NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Micevski, Tom; Hackelbusch, Andre; Haddad, Khaled; Kuczera, George; Rahman, Aatur. "Regionalisation of the parameters of the log-Pearson 3 distribution: a case study for New South Wales, Australia". Originally published in Hydrological Processes Vol. 29, Issue 2, p. 250-260 (2015)

Available from: <http://dx.doi.org/10.1002/hyp.10147>

This is the accepted version of the following article: Micevski, Tom; Hackelbusch, Andre; Haddad, Khaled; Kuczera, George; Rahman, Aatur. "Regionalisation of the parameters of the log-Pearson 3 distribution: a case study for New South Wales, Australia" Hydrological Processes Vol. 29, Issue 2, p. 250-260 (2015), which has been published in final form at <http://dx.doi.org/10.1002/hyp.10147>

Accessed from: <http://hdl.handle.net/1959.13/1058246>

Regionalisation of the Parameters of the Log-Pearson 3 Distribution: A case study for New South Wales, Australia

Tom Micevski¹, André Hackelbusch^{1,2}, Khaled Haddad³,
George Kuczera¹ and Ataur Rahman³

¹Discipline of Civil, Surveying and Environmental Engineering
University of Newcastle, Callaghan NSW 2308, AUSTRALIA

²Faculty of Civil- and Environmental Engineering, Ruhr-University Bochum
44870, Bochum, GERMANY

³School of Computing, Engineering and Mathematics
University of Western Sydney, Penrith, NSW 2751, AUSTRALIA

E-mail: a.rahman@uws.edu.au

Abstract

The index flood method is widely used in regional flood frequency analysis (RFFA) but explicitly relies on the identification of ‘acceptable homogeneous regions’. This paper presents an alternative RFFA method, which is particularly useful when ‘acceptable homogeneous regions’ cannot be identified. The new RFFA method is based on the region of influence (ROI) approach where a ‘local region’ can be formed to estimate statistics at the site of interest. The new method is applied here to regionalise the parameters of the log-Pearson 3 (LP3) flood probability model using Bayesian generalised least squares (GLS) regression. The ROI approach is used to reduce model error arising from the heterogeneity unaccounted for by the predictor variables in the traditional fixed-region GLS analysis. A case study was undertaken for 55 catchments located in eastern New South Wales, Australia. The selection of predictor variables was guided by minimizing model error. Using an approach similar to stepwise regression, the best model for the LP3 mean was found to use catchment area and 50-year, 12-hour rainfall intensity as explanatory variables, whereas the models for the LP3 standard deviation and skewness only had a constant term for the derived ROIs. Diagnostics based on leave-one-out cross validation show that the regression model assumptions were not inconsistent with the data and, importantly, no genuine outlier sites were

identified. Significantly, the ROI GLS approach produced more accurate and consistent results than a fixed-region GLS model, highlighting the superior ability of the ROI approach to deal with heterogeneity. This method is particularly applicable to regions which show a high degree of regional heterogeneity.

1. INTRODUCTION

Flood quantile estimation is required for the design of water infrastructure, floodplain mapping, flood insurance studies and many other water resources management tasks. For sites with reasonably long stream-flow data, at-site flood frequency analysis is the most commonly adopted method for flood quantile estimation. In many situations, however, there is no or little at-site flood data available to undertake meaningful flood frequency analysis. Under this situation, regional flood frequency analysis (RFFA) is often adopted, which attempts to transfer the flood frequency information contained in a group of gauged catchments to an ungauged one (Nezhad et al., 2010).

RFFA essentially consists of two steps: (1) formation of regions, and (2) development of regional estimation equations (e.g. Chebana and Ouarda, 2008). In RFFA, formation of regions can be based on geographic and administrative boundaries or in catchment characteristics data space. Acreman and Sinclair (1986) classified catchments in Scotland using cluster analysis and tested the homogeneity of catchment groups using a likelihood ratio test. The allocation of an ungauged catchment to regions formed in catchment characteristics data space is often problematic. For example, Nathan and McMahon (1990) used Andrews curves for identifying homogeneous groups for low-flow regionalisation in Australia, where the assessment of similarity for an ungauged catchment, with that of the candidate homogeneous groups, involved a degree of subjectivity.

Acreman and Wiltshire (1987) proposed regions without fixed boundaries. Subsequently, Burn (1990a, 1990b) and Zrinji and Burn (1994) proposed the region of influence (ROI) approach where each site of interest (i.e. catchment where flood quantiles are to be estimated) can form its own region. Tasker et al. (1996) compared five methods of developing regression models for ungauged catchments using data from 204 gauging stations in Arkansas. The formation of regions in these methods was based on proximity in geographical space (e.g. geographical regions) or catchment attributes space. They found the ROI approach was the best among the five methods considered. A key advantage of the ROI approach is that it can overcome the inconsistency in flood quantile estimates at the boundary of two neighbouring administrative regions (e.g. state borders). A recent study by Eng et al. (2005) compared the performance of ROI approaches based on predictor-variable similarity or geographical proximity for estimating the 50-year

peak discharge, using an ordinary least squares approach with 1091 sites in south-eastern USA. They found that using geographical proximity produced the smallest predictive errors over the study region. Similar results demonstrating the superiority of geographical proximity over predictor-variable similarity have been shown by others (e.g. Merz and Blöschl, 2005; Kjeldsen and Jones, 2007). Haddad et al. (2012a, b) applied the ROI approach to Australian regional flood studies using geographical proximity as a measure to form ROI regions.

For developing the regional estimation models, one of the most commonly adopted techniques is the index flood method, which is based on the concept of ‘homogeneous regions’ (e.g. Dalrymple, 1960; Wiltshire, 1986a, b; Chowdhury et al., 1991; Lu and Stedinger, 1992; Hosking and Wallis, 1993; Fill and Stedinger, 1995; Gupta and Dawdy, 2006; Cunderlik and Burn, 2006; Castellarin et al., 2008; Meshgi and Khalili, 2009a, b). Identification and treatment of outlier sites in the index flood method is often problematic (Saf, 2010). Furthermore, in some situations, ‘acceptable homogeneous regions’ (Hosking and Wallis, 1993) cannot be established; for example, Bates et al. (1998) could not establish regions (based on geographical proximity and catchment attributes) in south-east Australia that could be treated as ‘acceptably homogeneous’.

When an ‘acceptably homogeneous region’ cannot be identified, application of RFFA methods other than the index flood method need to be investigated. For instance, the quantile regression technique (Thomas and Benson, 1970) relaxes the requirement of an ‘acceptable homogeneous region’. Quantile regression based on generalized least squares (GLS) has been widely adopted in the US (e.g. Stedinger and Tasker, 1985, 1986; Kroll and Stedinger, 1998; Griffis and Stedinger, 2007; Eng et al., 2007). It has been long recognized that, under quite general conditions, GLS procedures produce unbiased and minimum variance estimates (e.g. Theil, 1971). Furthermore, GLS-based methods can account for differing record lengths, spatial cross-correlation in quantile estimates, and, importantly, can distinguish between sampling and model error. The selection of the final set of predictor variables for use with GLS regression usually involves a stepwise variable selection search based on ordinary least squares regression (e.g. Tasker et al., 1996). There is scope to improve upon predictor variable selection by searching for the set of variables that minimizes the model error variance, rather than the total (sampling plus model) error variance.

As an alternative to quantile regression, the parameters of a probability distribution can be regressed against catchment characteristics thus enabling the estimation of the flood frequency distribution at an

ungauged site. Regionalising probability model parameters offers three significant advantages over regionalising quantiles:

1. It ensures flood quantiles increase smoothly with increasing average recurrence interval (ARI), an outcome that may not always be achieved with quantile regression;
2. It is straightforward to combine any at-site flood information with regional estimates using the approach described by Micevski and Kuczera (2009) to produce more accurate quantile estimates; and
3. It permits quantiles to be estimated for any ARI in the range of interest.

Note that these advantages are independent of the probability distribution selected and quantile uncertainty increases with larger ARIs.

The main objective of this study is to develop a RFFA method that can be applied to region which lacks homogeneity. Here, a ROI approach is proposed to derive a local region for each of the statistics of interest and then prediction equations are derived for the statistics using Bayesian GLS regression.

2. REGIONAL REGRESSION MODEL

Following Tasker (1980) the multiple linear regression equation can be expressed as

$$\hat{y}_i = \beta_0 + \sum_{j=1}^P \beta_j x_{ij} + \delta_i \quad \text{for } i = 1, 2, \dots, N \quad (1)$$

where $\hat{y}_i$ is the estimate of the flood statistic at site i (with true value y_i), β_j is the j th regional regression parameter to be estimated, x_{ij} is the j th catchment characteristic at site i , N is the number of sites in the region, P is the number of catchment characteristics, and δ_i is the total error assumed to be independent and normally distributed with zero mean and a constant variance σ_R^2 . The total error comprises of (i) time-sampling error in the sample estimator of y and (ii) model error.

Typically the parameters of the regression model (1) are estimated using ordinary least squares (OLS) estimation which has some serious limitations as noted by Stedinger et al., (1993) and Haddad and Rahman (2012).

To overcome the shortcomings of the OLS regression, generalized least squares (GLS) methods were developed by Stedinger and Tasker (1985, 1986) and Tasker and Stedinger (1989). Following Reis et al. (2005), the linear regression model generalizes to (in matrix form)

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} \quad (2)$$

with $\boldsymbol{\eta}$ being the total error with zero mean and covariance matrix

$$E(\boldsymbol{\eta}\boldsymbol{\eta}^T) = \text{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\sigma_R^2) = \sigma_R^2 \mathbf{I} + \boldsymbol{\Sigma} \quad (3)$$

where $\boldsymbol{\Lambda}$ is the covariance matrix of the total error, $\boldsymbol{\Sigma}$ is the covariance matrix of the sampling errors $\boldsymbol{\varepsilon}$, whose form depends upon the flood statistic of interest and the underlying flood distribution (e.g. Griffis and Stedinger 2007; Micevski and Kuczera 2009) and σ_R^2 is the regional model error variance.

Consider a new site not used in the derivation of the regional model. Let $\mathbf{x}_o$ be the vector of characteristics at the new site. The expected value of the flood statistic y_o is $\mathbf{x}_o^T \hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the GLS estimate of $\boldsymbol{\beta}$ (using the GLS procedures detailed in Micevski and Kuczera (2009)). The predictive variance of y_o is (Reis et al., 2005, Eq. 13, Sec. 2.3):

$$\text{Var}(y_o) = \sigma_R^2 + \mathbf{x}_o^T (\mathbf{X}^T \boldsymbol{\Lambda}^{-1} \mathbf{X})^{-1} \mathbf{x}_o \quad (4)$$

The second term is the contribution of uncertainty in $\boldsymbol{\beta}$ to y_o .

3. CASE STUDY

A high-quality flood database encompassing 55 sites in eastern NSW, Australia, was used in this study. These catchments were selected because they had the highest quality data (REF?) and were minimally affected by regulation. The lengths of the annual maximum flood series in the catchments vary between 24 and 74 years, with a median of 30 years. The catchments are small to medium-sized with areas in the range of 8 to 1010 km² and a median of 203 km². Further information on data collation methods can be found in Haddad et al. (2010).

In addition, a total of 11 climatic and physical catchment characteristics were considered: A = catchment area [km²]; $I_{h,y}$ = rainfall intensity [mm/h] for y -year ARI and duration of h hours, where four values were used ($I_{1,2}$, $I_{1,50}$, $I_{12,2}$, and $I_{12,50}$); R = mean annual rainfall [mm/yr]; E = mean annual areal evaporation [mm/yr]; S_D = stream density [km/km²]; S_L = stream length [km]; S_{1085} = slope of the central 75% of mainstream [m/km]; and F = fraction of catchment area under forest [-].

3.1. Selection of catchment variables for use in regional regression

For the case study region, three predictor equations for the log-Pearson 3 (LP3) moments, mean μ , standard deviation σ , and skewness γ , are required. To identify the optimal set of predictor variables (or catchment characteristics), a procedure similar to stepwise regression was adopted using all 55 sites in a single fixed region. The search for the best set of predictor variables was started by performing a Bayes-

ian GLS regression with just a single, constant term. The model error variance and its posterior standard deviation, and the regression parameter (β_0) were recorded. Then the predictor variable area (A) was added and the regression repeated, with the model error variance and its posterior standard deviation and regression parameters (β_0, β_1) again recorded. This procedure was performed for different combinations of catchment characteristics, as summarised in Table 1. To avoid multicollinearity, the correlation matrix of the $\log_{10}$ catchment characteristics was computed and used to ensure highly correlated characteristics were excluded from combinations tested in the stepwise procedure.

The selection of the optimal predictor set was primarily guided by the model error variance (σ_R^2). The combinations of predictor variables with model error variances that were statistically indistinguishable from the lowest model error variance were then selected. From this subset, models with one or more regression parameters (other than the constant) that were less than two standard deviations from zero were rejected. If multiple models still remained, the simplest model (with the fewest predictor variables) was selected.

3.2. Test for regional homogeneity

To assess the degree of homogeneity for the study data set, a number of alternative regions were proposed as shown in Table 2 and the heterogeneity test proposed by Hosking and Wallis (1993) was applied. None of the regions were found to be ‘acceptably homogeneous’.

This result shows that catchments in eastern NSW exhibit a high degree of heterogeneity, so the index flood method should not be applied. This motivated the consideration of other RFFA methods for the study region.

3.3. Identification of parent distribution

To identify the most appropriate parent distribution for the data set, a goodness-of-fit test was performed on each site using a set of 7 candidate distributions following the approach of Haddad and Rahman (2011). The log-Pearson 3 [LP3] distribution was the best-fit distribution for the selected sites. This motivated adoption of the LP3 as the regional parent distribution for the study area and to then regionalise its three parameters for the new RFFA method. It should be noted here that LP3 is the recommended distribution in the Australian Rainfall and Runoff (I. E. Aust., 1987)

3.4. Selection of the ROI

The fixed-region GLS regression analysis identified the catchment attributes that best account for heterogeneity by minimizing model error variance. However, it is hypothesized that there remains spatial structure in the model error residuals. The existence of such structure implies that the model error variance within sub regions of the fixed region will be less than the fixed-region model error variance. It is in this context that the ROI approach was applied – to further reduce the heterogeneity unaccounted by the fixed-region GLS model – with an ROI approach similar to that of Tasker et al. (1996) being implemented.

The ROI approach uses the physical distance between sites as the distance metric (i.e. geographic proximity). In the first iteration, the 15 nearest stations to the site of interest are selected and a regional GLS regression is performed and the predictive variance (4) is recorded. The second iteration proceeds with the next five closest stations being added to the ROI and repeating the regression. This procedure terminates when all eligible sites have been included in the ROI. The ROI for the site of interest is then selected as the one which yields the lowest predictive variance.

This approach fundamentally differs from that of Tasker et al. (1996) in that it seeks to minimize the regression model's predictive error variance (4) rather than selecting an *a priori* fixed number of sites that minimize a distance metric in catchment characteristic space. The ROI criterion of Tasker et al. (1996) cannot guarantee minimum predictive variance. Moreover, the selection of sites that are minimally different in catchment characteristic space is likely to result in greater uncertainty in β , since these sites will not have a broad “spread” of catchment characteristics $\mathbf{X}$. It is important to note that the predictive variance (4) has two terms: the model error variance and the predictive variance arising from uncertainty in β . The first term is the posterior expected value of the model error variance taken from Micevski and Kuczera (2009) – this is always non-zero and guards against situations where the most likely value of the model error variance is zero. The second term guards against the ROI favouring a small number of sites to minimize model error variance – as the number of sites is reduced, any reduction in model error variance is likely to be offset by an increase in uncertainty in β .

3.5. Assessment of Model Adequacy

Central to this study is the assessment of the adequacy of the regional regression model. A good measure is the raw residual, namely the difference between the sample and regional estimates of the LP3 param-

ter. However, interpretation of the raw residual can be misleading if its variance is not constant. To counter this, a standardized residual R_i was used, which is the raw residual divided by its standard deviation defined as the square root of the sum of the predictive variance of the LP3 parameter and its sampling variance given by the appropriate diagonal element of the sampling covariance matrix Σ . This yields the definition

$$R_i = \frac{\mathbf{x}_i^T \hat{\boldsymbol{\beta}} - \hat{y}_i}{\sqrt{\text{Var}(y_i) + \Sigma_{i,i}}} = \frac{\mathbf{x}_i^T \hat{\boldsymbol{\beta}} - \hat{y}_i}{\sqrt{\sigma_R^2 + \mathbf{x}_i^T (\mathbf{X}^T \boldsymbol{\Lambda}^{-1} \mathbf{X})^{-1} \mathbf{x}_i + \Sigma_{i,i}}} \quad \text{for } i = 1, \dots, N \quad (5)$$

To assess the adequacy of the estimated flood quantiles, standardized residuals, referred to as Z-scores to differentiate from standardized residuals of LP3 parameters, were used. For site i and a given recurrence interval ARI, the Z-score is

$$Z_{\text{ARI},i} = \frac{\log_{10} Q_{\text{ARI},i} - \log_{10} \hat{Q}_{\text{ARI},i}}{\sqrt{\sigma_{\text{ARI},i}^2 + s_{\text{ARI},i}^2}} \quad (6)$$

where the numerator is the difference between the logarithms of the regional quantile and site quantile estimate and the denominator is the square root of the sum of log flood quantile variances for the regional ($\sigma_{\text{ARI},i}^2$) and site ($s_{\text{ARI},i}^2$) estimate. The regional estimates were obtained using the importance-sampling procedure described in Micevski and Kuczera (2009), while the at-site estimates were obtained using the Bayesian approach described in Kuczera (1999). Note that the use of log quantiles renders the difference normally distributed and, hence, enables the use of standard statistical tests.

Leave-one-out cross validation was used to validate the regional models (Haddad et al., 2013). The method of cross-validation leaves out the site of interest and develops regional equations for the mean μ , standard deviation σ , and skewness γ using the remaining sites. This is repeated for all stations considered in this study. This ensures the validation is always an independent test of model performance. By construction, the ROI approach satisfies the cross-validation requirements (since the site of interest was not included in the ROI formed).

4. RESULTS AND DISCUSSION

4.1. Selection of catchment characteristics

The stepwise procedure for selecting the best set of catchment characteristics resulted in the following equations for the LP3 mean (μ), standard deviation (σ), and skewness (γ):

$$\mu = \beta_0 + \beta_1 \log(A) + \beta_2 \log(I_{12,50}) \quad (7)$$

$$\sigma = \beta_0 \quad (8)$$

$$\gamma = \beta_0 \quad (9)$$

with Table 3 providing full details.

Table 1 summarises the model error variance and its uncertainty, as expressed by its posterior standard deviation, for the regional models of the three LP3 parameters and for each combination of catchment characteristics (using all sites in a single fixed region). To illustrate the stepwise procedure, consider Figure 1 which plots the model error variance for the LP3 mean regional model. Combination 1 used no catchment characteristics (only a constant term), combination 2 used only area, while combinations 3 to 15, 16 to 17, and 18 used two, three, and four characteristics respectively (see Table 1 for details of these combinations). The best-performing combinations are 5, 17, and 18, since these have the smallest model error variances. Combination 5, which used area and 50-year, 12-hour rainfall intensity as the catchment characteristics, was selected as the optimal combination. Even though combinations 17 and 18 had slightly lower model error variances, they were not significantly less than model error variance for combination 5 and had one or more regression parameters less than two standard deviations from zero.

Of interest was the selection of catchment characteristics for the skewness. Table 1 shows that use of a constant (combination 1) produced the lowest model error variance. What is striking is that the posterior standard deviation on the model error variance is much larger than the posterior expected value which is very small. This is a consequence of the fact that the total error in LP3 skewness regression is dominated by sampling error in the skewness.

Of particular significance is the conclusion that no catchment characteristics were identified for the LP3 standard deviation and skewness regression models. This finding is consistent with Gruber and Stedinger (2008) who found that a constant model for a regional skewness was the best model for a large region in the southeastern United States. On first glance, this finding appears to be consistent with the assumptions of the index flood method. However, this is not the case. The index flood method assumes that the stan-

dard deviation and skewness are held constant, while in this GLS procedure these parameters are not constant — there is intrinsic variability through the model error variance. Also of interest is the finding that the regional skewness has a posterior expected value of -0.36 with an expected model error standard deviation of 0.04 — this provides a very strong prior for the skewness in LP3 applications and should substantially reduce uncertainty in flood quantile estimates.

4.2. Region of Influence GLS Regression

Figure 2 summarizes the number of sites selected in the ROI for each site and each LP3 parameter. The ROIs for the mean typically had fewer sites than the ROIs for the standard deviation and skewness. On average, ROIs for the mean had 23 sites, 33 sites for the standard deviation, and 46 sites for the skewness. This suggests that the LP3 mean experiences the greatest heterogeneity of the three LP3 parameters. This heterogeneity is illustrated in Figure 3 which shows how the predictive variance (4) varies with the number of sites within the region of influence, for a typical site. If the ROI is made large then the regression for the LP3 mean has an excessive predictive variance, while a small ROI inflates the predictive variance of the skewness regression. This highlights the inherent weakness of a fixed-region regionalisation, which, if made too large, will have a model (and predictive) error inflated by heterogeneity unaccounted for by the catchment characteristics.

The standardized residuals are presented in Figure 4, which plots the standardised residuals for the three LP3 parameters as a function of catchment area. It shows that there is no apparent dependence between catchment area and the LP3 standardized residuals. Additional inspection of the spatial relationship of the standardised residuals (not shown) reveals that there is a relatively even spread of positive and negative residuals throughout the region. This relative lack of observed structure in the standardised residuals suggests that the regional regression equations have adequately removed heterogeneity not accounted for by the predictor variables — if there was strong clustering of positive or negative residuals, this would suggest that the ROI had failed to remove local heterogeneity.

Table 4 reports the results of statistical tests on the standardised residuals for the ROI GLS regressions of the LP3 mean, standard deviation and skewness. The Kolmogorov-Smirnov and Shapiro-Wilk normality tests, evaluated using the R software package (R Development Core Team, 2008), produce no outcomes significant at the 10% level (minimum p value is 0.124). This was also confirmed through inspection of QQ plots (not shown) for the standardised residuals of all three LP3 parameters, which all displayed no

genuine outliers. This suggests the regional equations can be used with considerable confidence with the knowledge that heterogeneity has been adequately accounted for.

The regional equations for the LP3 parameters represent an intermediate goal. The ultimate objective is to infer quantiles at an ungauged site. Tables 5 and 6 report statistical tests for Z scores computed by (6) — for 2-, 10- and 100-year ARI quantiles — for both the fixed-region and ROI GLS regressions, respectively.

These Tables reveal that no statistical test produced outcomes significant at the 10% level, suggesting that assumption of normality for the Z scores has been satisfied. Indeed, QQ plots of Z scores (not shown) indicate that the normality assumption is quite well satisfied with generally all points following a straight line. Now, if the Z scores were indeed normally and independently distributed with mean 0 and variance 1, then the slope of the QQ plot, which can be interpreted as the standard deviation of the sample, should approach 1 and the intercept, which is the mean of the sample, should approach 0 as the number of sites increases. The strength of this hypothesis can be tested using the sample mean and standard deviation of the Z scores computed for the 55 sites, as reported in Tables 5 and 6. A χ^2 test shows that there is a 2.5% chance that the sample standard deviation of the Z scores will exceed 1.177 and a 2.5% chance that it will be less than 0.804 (for 55 sites). Referring to Table 5, one concludes that only the 100-year-ARI quantiles for the fixed-region GLS appear inconsistent with the assumptions made in the regional analysis at the 5% significance level. While the 2-year-ARI quantiles for the ROI GLS regression are marginally outside these bounds, it must be remembered that the sites have differing record lengths (see Section 3) and are correlated, so this slight inconsistency is not overwhelming. Likewise, there is a 2.5% chance that the sample mean of the Z scores will exceed 0.264 and a 2.5% chance that it will be less than -0.264. In this case, both the ROI and fixed-region GLS results were found not to be inconsistent with the hypotheses made in the regional model at the 5% significance level.

These results indicate that the fixed-region GLS model overestimates the uncertainty in the 100-year quantiles. This may be because site heterogeneity was not adequately accounted for by the fixed-region regression model resulting in an inflated model error variance. It was noted that in the case of the LP3 mean, the ROI only used 23 sites, on average, to identify the model with minimum error variance. This is an important finding as it strengthens the case for a ROI approach in preference to a method solely based on a fixed region.

In Figure 5, flood quantiles for ARIs of 10 and 100 years are compared for four sites. For each ARI, the 5, 50 and 95th posterior percentiles of the flood quantile are presented for the site data, fixed-region GLS, and ROI GLS analyses. The regional 90% prediction limits (defined by the 5th and 95th posterior percentiles) for three sites (201001, 210076, and 222007) overlap the at-site 90% prediction limits indicating that the regional analyses are consistent with the at-site analysis. Note that site 203030 had the largest absolute Z-scores and standardised residuals and illustrates the worst case in the cross validation. Although the regional model performs poorly for this site, this is attributed to sampling variability rather than outlier characteristics – see Tables 5 and 6 to confirm that there are no genuine outliers. As expected, flood quantiles using site data have the lowest uncertainty (i.e. the width of the 90% prediction limits is the smallest). Estimates of the posterior median for the three analyses are of the same magnitude for all sites, except site 203030. However, if only the fixed-region and ROI GLS are compared, the ROI GLS predicts flood quantiles with lower uncertainty than does fixed-region GLS for all four sites considered here. Indeed, this is confirmed for the study region in Figure 6 which plots a histogram of the ratio of the 90% prediction-limit intervals (in log space) for the ROI GLS and the fixed-region GLS for all 55 sites. The ratio is consistently below one, indicating that ROI GLS has the smaller predictive uncertainty

5. CONCLUSION

A new RFFA method is proposed in this study which derives a local region using the ROI approach for flood statistics of interest to reduce the degree of heterogeneity. The method involves use of Bayesian GLS regression to regionalise the three parameters of the log-Pearson 3 distribution for 55 sites located in eastern New South Wales, Australia, with the aim of providing flood quantile estimates at ungauged catchments. The ROI GLS model was found to exhibit superior performance than the fixed-region GLS model primarily because it better deals with the heterogeneity unaccounted for by the explanatory variables.

The ROI was the set of sites that produced the smallest predictive variance at the site of interest. The ROI analysis produced ROIs which averaged 23 sites for the LP3 mean, 33 sites for the LP3 standard deviation, and 46 sites for the LP3 skewness. This suggests that the greatest regional heterogeneity was for the mean log-Pearson 3 parameter. Statistical tests on both the standardised residuals and Z scores showed no evidence of inconsistency with the model assumptions.

From the detailed comparison of predictive uncertainty for four sites — one of which had the worst Z-scores and standardised residuals — it showed that, as expected, that the use of at-site data produced the

most accurate quantiles. The worst-case site highlights the fact that the 90% prediction limits of the regional quantile estimates may not overlap with the at-site 90% prediction limits. That said, such cases are to be expected. The important finding was that the worst site was not a genuine outlier. The ROI GLS prediction intervals were consistently more compact than those of the fixed-region model.

The RFFA method developed in this study is general and can be applied to any region. It makes best use of the ROI approach and Bayesian GLS regression to develop regional prediction equations that avoid boundary inconsistency and consider inter-station correlation and sampling error in flood data explicitly. The method has enough flexibility to update the regional flood estimation using the at-site flood data, if available in future. ,

ACKNOWLEDGEMENTS

This work was supported by Project 5 (Regional flood methods for Australia) in relation to the update of Australian Rainfall and Runoff. The authors would like to acknowledge the financial contribution of the Department of Climate Change and Engineers Australia as well as the support of James Ball, Erwin Weinmann, Mark Babister and William Weeks.

REFERENCES

- Acreman, M.C., and C.D. Sinclair (1986). Classification of drainage basins according to their physical characteristics; An application for Flood Frequency Analysis in Scotland. *Journal of Hydrology*, 84(3-4), 365-380.
- Acreman, M.C., and S.E. Wiltshire (1987). Identification of regions for regional flood frequency analysis, Abstract, *EOS*, 68(44), 1262, 1987.
- Bates, B.C., A. Rahman, R.G. Mein, and P.E. Weinmann (1998). Climatic and physical factors that influence the homogeneity of regional floods in south-eastern Australia. *Water Resources Research*, 34(12), 3369-3382.
- Burn, D.H. (1990a). An appraisal of the “region of influence” approach to flood frequency analysis. *Hydrological Sciences Journal*, 35(2), 149-165.

Burn, D.H. (1990b). Evaluation of Regional Flood Frequency Analysis with a Region of Influence Approach. *Water Resources Research*, 26(10), 2257-2265.

Castellarin, A., D.H. Burn, and A. Braith (2008). Homogeneity testing: How homogenous do heterogeneous cross-correlated regions seem. *Journal of Hydrology*, 360(1-4), 67-96.

Chebana, F., and T.B.M.J. Ouarda (2008). Depth and homogeneity in regional flood frequency analysis. *Water Resources Research*, 44(11), W11422, doi:10.1029/2007WR006771.

Chowdhury, J.U., J.R. Stedinger, and L.-H. Lu (1991). Goodness of fit tests for regional flood distributions. *Water Resources Research*, 27(7), 1765-1776.

Cunderlik, J. M., and D.H. Burn (2006). Site-focused nonparametric test of regional homogeneity based on flood regime. *Journal of Hydrology*, 318(1-4), 276–291.

Dalrymple, T. (1960). Flood frequency analysis. *U.S Geol. Survey Water Supply Paper*, 1543-A, 11-51.

Eng, K., G.D. Tasker, and P.C.D. Milly (2005). An analysis of Region-of-Influence methods for Flood Frequency Regionalization in the Gulf-Atlantic rolling plains. *Journal of the American Water Resources Association*, 41(1), 135-143.

Eng, K., P.C.D. Milly, and G.D. Tasker (2007). Flood Regionalization: A Hybrid Geographic and Predictor-Variable Region-of-Influence Regression Method. *Journal of Hydrologic Engineering*, 12(6), 585-591.

Fill, H.D., and J.R. Stedinger (1995). Homogeneity tests based upon Gumbel distribution and a critical appraisal of Dalrymple's test. *Journal of Hydrology*, 166(1-2), 81-105.

Griffis, V.W., and J.R. Stedinger (2007). The use of GLS regression in regional hydrologic analyses. *Journal of Hydrology*, 344(1-2), 82-95.

Gruber, A.M., and J.R. Stedinger (2008). Models of LP3 Regional Skew, Data Selection, and Bayesian GLS Regression, paper presented at EWRI World Water & Environmental Resources Congress, American Society of Civil Engineers, Honolulu, HI, May 13-16, 2008, 10 pages.

Gupta, V.K. and Dawdy, D. R. (2006). Physical interpretations of regional variations in the scaling exponents of flood quantiles. *Hydrological Processes* 9(3–4): 347–361.

Haddad, K., Rahman, A., Weinmann, P.E., Kuczera, G. and Ball, J.E. (2010). Streamflow data preparation for regional flood frequency analysis: Lessons from south-east Australia. *Australian Journal of Water Resources*, 14, 1, 17-32.

Haddad, K. and Rahman, A. (2011). Selection of the best fit flood frequency distribution and parameter estimation procedure – A case study for Tasmania in Australia, *Stochastic Environmental Research & Risk Assessment*, 25, 415-428.

Haddad, K. and Rahman, A. (2012a). Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework – Quantile Regression vs. Parameter Regression Technique, *Journal of Hydrology*, 430-431 (2012), 142-161.

Haddad, K., Rahman, A., and Stedinger, J.R. (2012b). Regional Flood Frequency Analysis using Bayesian Generalized Least Squares: A Comparison between Quantile and Parameter Regression Techniques, *Hydrological Processes*, 26, 1008-1021.

Haddad, K., Rahman, A., Zaman, M., Shrestha, S. (2013). Applicability of Monte Carlo Cross Validation Technique for Model Development and Validation in Hydrologic Regression Analysis Using Ordinary and Generalised Least Squares Regression, *Journal of Hydrology*, <http://dx.doi.org/10.1016/j.jhydrol.2012.12.041>

Hosking, J.R.M., and J.R. Wallis (1993). Some statistics useful in regional frequency analysis. *Water Resources Research*, 29(2), 271-281.

Institution of Engineers Australia (I. E. Aust.) (1987). Australian Rainfall and Runoff, Vol. 1, The Institution of Engineers Australia, Canberra.

Kjeldsen, T.R., and D.A. Jones (2007). Estimation of an index flood using data transfer in the UK, *Hydrological Sciences Journal*, 52(1), 86-98.

Kroll, C.N., and J.R. Stedinger (1998). Generalized Least Squares Regression Procedures Revisited, *Water Resources Research*, 34(1), 121-128.

Kuczera, G. (1999). Comprehensive At-Site Flood Frequency Analysis Using Monte Carlo Bayesian Inference, *Water Resources Research*, 35(5), 1551-1557, doi:10.1029/1999WR900012.

Lu, L.-H., and J.R. Stedinger (1992). Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test, *Journal of Hydrology*, 138(1-2), 223-245.

Meshgi, A., and D. Khalili (2009a) Comprehensive evaluation of regional flood frequency analysis by L-and LH-moments. 1. A re-visit to regional homogeneity. *Stoch Environ Res Risk A* 23:119-135. doi:10.1007/s00477-007-0201-7

Meshgi, A., and D. Khalili (2009b) Comprehensive evaluation of regional flood frequency analysis by L-and LH-moments. II. Development of LH-moments parameters for the generalized Pareto and generalised logistic distributions. *Stoch Environ Res Risk A* 23:137-152. doi:10.1007/s00477-007-0201-7

Merz, R., and G. Blöschl (2005). Flood frequency regionalisation—spatial proximity vs. catchment attributes, *Journal of Hydrology*, 302(1-4), 283-306.

Micevski, T., and G. Kuczera (2009). Combining site and regional flood information using a Bayesian Monte Carlo approach, *Water Resources Research*, 45(4), W04405, doi:10.1029/2008WR007173.

Nathan, R.J., and T.A. McMahon (1990). Identification of homogeneous regions for the purpose of regionalisation. *Journal of Hydrology*, 121(4), 217-238.

Nezhad, M.K., Chokmani, K., Ouarda, T.B.M.J., Barbet, M., Bruneau, P. (2010). Regional flood frequency analysis using residual kriging in physiographical space. *Hydrological Processes* 24(:): 2045–2055, DOI: 10.1002/hyp.7631.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0.

Reis Jr, D.S., J.R. Stedinger, and E.S. Martins (2005). Bayesian Generalized Least Squares Regression with Application to Log Pearson Type 3 Regional Skew Estimation. *Water Resources Research*, 41(10), W10419, doi:10.1029/2004WR003445.

Saf, B. (2010). Assessment of the effects of discordant sites on regional flood frequency analysis, *Journal of Hydrology*, 380(3-4), 362-375.

Stedinger, J.R., R.M. Vogel, and E. Foufoula-Georgiou (1993). Frequency Analysis of Extreme Events, in *Handbook of Hydrology*, edited by D.R. Maidment, chapter 18, McGraw-Hill, New York.

Stedinger, J.R., and G.D. Tasker (1985). Regional Hydrologic Analysis. 1. Ordinary, Weighted, and Generalized Least Squares Compared. *Water Resources Research*, 21(9), 1421-1432. (Correction: *Water Resources Research*, 22(5), 844, 1986).

Stedinger, J.R., and G.D. Tasker (1986). Regional Hydrologic Analysis. 2. Model-Error Estimators, Estimation of Sigma and Log-Pearson Type 3 Distributions. *Water Resources Research*, 22(10), 1487-1499.

Tasker, G.D., and J.R. Stedinger (1989). An Operational GLS Model for Hydrologic Regression, *Journal of Hydrology*, 111(1-4), 361-375.

Tasker, G.D. (1980). Hydrologic Regression With Weighted Least Squares, *Water Resources Research*, 16(6), 1107-1113.

Tasker, G.D., S.A. Hodge, and C.S. Barks (1996). Region of Influence regression for estimating the 50-year flood at ungauged sites, *Water Resources Bulletin*, 32(1), 163-170.

Theil, H. (1971), *Principles of Econometric*, Wiley.

Thomas, D.M., and M.A. Benson (1970). Generalization of streamflow characteristics from drainage basin characteristics. *U.S. Geol. Survey Water Supply Paper*, 1975.

Wiltshire, S.E. (1986a). Identification of homogeneous regions for flood frequency analysis. *Journal of Hydrology*, 84(3-4), 287-302.

Wiltshire, S.E. (1986b). Regional flood frequency analysis I: Homogeneity statistics. *Hydrological Sciences Journal*, 31(3), 321-333.

Zrinji, Z., and D.H. Burn (1994). Flood Frequency analysis for ungauged sites using a region of influence approach. *Journal of Hydrology*, 153(1-4), 1-21.

CAPTIONS

Table 1 Summary of the catchment characteristics used in the stepwise regression (single fixed region).

Table 2 Results from Hosking and Wallis (1993) test to assess the degree of heterogeneity^a for various candidate regions in eastern New South Wales (Australia)

Table 3 Summary of the final GLS regression results (single fixed region)

Table 4 Summary of the standardised residual statistics for the ROI GLS regressions.

Table 5 Summary of the Z-score statistics for the fixed-region cross-validation GLS regressions.

Table 6 Summary of the Z-score statistics for the ROI GLS regressions.

Figure 1. Selecting explanatory variables for the mean (μ) GLS regression model. The lighter/darker bars denote the posterior mean/standard deviation of the regional model error variance term σ_R^2 . Refer to Table 1 for a description of the combinations used.

Figure 2 Number of sites which produced the lowest predictive variance for the (a) mean, (b) standard deviation, and (c) skewness ROI GLS regression models.

Figure 3. Typical effect of the size of the region of influence on the predictive variance. The mean (m) and standard deviation (s) use the left-hand axis, while the skewness (g) uses the right-hand axis.

Figure 4 Plot of the relationship between the ROI GLS standardised residual and catchment area for the mean (m), standard deviation (s), and skewness (g) regression models.

Figure 5 Posterior distributions of 10- and 100-year flood quantiles for four sites. The 5, 50 and 95th posterior percentiles of the flood quantiles are presented for the site data (Site), fixed-region cross-validation GLS (CV), and the ROI GLS (ROI).

Figure 6 Histogram of the ratio of 90% prediction-limit intervals for the ROI GLS to the fixed-region GLS. Prediction limits for the (log space) 100-year flood quantile are used.

TABLES

Table 1 Summary of the catchment characteristics used in the stepwise regression (single fixed region).

Combination	Catchment Characteristics ^a	Posterior mean (standard deviation) of model error variance		
		LP3 parameter		
		Mean	Standard deviation	Skewness
1	K	1.1652 (0.2437)	0.1203 (0.0291) ^b	0.0019 (0.0069) ^b
2	K, A	0.8285 (0.1799)	0.1208 (0.0293)	0.0052 (0.0102)
3	$K, A, I_{1,2}$	0.3760 (0.0891)	0.1113 (0.0272)	0.0050 (0.0106)
4	$K, A, I_{1,50}$	0.3943 (0.0927)	0.1168 (0.0288)	0.0051 (0.0107)
5	$K, A, I_{12,50}$	0.3150 (0.0761) ^b	0.1101 (0.0269)	0.0052 (0.0107)
6	K, A, R	0.3918 (0.0884)	0.0870 (0.0228)	0.0048 (0.0107)
7	K, A, F	0.8420 (0.1804)	0.1188 (0.0281)	0.0030 (0.0083)
8	K, A, E	0.5861 (0.1289)	0.1092 (0.0265)	0.0071 (0.0113)
9	K, A, S_{1085}	0.8104 (0.1755)	0.1242 (0.0303)	0.0051 (0.0109)
10	K, A, S_L	0.8410 (0.1807)	0.1131 (0.0280)	0.0047 (0.0106)
11	$K, I_{12,50}, R$	0.7076 (0.1551)	0.0859 (0.0216)	0.0021 (0.0063)
12	K, S_L, R	0.4665 (0.1053)	0.0867 (0.0224)	0.0048 (0.0107)
13	K, S_D, R	0.8652 (0.1844)	0.0863 (0.0224)	0.0045 (0.0105)
14	K, F, R	0.9068 (0.2011)	0.0860 (0.0224)	0.0025 (0.0076)
15	K, S_{1085}, F	0.9958 (0.2120)	0.1216 (0.0301)	0.0050 (0.0106)
16	$K, A, I_{12,50}, S_D$	0.3245 (0.0801)	0.1123 (0.0288)	0.0033 (0.0087)
17	$K, A, I_{12,50}, R$	0.2922 (0.0687)	0.0885 (0.0231)	0.0037 (0.0087)
18	$K, A, I_{12,50}, R, E$	0.2893 (0.0692)	0.0891 (0.0227)	0.0034 (0.0095)

^a K is a constant term. Refer to text in Section 3 for a full description of the characteristics.

^bDenotes the selected combination.

Table 2 Results from Hosking and Wallis (1993) test to assess the degree of heterogeneity^a for various candidate regions in eastern New South Wales (Australia)

Candidate regions	No. of sites	H(1)	H(2)	H(3)	No. of discordant sites	Selected distribution
Total data set	55	10.91	8.35	5.89	2	LP3
	53	10.58	8.00	5.85	0	LP3
North-eastern NSW	21	7.31	5.05	3.49	0	LP3
South-eastern NSW	34	9.66	5.92	3.44	1	LP3
	33	7.16	5.55	3.67	2	LP3
	31	6.33	5.24	7.83	1	LP3
	30	7.28	5.69	4.06	0	LP3
Area range 0-100 km ²	12	8.51	6.90	3.29	0	LP3
Area range 101 - 300 km ²	24	7.61	5.59	2.98	1	LP3
	23	5.91	4.70	2.53	0	LP3
Area range 301 km ² and greater	19	6.60	5.09	4.37	0	LP3

^aRegions are “acceptably homogeneous” for heterogeneity statistics $H < 1$.

Table 3 Summary of the final GLS regression results (single fixed region)

GLS Regression Model	Regression Parameter	Posterior Moment	
		Mean	Standard Deviation
Mean μ	σ_R^2	0.3150 = 0.5612 ²	0.0761
	β_0 (constant)	0.3326	0.7898
	β_1 ($\log_{10} A$)	1.3222	0.1797
	β_2 ($\log_{10} I_{12,50}$)	4.7182	0.5595
Standard Deviation σ	σ_R^2	0.1202 = 0.1096 ²	0.0291
	β_0 (constant)	1.2318	0.0716
Skewness γ	σ_R^2	0.0019 = 0.0436 ²	0.0069
	β_0 (constant)	-0.3567	0.0971

Table 4 Summary of the standardised residual statistics for the ROI GLS regressions.

GLS Regression Model	Sample		Test p value		
	Mean m	Standard deviation s	K-S test ^a		Shapiro-Wilk test
			N(m,s ²)	N(0,1 ²)	
Mean μ	-0.102	1.21	0.857	0.857	0.343
Standard Deviation σ	-0.0259	1.13	0.821	0.821	0.75
Skewness γ	-0.0475	0.806	0.714	0.75	0.124

^aThe N(m,s²) and N(0,1²) Kolmogorov-Smirnov tests respectively compare the observed distributions to the corresponding sampled (m and s²) and standard normal distributions.

Table 5 Summary of the Z-score statistics for the fixed-region cross-validation GLS regressions.

ARI	Sample		Test p value		
	Mean m	Standard deviation s	K-S test		Shapiro-Wilk test
			N(m,s ²)	N(0,1 ²)	
2	-0.0165	1.01	0.964	0.964	0.799
10	-0.0851	0.813	0.75	0.786	0.971
100	-0.198	0.725	0.607	0.607	0.648

Table 6 Summary of the Z-score statistics for the ROI GLS regressions.

ARI	Sample		Test p value		
	Mean m	Standard deviation s	K-S test		Shapiro-Wilk test
			N(m,s ²)	N(0,1 ²)	
2	0.0312	1.2	0.893	0.893	0.219
10	-0.0827	1.06	0.893	0.857	0.713
100	-0.223	0.885	0.714	0.714	0.343

FIGURES

Figure 1. Selecting explanatory variables for the mean (μ) GLS regression model. The lighter/darker bars denote the posterior mean/standard deviation of the regional model error variance term σ_R^2 . Refer to Table 1 for a description of the combinations used.

Figure 2. Number of sites which produced the lowest predictive variance for the (a) mean, (b) standard deviation, and (c) skewness ROI GLS regression models.

Figure 3. Typical effect of the size of the region of influence on the predictive variance. The mean (m) and standard deviation (s) use the left-hand axis, while the skewness (g) uses the right-hand axis.

Figure 4. Plot of the relationship between the ROI GLS standardised residual and catchment area for the mean (m), standard deviation (s), and skewness (g) regression models..

Figure 5. Posterior distributions of 10- and 100-year flood quantiles for four sites. The 5, 50 and 95th posterior percentiles of the flood quantiles are presented for the site data (Site), fixed-region cross-validation GLS (CV), and the ROI GLS (ROI).

Figure 6. Histogram of the ratio of 90% prediction-limit intervals for the ROI GLS to the fixed-region GLS. Prediction limits for the (log space) 100-year flood quantile are used.