Flow Characteristics at Gauging Stations and Through the Use of Base-Flow Measurements. The United States - People's Republic of China Bilateral Symposium on Droughts and Arid-Region Hydrology, Tucson, Arizona, United States Geological Survey.

Toebes, C.; Morrissey, W.B. 1970: Representative Basins of New Zealand 1970. Miscellaneous Hydrological Publication No. 7. NWASCO, Wellington, New Zealand, 291 p.

Toebes, C.; Palmer, B.R. 1969: Hydrological Regions of New Zealand. Miscellaneous Hydrological Publication No. 4, NWASCO, Wellington, New Zealand. 45 p + 2 maps.

Walter, K. M. 2000: Index to Hydrological Recording Sites in New Zealand. NIWA Technical Report 73, NIWA, Christchurch, New Zealand. 216 p.

Waugh, J.R. 1970a: The Relationship Between Summer Low Flows and Geology in Northland, New Zealand. Miscellaneous Hydrological Publication No. 6, NWASCO, Wellington, New Zealand. 21 p.

Waugh, J.R. 1970b: Base Flow Recessions as an Index of Representativeness in the Hydrological Regions of Northland, New Zealand. In: Symposium on the results of research on representative and experimental basins, IAHS Publication No. 96, pp 602-613.

Whitehouse, I.E.; McSaveney, M.J. *et al.* 1983: Spatial Variability of Low Flows Across a Portion of the Central Southern Alps, New Zealand. *Journal of Hydrology (NZ) 22(2)*: 123-137.

Wilson, J.T. 2000: Evaluation of a Method of Estimating Low-Flow Frequencies from Base-Flow Measurements at Indiana Streams. Water-Resources Investigations Report 00-4063, United States Geological Survey, Denver, 53 p.

Appendix
Correction to geometric mean regression error term

Calculations using Ricker's (1973) equation for variance of an estimate of Y gave smaller variance for small N than for large N. This does not seem realistic, so an investigation of the geometric mean regression and ordinary least-squares formulae was made, assuming that the variance terms should be similar at the mean value $(X = \overline{X})$.

For ordinary least-squares, the variance of an estimate of Y is:

$$Variance = \left(s_{y.x} \sqrt{1 + \frac{1}{N} + \frac{(X - \overline{X})^2}{(N-1)s_x^2}} \right)^2$$

from equation 1.

At $(X = \overline{X})$ the last term is 0, and since

$$s_{y.x}^2 = \left(\frac{\sum y^2 - \left(\sum xy\right)^2 / \sum x^2}{N-2} \right), \text{ then}$$

$$Variance = \left(\frac{\sum y^2 - \left(\sum xy\right)^2 / \sum x^2}{N-2} \right) \frac{N+1}{N}$$

and thence:

$$Variance = \left(\frac{N-1}{N-2} \right) \left(\frac{N+1}{N} \right) \left(\frac{\sum y^2}{N-1} - \frac{\left(\sum xy\right)^2}{(N-1)\sum x^2} \right) \quad (A.1)$$

For geometric mean regression, the variance of an estimate of Y, is

$$Variance = \left(\sqrt{\frac{\sum y^2 (1 - r^2)}{N-1} + \hat{\lambda}^2 (1-r)^2 (X - \overline{X})^2} \right)^2$$

from equation 2 with $y = Y - \overline{Y}$.

At $(X = \overline{X})$ the second term is 0, and since

$$r^2 = \frac{\left(\sum xy\right)^2}{\sum x^2 \sum y^2} \text{ where } x = X - \overline{X} \text{ then}$$

$$Variance = \frac{\sum y^2}{N-1} \left(1 - \frac{\left(\sum xy\right)^2}{\sum x^2 \sum y^2} \right) \text{which expands}$$

to:

$$Variance = \frac{\sum y^2}{N-1} - \frac{\left(\sum xy\right)^2}{(N-1)\sum x^2}. \quad (A.2)$$

Equations A.1 and A.2 differ by the factor $\frac{(N-1)(N+1)}{(N-2)N}$. Table 4 shows the differences obtained at small N when this factor is applied.

Table 4 – Standard error adjustment factor for geometric mean regression

N	$\frac{(N-1)(N+1)}{(N-2)N}$	Factor
3	(2/1)*(4/3)	2.67
4	(3/2)*(5/4)	1.88
5	(4/3)*6/5)	1.60
6	(5/4)*(7/6)	1.46
8	(7/6)*(9/8)	1.31
10	(9/8)*(11/10)	1.24
20	(19/18)*(21/20)	1.11
50	(49/48)*(51/50)	1.04
100	(99/98)*(101/100)	1.02
1000	(999/998)*(1001/1000)	1.00

Manuscript received 25 January 2003; accepted for publication 13 March 2003