Regionalisation of rainfall statistics for the IFD Revision Project

Fiona Johnson
Hydrologist, Climate and Water Division, Bureau of Meteorology, Sydney, Australia

Karin Xuereb
Senior Meteorologist, Climate and Water Division, Bureau of Meteorology, Melbourne, Australia

Erwin Jeremiah
Hydrologist, Climate and Water Division, Bureau of Meteorology, Sydney, Australia

Janice Green
IFD Revision Project Manager, Climate and Water Division, Bureau of Meteorology, Canberra, Australia

The Australian Bureau of Meteorology has revised the Intensity-Frequency-Duration (IFD) design rainfall estimates for Australia. The revised IFDs cover Annual Exceedance Probabilities (AEPs) ranging from 50% to 1%. However, as even daily rainfall records are often much shorter than 100 years, there is significant uncertainty in the estimates of less frequent events. Regionalisation is an approach that overcomes this problem by assuming that information can be combined from multiple stations to give more accurate estimates of the parameters of the extreme value probability distributions

A Region of Influence (ROI) approach has been adopted for the IFD revision project. This paper describes the process used to define the ROIs and to determine the optimum region sizes. A number of approaches for the definition of the ROIs were considered including spatial proximity, using ellipses to define a preferred direction for spatial proximity, and using non-geographical site characteristics to define the closest stations. It was found that a simple measure of spatial proximity which included elevation as well as latitude and longitude gave the best results. In general it was found that predictive uncertainty in the rainfall quantiles was minimised when regions were sized to include 500 station-years of data or 8 neighbours. This is consistent with earlier studies that have suggested an appropriate rule of thumb is a requirement for the size of a region to be approximately five times the largest quantile of interest.

1. INTRODUCTION

The Bureau of Meteorology has revised the Intensity-Frequency-Duration (IFD) design rainfall data for Australia. This revision will improve on the current IFD data (Pilgrim, 2001) through the use of current statistical analysis techniques and by the improved station data density and record lengths available since the earlier IFDs were published. The increase in record lengths is important because short records can bias the estimates of rainfall statistics and the resulting rainfall quantile estimates will have large uncertainties associated with them. However even with the increased data available for the current revision, the majority of stations have periods of record that are shorter than the exceedance probabilities of interest for engineering design. One approach that is widely used to reduce the uncertainty and overcome bias in estimating rainfall quantiles is regional frequency analysis. Regional frequency analysis, also known as regionalisation, recognises that for stations with short records, there is considerable uncertainty in estimating the parameters of probability distributions. To overcome this, it is assumed that by combining data from different stations, more accurate estimates of the probability distribution parameters can be made.

This paper presents a brief literature review on regionalisation approaches which was used to determine options to be tested for the IFD revision project. Regionalisation has been used in other studies of both rainfall quantile estimates as well as streamflow estimates. Literature from both fields can provide useful guidance for this project. The statistics used to evaluate the different methods are also presented. The results of the analysis and recommended strategy for regionalisation for the IFD revision project are then presented. This paper only discusses the regionalisation of daily rainfall events (1 day to 7 days) for the IFD revision project. For sub-daily durations the regional frequency analysis includes an additional step of Bayesian Generalised Least Squares Regression which has been reported separately (Haddad and Rahman, 2012; Johnson *et* al., 2012).

2. METHODOLOGY

The assumption behind regionalisation is that stations (or catchments in the case of streamflow) can be grouped into relatively homogeneous regions of "similar stations" such that a particular statistic of interest is equal at all stations (e.g. skewness) or alternatively there is a probability distribution that is common to all stations once scaled by a site specific factor. A number of decisions must be made in defining the homogenous regions (Kjeldsen and Jones, 2009); these include the optimum size of the region, the characteristics and weightings to be used to define similarity between sites and finally the statistics to be used to evaluate the regionalised rainfall quantiles. Approaches to address each of these issues are discussed below to identify options for testing the daily rainfall data for the IFD revision project.

2.1. Regionalisation approaches

Many regionalisation approaches are based around the "index flood" concept such that all stations in a homogenous region have an identical probability distribution after a site specific scaling factor is applied. This is called an "index flood" approach, as it was originally developed for flood frequency analysis but it can be used with any other type of data (Hosking and Wallis, 1997). In this article, it is defined as the "index rainfall" approach, as the regionalisation is applied with rainfall data. Testing of possible probability distributions to fit the Annual Maximum Series (AMS) from Australian rainfall data has shown that the Generalised Extreme Value (GEV) distribution is appropriate for both site (Green *et* al*.,* 2010) and regionalised data at the majority of stations and this distribution is used for the testing reported in this paper.

The homogenous regions for the frequency analysis can be defined in a number of ways. Cluster and partitioning methods divide the set of all stations into a fixed number of homogenous groups (Hosking and Wallis, 1997) where generally every site is assigned to one group. Alternatively, a Region of Influence (ROI) approach (Burn, 1990) can be adopted, such that for each station an individual homogenous region is defined. Each ROI will contain a potentially unique set of sites, with each site possibly contributing to multiple ROIs. The advantage of this approach is that the region sizes can be easily varied according to station density and the available record lengths. Jakob *et* al*.* (2005) recommended that the ROI approach should be used for the IFD pilot study due to its ease of application, consistency across different rainfall event durations and slightly better validation statistics than a clustering (fixed region) approach. ROIs have thus been used for the Australia wide regionalisation. The following discussion and results detail how the ROIs have been defined for the project.

2.2. Optimum region sizes

In the application of the ROI method, it is first necessary to establish how big the ROIs should be. The size of the ROI can be defined in two ways; either using the number of stations included in the region or alternatively by calculating the total number of station-years in the region as the sum of the record lengths of the individual stations included in the ROI. Kjeldsen and Jones (2009) found that there was no difference in using either methods to define the region size. The optimum region size needs balance the competing demands of sites being similar (ideally a small number of stations) against the requirements to estimate infrequent events (ideally a large number of stations) (Kjeldsen and Jones,

2009). If there are too few sites, the variance of the parameter estimates will be large and the advantages of regionalisation not realised. With too many sites there will be too much heterogeneity in the region and the assumptions of homogeneity will be violated. In addition the quantile estimates can be biased by the stations that are too different from the station of interest.

Burn (1997) and Thompson (2011) used initial region sizes of 20 to 25 stations. Based on simulation studies, Hosking and Wallis (1997) found that regions did not need to be larger than around 20 sites unless extreme quantiles were to be estimated. Jakob *et* al. (1999a) recommended that a region should contain a total number of station-years equal to approximately five times the return period of interest. Kjdelsen and Jones (2009) also suggest that the use of regions with 500 station-years is appropriate regardless of the target return period.

The first assessment of the regionalisation for the IFD revision therefore investigates the sensitivity of the results to the size of the region. In this case, the simplest definition for the ROI has been used where a circular ROI is defined based on spatial proximity. The purpose of this assessment is to find the optimum region sizes and allow testing of the decision points in the regionalisation.

2.3. Similarity measures

After defining the number of stations in each ROI, a similarity measure is required to determine which stations should be included in the region for the station of interest. The similarity measure can be based on spatial proximity i.e. those stations that are physically closest to the station of interest are chosen to form the ROI. Alternatively the proximity can be defined using non-geographical measures. In this case, the similarity measure is based on the distance from a site to the site of interest in a multi-dimensional space where the dimensions are measures of site characteristics that are known to affect large rainfall events.

Examples of non-geographic similarity measures include Mean Annual Rainfall (MAR) which was found to give unbiased estimates of the design rainfall (Di Baldassarre *et* al., 2006). Jakob *et* al. (1999a; 1999b) considered over 30 catchment characteristics for their influence on the flood statistics. It was found that the best similarity measure was comprised from catchment area, standard average annual rainfall and the base flow index for the site. Flood seasonality has been used by Cunderlik and Burn (2006) and Burn (1997) to define homogeneous regions.

For the IFD revision project, ROIs using spatial proximity have been tested in a number of ways:

- Distance between sites (in kms) defined using latitude and longitude
- Euclidean distance between sites where distance was defined using latitude, longitude and scaled elevation (Hutchinson, 1995)
- Nearest neighbours defined using distance in kilometres inside an elliptical ROI.

The rationale for the third geographic similarity measure is that stations along the coast may be more meteorologically similar to the target station than sites located closer in distance to the targeted site but further inland. In particular, this is thought to be the case for the eastern seaboard of Australia where the Great Dividing Range runs approximately parallel to the coast. It is therefore considered that an elliptical region of influence may lead to more meteorologically similar sites being selected in the region than a circular ROI.

To implement the third geographic similarity measure, the ellipse parameters are required including the ratio of the major to the minor axis of the ellipse (eccentricity) representing the orientation and size of the ellipse. The ellipse eccentricity has been allowed to smoothly vary depending on the distance of the station from the coast with the maximum eccentricity for stations within 1 degree of the coast and zero eccentricity (i.e. a circle) for stations more than 3 degrees from the coast. The ellipse orientation was calculated in ArcGIS by calculating the angle from each station to the nearest point on the mainland coast. The ellipse orientation is then calculated to be perpendicular to this angle. The angles from the nearest three stations are averaged to ensure that minor variations in the coast orientation did not influence the overall ellipse orientation. Finally the ellipse area is increased incrementally until the region contained the required number of stations or station-years.

In considering non-geographic similarity measures, characteristics believed to possibly exert some influence on the properties of large rainfall events at a site were identified (Green *et* al., 2011). Based on the recommendations of Cunderlik and Burn (2006) the seasonality of the AMS has also been considered as a useful predictor. The site characteristics that have been used for the analyses are:

- Location (latitude and longitude)
- Elevation
- Mean Annual Rainfall (MAR)
- Aspect (Asp)
- Slope
- Distance from the coast (Dist)
- Mean date of AMS (seasonality)
- Variability of AMS occurrence (seasonality)

Once the important predictors are established, then the distance from the site of interest to the other stations can be calculated using the Euclidean distance in the multi-dimensional space defined by the predictors. For the initial analysis, the "coordinate system" for this multi-dimensional space uses equal weightings. An optimised weighting for the best combination of predictors was also considered.

2.4. Evaluation of the regionalisation strategies

The inherent difficulty with testing regionalisation approaches is that the site estimates themselves have uncertainty so there is no "true" rainfall quantile to which the regionalised estimates can be compared. The Pooled Uncertainty Measure (PUM) (Jakob *et* al., 1999b; Kjeldsen and Jones, 2009) measures the average bias of the regionalised estimates compared to the at-site estimates over all sites. The PUM is defined as shown in Equation 1 with the station weighting as proposed by Kjeldsen and Jones (2009):

$$PUM_T = \left(\frac{\sum_{i=1}^{M} h_i \left(\log z_{T,i} - \log z_{T,i}^{p} \right)^2}{\sum_{i=1}^{M} h_i} \right)^{1/2} \quad \text{where } h_i = \frac{n_i}{1 + n_i/16} \tag{1}$$

Where $z_{T,i}$ is the at-site 1 in T Annual Exceedance Probability (AEP) rainfall quantile for the station i, the superscript p indicates the regionalised rainfall quantile, h_i is the weight assigned to station i and n_i is the record length in years. In general stations with longer record lengths are expected to have smaller uncertainties in their site estimates. They are therefore given a higher weighting in the calculation of the PUM evaluation statistic since the weighting factor h_i is defined using the station record length.

In addition to calculating PUM, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) have been used as additional evaluation statistics. Because the three statistics have different transformations for the rainfall quantiles (logarithm for PUM, square for RMSE and untransformed for MAE), they will each differently emphasise errors in the magnitudes of rainfall quantiles. Generally MAE is preferred to RMSE because the squared differences in the RMSE can be dominated by a small number of large errors (Lettenmaier and Wood, 1993).

Clearly, the minimum value for any of the evaluation statistics will occur for the case where the ROIs only contain the sites of interest (i.e z_T and z_T^p are the same). However this does not account for the error in the site estimates. To overcome this problem, all method evaluations assume an "ungauged site", such that the site of interest is not included in the ROI. The estimates of the $z_{T,I}$ and $z^p_{T,I}$ are then independent and the different regionalisation strategies can be evaluated fairly. It is important to note that once the testing of the regionalisation is complete, the final regionalised estimates will of course use the site of interest as part of its own ROI.

3. RESULTS AND DISCUSSION

The following section presents results for each of the regionalisation strategies that are being considered for the IFD revision project.

3.1. Optimum region sizes

Region sizes from 1 to 50 stations and from 50 to 5000 station-years were investigated to establish the optimum ROIs for estimating rainfall quantiles across Australia using a simple circular ROI. Figure 1 illustrates the variation of PUM with region size for the 24 hour duration rainfall event. For all rainfall quantiles, the predictive error decreases quickly as the region size gets larger until a minimum PUM value is reached. The PUM values then increase slowly as the regions get too large with too much heterogeneity. The minimum PUM values occur where there is an optimum size in the trade off between bias and variance of the regionalised estimates. It was found that a region with 500 station-years generally leads to the minimum PUM value. When considering the region defined using the number of stations it is found that a region of 8 stations performs best. Given that the average record length for stations used in the analysis is 66 years, a region of 8 stations will have on average 528 years of data and this is consistent with the region size using the station-year criteria. The findings are generally independent of rainfall event duration and return period.

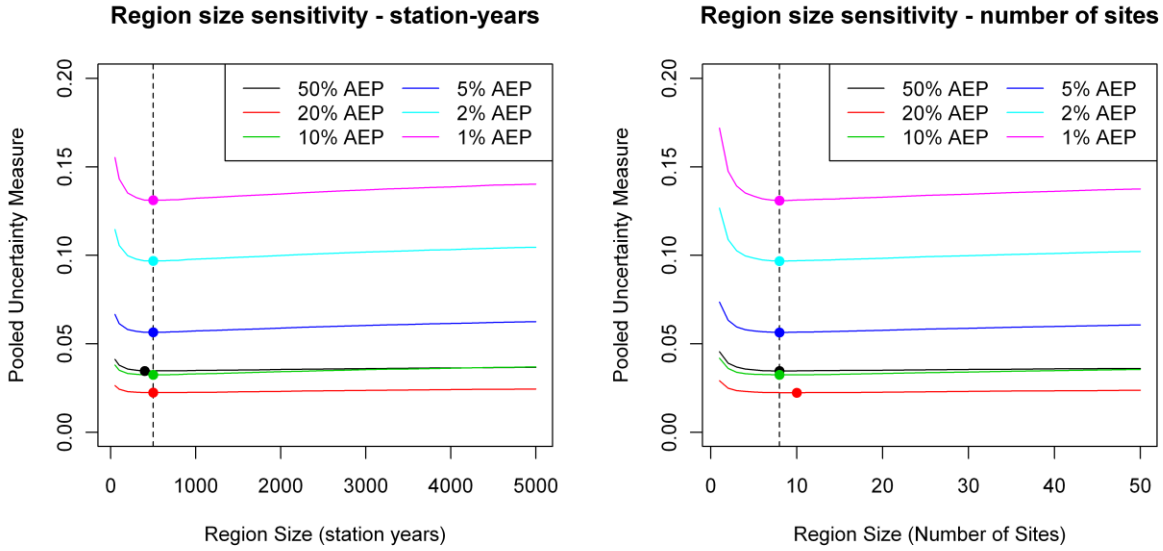

Figure 1 Optimum region size for 24 hour duration rainfall event where the region size is defined in station-years (left panel) and number of stations (right panel).

These results agree well with the findings carried out in Jakob *et* al. (1999a), where it is recommended that region sizes defined in station-years should be approximately five times the return period of interest. If the largest event of interest is the 1% AEP event, this would suggest that 500 station-years of data is required in a region. Kjdelsen and Jones (2009) suggested that a region with 500 station-years is appropriate regardless of the target return period. Defining regions in terms of station-years is attractive as this approach can adapt to different station densities and station record lengths. Given the similar results from both methods, the station-years definition for the region size is used for the remainder of the testing.

3.2. Similarity measures

After finalising the optimum region, different alternatives for defining the ROIs using geographical similarity are investigated. Figure 2 shows the variation in predictive error for each rainfall quantile for the three methods of defining geographical similarity (i.e. circular, circular with elevation, and Ellipse) all having ROIs of 500 station-years. It is clear that all three regionalisation approaches give similar results, although the smallest predictive error in all cases comes from simple circle ROI based on

geographic distance in kilometres. The added complexity of defining regions using elliptical ROIs does not appear to be justified based on these results. Further testing in areas of high topographic relief, not reported here, using the circular ROI with elevation found that the prediction errors are reduced slightly in some cases with this approach.

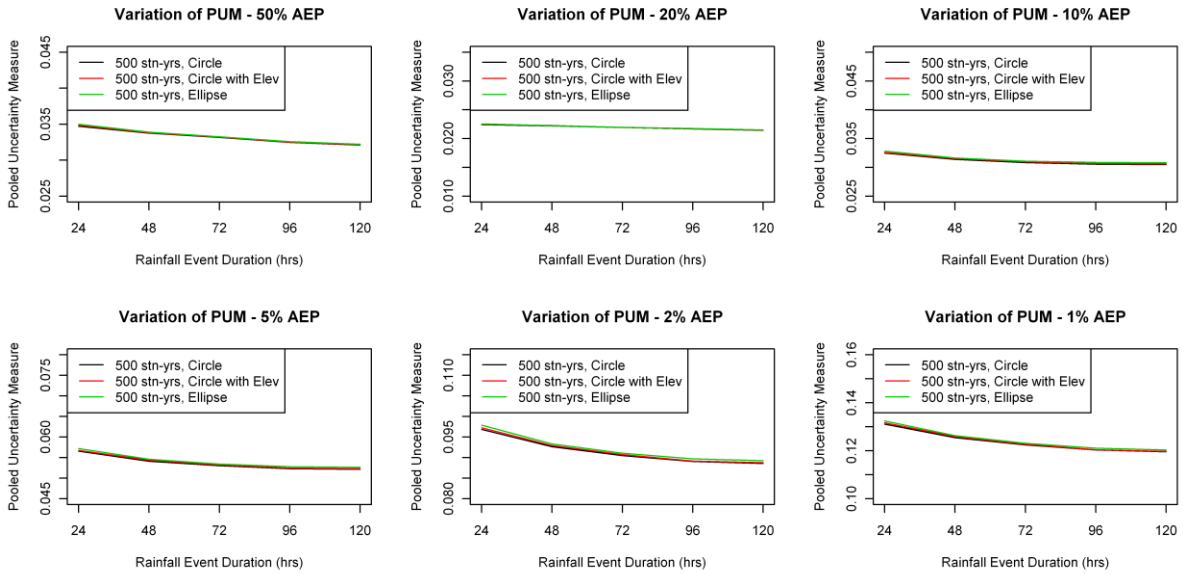

Figure 2 Variation of PUM for different geographic similarity measures

Given the number of site characteristics being considered for defining non-geographic similarity, it is important to understand which characteristics are strongly related to the rainfall statistics of interest. The index rainfall approach to regionalisation assumes that the L-CV and L-skewness for all sites in the region are equal once scaled by the mean of the AMS. For a characteristic to be useful in regionalising L-CV, it can be expected that it will also be a good predictor of L-CV. Therefore linear regression has been used to identify a reduced subset of characteristics that are important in explaining the variation of L-CV.

A best subset approach was used with 10 fold cross validation to find the best combination of predictors for the 24-hour rainfall L-CV. Table 1 shows the best combination of predictors for each subset size. The cross validated prediction errors indicated that the model with 8 predictors was the best. The R^2 values for the best linear regression models are summarised in Table 1. The R^2 values are comparable to those found for linear regressions to predict L-CV for flood frequency analyses in the United Kingdom (Jakob *et* al*.,* 1999b; Kjeldsen and Jones, 2009). The models with 3 predictors (MAR, Latitude and Distance to Coast, denoted as MLD) and 4 predictors (MAR, Latitude, Aspect and Distance to Coast) also appear to be good compromises between model complexity and minimised prediction error.

Table 1 Summary of the best performing linear regression for each subset size

Subset Size	Lat	Long	Elev	MAR	Slope	Asp	Dist to coast	AMS Mean Date	AMS Date Var.	R^2
1				X						0.17
2	X			X						0.30
3	X			X			X			0.34
4	X			X		X	X			0.35
5	X		X	X		X	X			0.36
6	X		X	X		X	X		X	0.37
7	X		X	X	X	X	X		X	0.37
8	X		X	X	X	X	X	X	X	0.38
9	X	X	X	X	X	X	X	X	X	0.38

Testing of all three models (3, 4 and 8 predictors) as non-geographic similarity measures shows that

the 3 predictors model (MLD) provides the minimum prediction error in all cases, apart from rainfall durations longer than 72 hours for the 10% AEP event. For these regionalisation tests, the characteristics were used in the multi-dimensional coordinate system with equal weights. Kjeldsen and Jones (2009) note that this may lead to one of the variables having too much weight in determining the regionalisation. As an alternative, a differential weighting scheme was also developed for the MLD model. Following the method of Kjeldsen and Jones (2009), the weight for the first predictor (MAR) was varied from 0 to 5 (limits chosen arbitrarily), whilst holding the weight for the other two predictors at one. The weighting that resulted in the minimum PUM was then adopted for the first variable. This variable was then kept with this weighting, whilst the weights for the second variable (Latitude) were tested and then the process repeated for the third variable. The optimum weights were found to be 0.13 MAR, 2.61 Latitude and 0.26 Distance to the Coast.

Figure 3 and Table 2 compare the results from the regionalisation using the geographic and non-geographic similarity measures with ROIs of 500 station-years. It is clear that the geographic measure using the circular ROI minimises the PUM value as well as the RMSE and MAE for the estimates of the 10% AEP and 5% AEP rainfall quantiles as shown in Table 2.

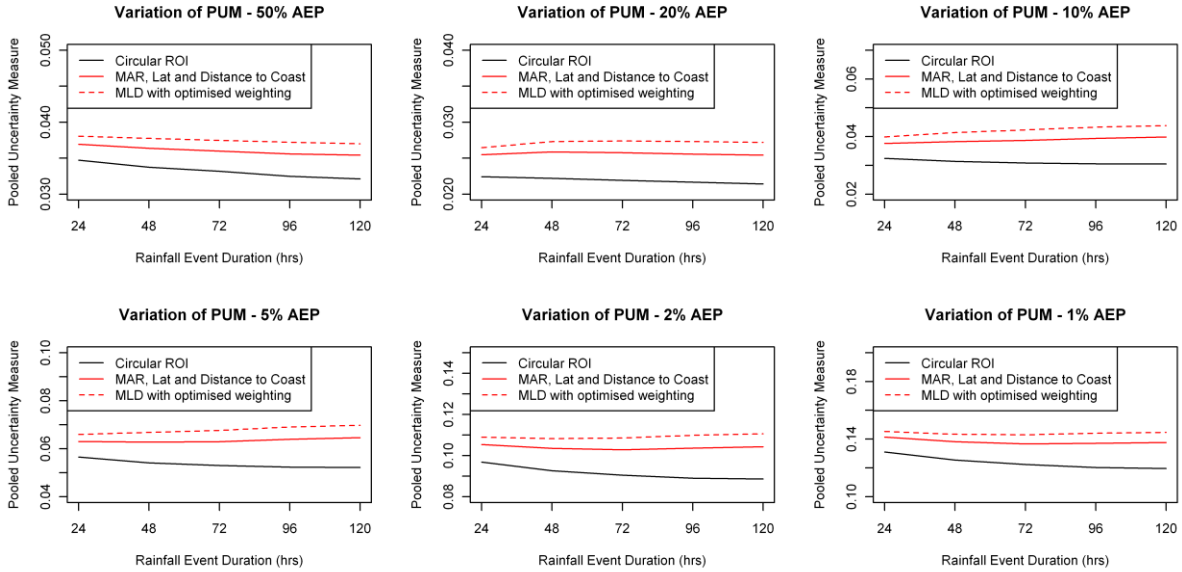

Figure 3 Predictive errors for best performing geographic and non-geographic similarity measures for all rainfall event durations

Table 2 Comparison of evaluation statistics for the best performing geographic and non-geographic similarity measures for 10% and 5% AEPs for the 24 hour rainfall event.

ROI Method	10% AEP			5%AEP		
	RMSE	MAE	PUM	RMSE	MAE	PUM
Circle	4.49	2.92	0.032	9.21	6.06	0.056
Circle with elevation	4.54	2.93	0.033	9.32	6.10	0.057
Ellipse	4.52	2.96	0.033	9.25	6.13	0.057
MLD	5.16	3.34	0.038	10.13	6.70	0.063
Weighted MLD	4.92	3.19	0.036	9.80	6.47	0.061

4. CONCLUSIONS

The analyses and results reported in this paper have been used to recommend the optimum regionalisation strategy for daily rainfall data for the IFD revision project. ROIs of 500 station-years led to the minimum prediction error in all rainfall quantiles from the 50% AEP to the 1% AEP event. Defining the region using the 8 nearest stations gave very similar results but the approach using station-years is recommended due to its flexibility in application across Australia where station record lengths and spatial densities vary considerably.

A number of geographic and non-geographic similarity measures were investigated as methods to define membership of each ROI. It was found that using geographic similarity with a circular ROI provided the best results. The results using distance in either two dimensions (i.e. latitude and longitude) or three dimensions (i.e. latitude, longitude and elevation) were very similar when considered Australia-wide but additional testing in areas of high topographic relief showed that using elevation could provide a slight benefit.

The regionalisation for the IFD revision project will thus be implemented using circular ROIs of 500 station-years with distance defined in three dimensions. Homogeneity tests will then be used to identify stations where the default regionalisation approach needs further refinement.

5. REFERENCES

Burn, D. H. (1990), *Evaluation of regional flood frequency analysis with a region of influence approach*, Water Resour. Res. 26(10): 2257-2265.

Burn, D. H. (1997), Catchment *similarity for regional flood frequency analysis using seasonality measures*, Journal of Hydrology 202: 212-230.

Cunderlik, J. M. and Burn, D. H. (2006), *Switching the pooling similarity distances: Mahalanobis for Euclidean*, Water Resour. Res. 42(W03409).

Di Baldassarre, G., Castellarin, A. and Brath, A. (2006), *Relationships between statistics of rainfall extremes and mean annual precipitation: an application for design-storm estimation in northern central Italy*, Hydrology and Earth System Sciences 10: 589-601.

Green, J., Johnson, F., Taylor, B. and Xuereb K. (2010), *IFD Revision - Proposed Method Draft Report*.

Green, J., Johnson, F. and The, C. (2011), *Revision of the Short Duration Intensity-Frequency-Duration (IFD) Design Rainfall Estimates for Australia,* in Proceedings of 34th IAHR World Congress - Balance and Uncertainty Water in a Changing World. Brisbane, Australia, Engineers Australia.

Haddad, K. and Rahman, A. (2012), *A Pilot Study on Design Rainfall Estimation in Australia using L-moments and Bayesian Generalised Least Squares Regression: Comparison of Fixed Region, Facets and Region of Influence Approach,* EnviroWater Sydney.

Hosking, J. R. M. and Wallis, J. R. (1997), *Regional frequency analysis: an approach based on L-moments,* Cambridge, Cambridge University Press.

Hutchinson, M. F. (1995), *Interpolating mean rainfall using thin plate smoothing splines*, International Journal of Geographical Information Systems 9(4): 385-403.

Jakob, D., Reed, D. W. and Robson, A. J. (1999a), *Selecting a pooling-group (A). Flood Estimation Handbook, Statistical procedures for flood frequency estimation,* Wallingford, Institute of Hydrology. 3: 28 - 39.

Jakob, D., Reed, D. W. and Robson, A. J. (1999b), *Selecting a pooling-group (B). Flood Estimation Handbook, Statistical procedures for flood frequency estimation,* Wallingford, Institute of Hydrology. 3: 153-180.

Jakob, D., Taylor, B. and Xuereb, K. (2005), *A Pilot Study to Explore Methods for Deriving Design Rainfalls for Australia - Part 1*, Hydrometeorological Advisory Service, Bureau of Meteorology. HRS10.

Johnson, F., Green, J., Haddad, K. and Rahman, A. (2012), *Application of Bayesian GLSR to estimate sub daily rainfall parameters for the IFD Revision Project,* Hydrology and Water Resources Symposiusm 2012. Sydney, Australia, Engineers Australia.

Kjeldsen, T. R. and Jones, D. A. (2009), *A formal statistical model for pooled analysis of extreme floods,* Hydrology Research 40(5): 465-480.

Lettenmaier, D. P. and Wood, E. F. (1993), *Hydrologic forecasting,* in Maidment, D. R., "Handbook of Hydrology", New York, McGraw-Hill.

Pilgrim, D. H., Ed. (2001), *Australian Rainfall and Runoff*, The Institution of Engineers, Australia.

Thompson, C. (2011), *HIRDS.V3: High Intensity Rainfall Design System - The method underpinning the development of regional frequency analysis of extreme rainfalls for New Zealand*, Revision 1, Retrieved 5th January 2012, 2012, from http://www.niwa.co.nz/sites/default/files/niwa_hirdsv3_method-rev1.pdf.