

# Application of Bayesian GLSR to estimate sub daily rainfall parameters for the IFD Revision Project

Fiona Johnson

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

Khaled Haddad

School of Computing, Engineering and Mathematics, University of Western Sydney, Sydney, Australia

Ataur Rahman

School of Computing, Engineering and Mathematics, University of Western Sydney, Sydney, Australia

Janice Green

IFD Revision Project Head, Climate and Water Division, Bureau of Meteorology, Canberra, Australia

*The Australian Bureau of Meteorology has recently revised the Intensity-Duration-Frequency (IFD) design rainfall estimates for Australia. Although there is much better coverage of sub-daily rainfall stations than for the previous IFD data, there are still far fewer continuous rainfall stations in Australia than daily rainfall stations. Bayesian Generalised Least Squares Regression (BGLSR) has been used to estimate sub-daily rainfall parameters based on site characteristics and daily rainfall statistics to improve the spatial coverage of sub-daily data.*

*The BGLSR was applied differently across Australia depending on the density of the spatial coverage of the sub-daily rainfall stations. For areas with good spatial coverage, a process akin to a Region of Influence approach was used where the regression equations vary in space. For parts of Australia with sparser data coverage, fixed areas of analysis were used to provide stable estimates of the regression equation coefficients.*

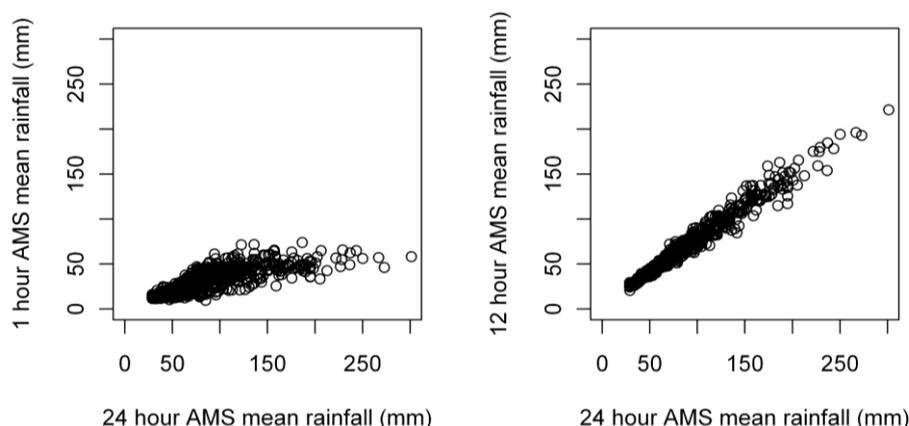
*L-moments were used to summarise the parameters of the Annual Maxima Series (AMS) data, and thus the BGLSR provides predictions of the mean, L-coefficient of variation and L-skewness. It was found that the BGLSR provided excellent estimates of the mean of the AMS, with the most important predictors being the 24 hour AMS mean rainfall and the latitude of the station. This paper describes the application of the BGLSR for the project and presents the results for one analysis area in detail.*

## 1. INTRODUCTION

The Bureau of Meteorology (the Bureau) has recently revised the Intensity-Duration-Frequency (IFD) design rainfall data for Australia. Of particular interest for infrastructure design in urban areas and for small catchments are the design rainfall estimates for sub-daily durations. Information on sub-daily rainfall is hampered by the relative short records and lower station densities when compared to daily rainfall stations. The average record length of Australian daily stations is approximately 65 years whereas for sub-daily rainfall stations it is 19 years. The lower spatial densities of sub-daily data mean that there will be more uncertainty in the gridded design rainfall data and therefore methods to improve the coverage of sub-daily rainfall information are required.

It is known that daily rainfall data can provide some information on properties of sub-daily data. The relationship between the mean of the one day Annual Maximum Series (AMS) and the mean of the

one hour and twelve hours AMS values are shown in Figure 1. It is clear that daily rainfall stations have some information that can be used to improve the coverage of sub-daily rainfall statistics.



**Figure 1 Relationship of daily to sub-daily rainfall statistics for 1 hour and 12 hour durations.**

For the IFD revision project, Bayesian Generalised Least Squares Regression (BGLSR) has been used to transfer this information from daily rainfall to sub-daily rainfall statistics. A Generalised Least Squares Regression (GLSR) approach was chosen because it allows for unequal variances at different sites and can account for spatial correlations between nearby stations, both of which are known to be issues with Australian rainfall data. The advantage of the Bayesian application of the GLSR is that the modelling and sampling error variances can be calculated separately. This is important in the case of rainfall statistics where the sampling error variances tend to be much higher than the model error variance. In the traditional GLSR application this can lead to undesirable negative model error estimates and these can be avoided with the use of the BGLSR approach.

This paper provides summary details of the BGLSR approach, a description of the calibration and model selection for the predictors for the BGLSR model and finally results from application of the model across Australia.

## 2. BAYESIAN GENERALISED LEAST SQUARES REGRESSION

GLSR has been used as a regionalisation approach in a range of studies for streamflow and rainfall data (Tasker and Stedinger, 1989; Madsen and Rosbjerg, 1997; Griffis and Stedinger, 2007) with extensions using a Bayesian framework by Reis *et al.* (2005). In Australia, Haddad *et al.* (2012) used BGLSR to obtain regional relationships to estimate peak streamflow in ungauged catchments and for pilot studies for the IFD revision project (Haddad and Rahman, 2009; Haddad *et al.*, 2011). GLSR is an extension of ordinary least squares (OLS) regression such that the predictand (dependent variable) is calculated from a linear combination of a number of predictor variables (independent variables) with a suitable error model. A number of assumptions are required in an OLS approach, namely that the errors at different sites are independent and that the variances are equal. GLSR does not require these assumptions to be made as an error covariance matrix is constructed that can allow for both unequal variances and cross correlations amongst the sites. It therefore is a more flexible approach to estimate rainfall statistics since it is known that nearby sites are correlated due to the spatial scales of meteorological systems and also that the error variances are likely to be unequal because of the different record lengths at each station.

The basic equations for the BGLSR are presented below with readers referred to Haddad and Rahman (2012) for detailed derivation of the model. In general the predictions for the rainfall statistic,  $y$ , of interest for site  $i$ , are made according to equation (1).

$$y_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ij} + \varepsilon_i + \delta_i \quad (1)$$

Where  $X_{ij}$  ( $j = 1, \dots, k$ ) are the  $k$  predictor variables,  $\beta_j$  are the parameters of the model that must be estimated,  $\varepsilon$  is the sampling error and  $\delta$  is the model error. In contrast to OLS, the errors in the GLSR model are assumed to have zero mean and the covariance structure described in Equation (2).

$$\text{Cov}\{\varepsilon_i, \varepsilon_j\} = \begin{cases} \sigma_{\varepsilon_i}^2, & i = j \\ \sigma_{\varepsilon_i} \sigma_{\varepsilon_j} \rho_{\varepsilon_{ij}}, & i \neq j \end{cases}; \text{Cov}\{\delta_i, \delta_j\} = \begin{cases} \sigma_{\delta}^2, & i = j \\ 0, & i \neq j \end{cases} \quad (2)$$

Where  $\sigma_{\varepsilon_i}^2$  is the sampling error variance at site  $i$ ,  $\rho_{\varepsilon_{ij}}$  is the correlation coefficient between sites  $i$  and  $j$ ,  $\sigma_{\delta}^2$  is the model error variance. For the Bayesian framework introduced by Reis *et al.* (2005), the parameters of the model ( $\beta$ ) are modelled with a multivariate normal distribution using a non informative prior. A quasi analytic approximation to the Bayesian formulation of the GLSR has been developed by Reis *et al.* (2005) to solve for the posterior distributions of the mean and variance for  $\beta$ .

A key step in the application of the BGLSR is the estimation of the error covariance matrix to model the sampling errors. A complication in its estimation is that the error covariance estimator should be independent of the parameter estimates (i.e.  $y_i$ ) (Stedinger and Tasker, 1985). To overcome this difficulty, the error variances (i.e. the diagonal terms of the error covariance matrix) are estimated using the common mean from all sites of the statistic of interest, with the estimate at site  $i$ , weighted by the relative record length of that station (Madsen and Rosbjerg, 1997; Reis *et al.*, 2005). For the inter-site correlations (i.e. off diagonal terms) direct sample estimates cannot be used because of the large sampling uncertainty and the likelihood that the resulting error covariance matrix will not be invertible. Thus the approach of Madsen *et al.* (2002) is used such that the correlations between concurrent years at sites  $i$  and  $j$  are first calculated. These correlations are used to fit an exponential decay curve to the distance-correlation relationship, with this relationship used to estimate all off-diagonal terms.

### 3. DEVELOPING THE BGLSR MODELS FOR THE IFD REVISION PROJECT

The aim of the BGLSR is to predict sub-daily rainfall statistics at the location of daily rainfall stations. L-moments are being used in the IFD revision project to summarise the statistical properties of the AMS data (Hosking and Wallis, 1997) because L-moments are relatively robust against outliers in the datasets. The statistics that are required for the project are:

- Mean of the AMS (also called the index rainfall)
- L-coefficient of variation (L-CV) – defined as the ratio of the L-scale to the mean of the distribution
- L-skewness – defined as the third L-moment divided by the L-scale

These three statistics can then be used to define the parameters of any appropriate probability distribution. It has been shown that the Generalised Extreme Value (GEV) distribution provides a reasonable fit to the majority of Australian AMS data (Green *et al.*, 2010).

The initial work required to apply the BGLSR was to determine the appropriate predictors (i.e.  $X$  from Equation 1) to estimate the three rainfall statistics listed above. A review of literature and meteorological causative mechanisms selected a number of site and rainfall characteristics for use as possible predictors as reported in Green *et al.* (2011). These predictors were:

- Latitude and longitude
- Elevation
- Slope
- Aspect
- Distance from the coast
- Mean annual rainfall
- Rainfall statistics (mean, L-CV and L-skewness) for the 24, 48 and 72 hour duration events

As well as determining the optimum combination of predictor variables, the testing for the BGLSR needed to determine the number of stations to contribute to each regression equation. The maximum number of stations used in the regression is limited to approximately 100 because of the requirement for the error covariance matrix to be invertible. Thus it is necessary to divide the country up into a number of analysis areas. For each analysis area, a regression relationship is developed which can be applied to all stations within the analysis area. Where the density of stations was high, a Region of Influence (ROI) approach (Burn, 1990) was adopted such that each station has its own ROI. This allows the regression equations to smoothly vary across the data dense analysis areas. For sparser

analysis areas, a clustering, or fixed region, approach was adopted such that stations were grouped by spatial proximity into analysis areas with rigid boundaries. All stations in each analysis area were used to derive one regression equation that was then adopted for the predictions at those stations.

Testing of the appropriate predictors and the region sizes was carried out for two study areas - Tasmania and South East Queensland. These were chosen due to the good station density of continuous rainfall stations as well as varied topography which would demonstrate if aspect, slope or elevation were useful predictors. Haddad and Rahman (2012) provide extensive details of the cross validated predictor selection process for each of the study areas. It was found that the most important predictor is the 24 hour rainfall statistic. However performance of the model was not changed significantly by including all predictors so this approach was adopted.

The standard error of prediction (SEP) of the cross validated models decreased as the number of stations used in the ROI increased. This is contrary to the results found for the regionalisation of daily rainfall data reported in Johnson *et al.* (2012) where the regionalisation was optimised for regions of approximately 8 stations. The difference in this case is that the BGLSR is attempting to determine the optimum regression equation rather than neighbours with the same probability distribution so it is reasonable that uncertainty in the regression equations may be reduced through including more stations. For the fixed analysis areas, the results for Tasmania indicated that the number of stations should be maximised. However there is clearly a trade off between the size of the analysis area and the amount of variation expected in the rainfall statistics. Thus the Australia wide analysis focussed on defining a number of analysis areas where meteorological conditions should be reasonably constant.

## 4. APPLICATION OF THE BGLSR

Following the testing of the BGLSR in the cross validation framework as described above, BGLSR models were set up across Australia and used to make predictions for sub daily rainfall durations from 1 hour minutes to 12 hours for the three rainfall statistics of interest (mean, L-CV and L-skewness). This section describes the data sets, delineation of analysis areas and results from one analysis area.

### 4.1. Data set

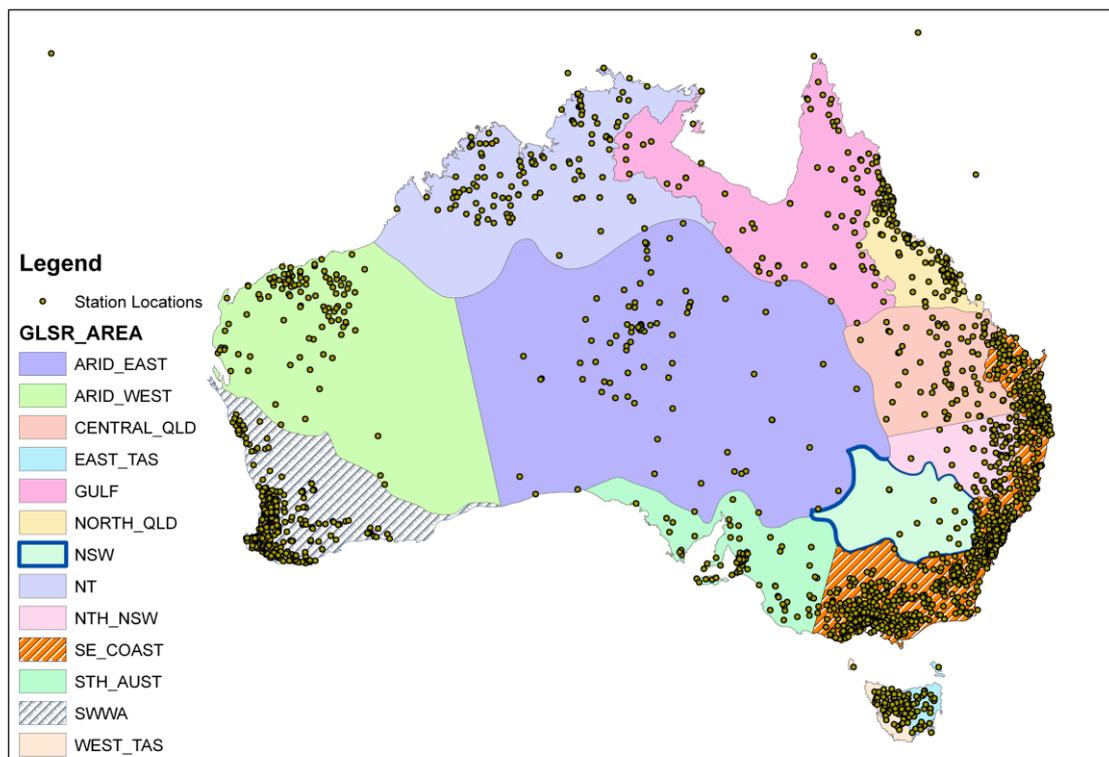
There are 774 Bureau owned rainfall stations that measure sub daily rainfall. Under the *Water Regulations 2008*, the Bureau of Meteorology now has access to water data collected by over 200 other agencies in Australia. From these additional stations, a set of 1575 stations across Australia was identified that met the record length and quality requirements for the project. This data was extensively quality controlled by comparing aggregated daily totals to those from nearby stations and also through statistical analysis of the AMS data sets for step changes, outliers and other errors. By combining the Bureau and other agencies' stations, a total of 2349 stations were available for the BGLSR analysis.

Previous analysis for the IFD Revision Project has indicated that using stations with the long periods of record leads to improved accuracy in the estimates of the rainfall statistics of interest, in particular the higher order moments of L-CV and L-skewness (Jakob *et al.*, 2005; Jakob *et al.*, 2009). When considering the daily rainfall stations a threshold of 30 years of data was adopted for the analysis. However for the sub-daily stations, this would significantly reduce the number of stations that could be used in the project by approximately 80%. To maximize the number of stations and spatial coverage of those stations, a threshold of 8 years has been adopted for continuous rainfall stations to be included in the analysis. The spatial distribution of the stations across Australia is shown in Figure 2.

### 4.2. Analysis area delineation

Ideally the number of stations in each analysis area would be maximised to improve the accuracy of the regression equations. However as discussed above the number of stations is limited by the requirement for the error covariance matrices to be invertible. The delineation of the analysis areas thus needs to balance these two competing requirements. It is also important that stations are grouped into analysis areas where the causative mechanisms for large rainfall events are similar. The

rainfall stations were grouped primarily according to climatic zones by considering the seasonality of rainfall events and mean annual rainfalls. Australian drainage divisions were also used to guide the division of larger climatic zones into smaller areas over which the BGLSR calculations are tractable, such as in the northern tropics where three analysis areas have been adopted (NT, GULF and NORTH\_QLD). The final analysis areas are shown in Figure 2. A 0.2 degree buffer has been used in assigning stations to each analysis area to provide a smooth transition between adjacent areas.



**Figure 2 Analysis areas adopted for the BGLSR – fixed analysis areas are shown with solid fill, ROI areas are hatched. Locations of continuous rainfall stations are also shown.**

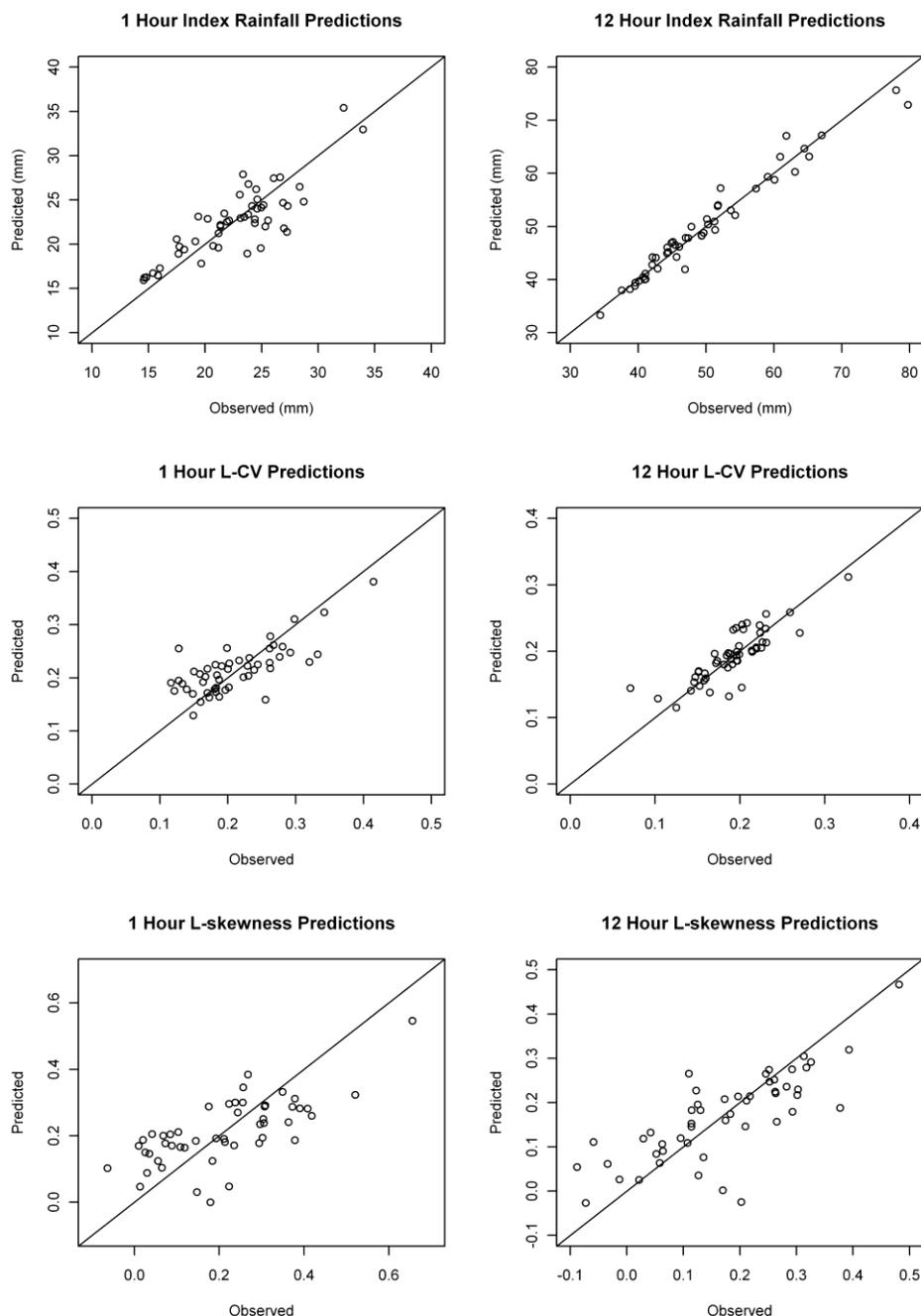
### 4.3. Transformation of predictor variables

To improve the predictions from the BGLSR it is desirable that the distribution of each predictor variable is relatively symmetric and preferably approximately normally distributed. For each analysis area the distribution of the predictor variables from all sites in the area was examined using histograms and quantile-quantile plots. For predictors that appeared to be strongly skewed, a range of transformations were trialled to attempt to reduce the skewness of the variable. The transformations included a natural logarithm, square root transformation and Box-Cox (i.e. power) transformation. In general the log transformation and the Box-Cox transformation were successful in reducing the skewness of the predictors.

### 4.4. BGLSR results

Due to the large number of analysis areas, stations and rainfall event durations this section presents the results for just a single analysis area – the New South Wales area (NSW), shown with a thick blue border in Figure 2. This area has been chosen for the reporting as it clearly demonstrates the benefits of using BGLSR for the IFD revision project when the densities of the sub daily to daily stations are compared. There are 50 continuous rainfall stations in this area, of which 21 are Bureau owned and 29 are owned by other water agencies. In contrast, there are 569 Bureau owned daily read rainfall stations in this area. The additional information that these daily rainfall stations can provide is useful in reducing the uncertainty in the final gridded rainfall estimates.

Figure 3 compares the predicted rainfall statistics and observed data at the continuous rainfall stations in the NSW analysis area for the 1 hour and 12 hours event durations. As found by Haddad and Rahman (2012), the best predictions are obtained for the index rainfalls and for the 12 hour duration. The higher order moments (L-CV and L-skewness) are predicted with less precision. The range of the predicted values captures the range seen in the observations.



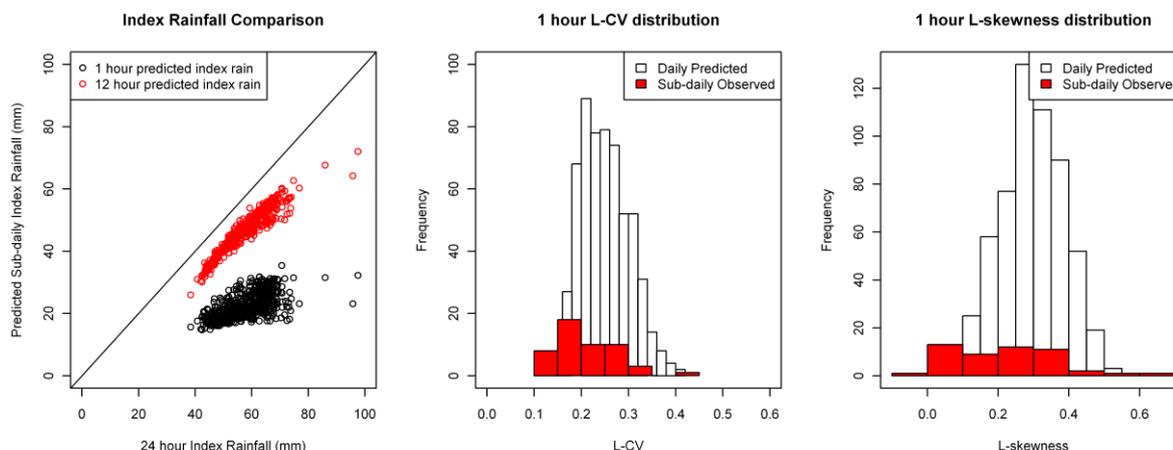
**Figure 3 Comparison of predicted vs observed rainfall statistics for NSW analysis area for the 1 hour and 12 hour duration rainfall events**

#### 4.5. Predictions at daily station locations

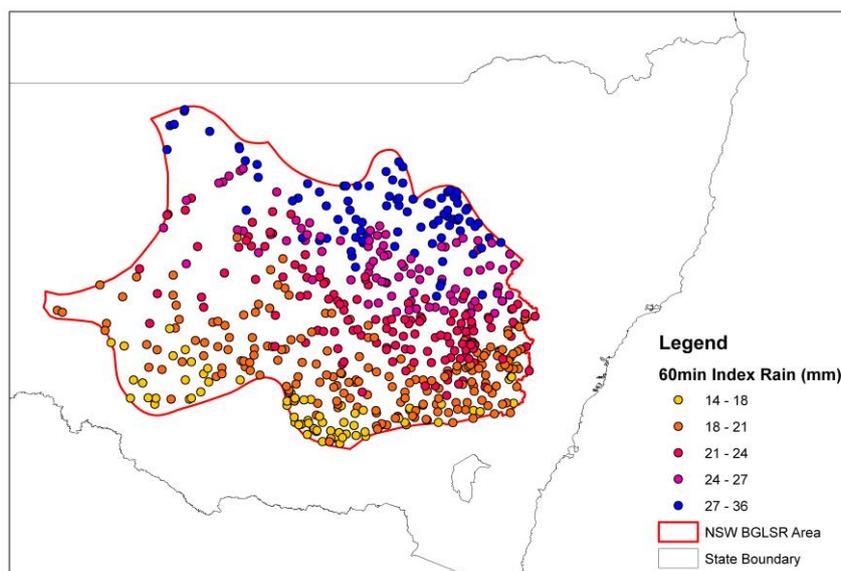
After determining the regression coefficients for the analysis areas, these coefficients are combined with the set of predictors for the daily station locations to produce the estimates of the sub-daily rainfall statistics. There are no observations of the sub-daily rainfall statistics to which these predictions at daily sites can be compared. However “sanity” checks on the values can be carried out by comparing

the estimates to the 24 hour rainfall statistics and to the possible range of values for L-CV and L-skewness (both limited to -1 to 1).

Figure 4 shows the 1 hour and 12 hour index rainfall predictions compared to the observed 24 hour index rainfall at all daily sites in the analysis area. This shows that in all cases the predicted values are consistent with (i.e. lower than) the 24 hour index values. Also shown in Figure 4 are the distributions of the predicted 1 hour rainfall statistics for the daily sites to the observed distributions of these statistics at the sub-daily stations. It is clear that the predicted range of values is consistent with the observed range, although the daily distributions are slightly more symmetric than the sub-daily data..



**Figure 4 Comparison of predicted index rainfalls to observed 24 hour index rainfall and predicted and observed distributions of 1 hour L-CV and L-skewness for NSW analysis area**



**Figure 5 Predicted 1 hour index rainfall values for NSW analysis area**

Finally, Figure 5 shows the spatial patterns of the 1 hour index rainfall values across the NSW analysis area. The first point to note is the difference in the density of the daily station data in this analysis area compared to the sub-daily stations shown in Figure 2. The index rainfall values show a strong south-north increasing gradient as well as a smaller west-east increasing gradient. This is similar to the patterns of both the 24 hour index rainfall and MAR through this analysis area.

## 5. CONCLUSION

BGLSR has been adopted for the IFD Revision Project to provide estimates of sub-daily rainfall

statistics at sites where only daily rainfall data has been recorded. This achieved by regressing site characteristics and rainfall statistics for 24 hour to 72 hour events against the sub-daily rainfall statistics. The rainfall statistics of interest are the mean, L-CV and L-skewness of the AMS. The advantage of the BGLSR approach is that it enables inter-station correlations and unequal site variances, both of which are known to affect Australian rainfall statistics. In addition the method provides full specification of the uncertainty of the estimated parameters and predicted values.

The benefit of using estimated sub-daily rainfall statistics for the project is that the number of locations with sub-daily information is increased from approximately 2300 to approximately 9700 when both the daily and continuous rainfall stations locations are used. This substantially increased density of sub-daily rainfall data will assist in the subsequent gridding of the rainfall quantiles across Australia.

## 6. REFERENCES

- Burn, D. H. (1990), *An appraisal of the "region of influence" approach to flood frequency analysis*, Hydrological Sciences - Journal-des Sciences Hydrologiques 35(2): 149-165.
- 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*. Proceedings of 34th IAHR World Congress - Balance and Uncertainty Water in a Changing World. Brisbane, Australia, Engineers Australia.
- Griffis, V. W. and Stedinger, J. R. (2007), *The use of GLS regression in regional hydrologic analyses*, Journal of Hydrology 344: 82-95.
- Haddad, K. and Rahman, A. (2009). *A pilot study on design rainfall estimation using generalised least squares regression*. Research Report prepared for Australian Bureau of Meteorology, School of Engineering, UWS, 85 pp.
- 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.
- Haddad, K., Rahman, A. and Green, J. (2011), *Design Rainfall Estimation in Australia: A Case Study using L moments and Generalized Least Squares Regression*, Stochastic Environmental Research & Risk Assessment, 25, 6, 815-825.
- Haddad, K., Rahman, A. and Stedinger, J. R. (2012), *Regional Flood Frequency Analysis using Bayesian Generalized Least Squares: A Comparison between Quantile and Parameter Regression Techniques*, Hydrological Processes 26(7): 1008-1021.
- Hosking, J. R. M. and Wallis, J. R. (1997), *Regional frequency analysis: an approach based on L-moments*. Cambridge, Cambridge University Press.
- 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.
- Jakob, D., Xuereb, K., Meighen, J. and Taylor, B. (2009), *A Pilot Study to Explore Methods for Deriving Design Rainfalls for Australia - Part 2*, Hydrometeorological Advisory Service, Bureau of Meteorology. HRS11.
- Johnson, F., Xuereb, K., Jeremiah, E. and Green, J. (2012), *Regionalisation of rainfall statistics for the IFD Revision Project*. Hydrology and Water Resources Symposium 2012. Sydney, Engineers Australia.
- Madsen, H. and Rosbjerg, D. (1997), *Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modelling*, Water Resources Research 33(4): 771-781.
- Madsen, H., Mikkelsen, P., Rosbjerg, D. and Harremoes, P. (2002), *Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series statistics*, Water Resources Research 38(11): 1239.
- Reis, D. S. J., Stedinger, J. R. and Martins, E. S. (2005), *Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation*, Water Resources Research 41: W10419.
- Stedinger, J. R. and Tasker, G. D. (1985), *Regional Hydrologic Analysis 1. Ordinary, Weighted and Generalized Least Squares Compared*, Water Resources Research 21(9): 1421-1432.
- Tasker, G. D. and Stedinger, J. R. (1989), *An Operational GLS Model for Hydrologic Regression*, Journal of Hydrology 111: 361-375.