Patching and Disaccumulation of Rainfall Data for Hydrological Modelling

Bennett, N.D. [1], L.T.H. Newham [1] B.F.W. Croke[1, 2] and A.J. Jakeman [1]

[1] Integrated Catchment Assessment and Management Centre, The Fenner School of Environment and Society, The Australian National University, Canberra
[2] Department of Mathematics, The Australian National University, Canberra
Email: u4115489@anu.edu.au

Keywords: *rainfall data, patching, disaccumulation, error reduction*

EXTENDED ABSTRACT

Hydrological models are increasingly relied upon for decision-making in catchment planning and management. But the impacts of errors in input data are seldom fully considered in the analysis of these and many other environmental models, even though they can be a potentially large source of model uncertainty.

Precipitation data are an essential input for many hydrological modelling investigations but there are potentially significant errors in catchment areal rainfall estimates. Estimating rainfall at catchment scales is a difficult process, due to variations of rainfall both in time and space. Identifying, and preferably reducing, the errors present in catchment rainfall estimates is a necessary step towards determining and improving the accuracy of hydrological models. Sources of errors within areal rainfall estimates can include: extrapolating point measurements to spatial estimates, missing data, accumulation errors (rainfall is allowed to accumulate in the gauge for several days before it is measured), along with other measurement and timing errors.

The focus of this paper is on patching and disaccumulation errors. Patching errors occur when missing rainfall within a record is estimated. Disaccumulation errors occur when accumulated rainfall is allocated to the preceding days within the period. In this paper we describe different rainfall patching and disaccumulation methods for dealing with accumulated and missing rainfall. We focus on conceptually simple and computationally inexpensive methods. The patching methods considered in the paper include:

- Nearest Neighbour by Distance (ND): selecting the closest gauge with data.
- Nearest Neighbour by Correlation (NC): selecting the neighbouring gauge that has the highest correlation with the gauge to be patched.
- Inverse Distance Weighted (IDW): Using multiple neighbouring gauges weighted by distance.
- Average of Gauges Selected by Correlation (A): all gauges with a correlation larger than 0.7 are averaged.
- Weighted Average of Gauges Selected by Correlation (WA): all gauges with a correlation larger than 0.7 are averaged, weighted by the correlation level.

For disaccumulation tests, the following methods were tested:

- Allocating total rainfall accumulated equally over the days in the accumulation period.
- Observing the temporal distribution of rainfall at a neighbouring station (selected by distance or correlation) during the accumulation period and using this information to allocate accumulated rainfall.

These techniques are tested on catchments on the southeast coast of Australia. Of the techniques trialled, correlation techniques were observed to be an ideal station selection variable for patching processes, with the lowest root mean square error achieved using the WA method. Scaling of data using long-term means is not recommended particularly under conditions when the rainfall is highly variable, as observed in the test catchments. Further studies with short term means need to be completed to see if the method is a viable alternative. Using the total rainfall over an accumulation coupled with the rainfall distribution from a neighbouring station selected by correlation was found to achieve the lowest error for disaccumulation operations.

1. INTRODUCTION

Hydrological models are increasingly relied upon for decision-making in catchment planning and management (Croke and Jakeman, 2001). It is therefore necessary to understand not just errors that occur as part of the hydrological modelling process, but also input errors. The impacts of systematic and random errors are seldom fully incorporated in the analysis of environmental models but they can be a potentially large source of error. Careful consideration as to their potential effects and of methods for their control are needed. Treatment of input errors is all the more important for prediction in ungauged basins where models cannot be calibrated directly from gauged data and where input errors may have a greater impact.

Historical rainfall records are measured at typically widely separated rain gauges on a daily basis. The availability of these measurements varies in time and space, with not all gauges operating at the same time. Using this data to generate estimates of a catchment's rainfall can result in significant errors. This can include systematic measurement error in the data for individual gauges, systematic errors caused by spatial and elevation variations, as well as random errors affected by gauge density, measurement mistakes and how representative the gauges are for their surrounding area. (Hall and Barclay, 1975) This paper looks at two of these errors: (i) those caused by missing data in gauges and (ii) rainfall accumulation.

Missing rainfall records often need to be estimated to complete rainfall records or assist in areal rainfall calculations. This however introduces errors into the records as the true measurement of rainfall is unknown. The magnitude of the error depends on the spatial and temporal variability of the rainfall and the density of gauges in the vicinity. Further errors can also be introduced into rainfall records when data accumulation occurs. Data accumulation is a form of missing data, occurring when rainfall over a number of days is allowed to accumulate in a gauge and is then allocated to the first day the gauge is next read.

The objective of this study is to evaluate rainfall patching and disaccumulation methods with a particular focus on simple methods that can be easily computed. Each method is tested to determine its relative accuracy, a necessary step before any method can be used in modelling processes. Section 2 provides a brief background on previous studies. Sections 3 and 4 describe the methods used in the study and their application in the test catchments. Key results and a discussion of those results are presented in Sections 5 and 6, respectively.

2. REVIEW OF PATCHING AND DISACCUMULATION PROCESSES

The estimation of missing rainfall data has long been a problem in hydrology (e.g. Wei and McGuiness, 1973; and Simanton and Osborn, 1973) and the problem has been extensively investigated. However, it is unlikely there is one method best for all applications and optimisation for each use may be required (Teegavarapua and Chandramoulia, 2005). Traditional approaches to patching of rainfall include nearest neighbour or inverse-distance weighted mean where the closest rain gauge or the weighted average of the nearby gauges are used to fill the missing data. Problems with these methods are generally related to the use of distance as a station selection variable, which may be a poor choice due to complex topographic and orographic effects on rainfall (McGuen, 1998). Research on patching methods has focused on different weighting schemes for inverse distance weighting or through improving methods of station selection. Research has extended to the use of the relatively complicated coordinate system methods where the gauge closest to the origin in each quadrant is averaged (McGuen, 1998). Other options to replace distance for station selection identified include the correlation of gauges, which should theoretically enable the optimal selection of gauges based on their relationship to each other. Teegavarapua and Chandramoulia (2005) tested this method on a region within Kentucky, USA and found it to be a significant improvement on distance-based selection techniques and recommended its use.

More complicated methods for rainfall patching have also been tested. Makhuva et al. (1997) used techniques based on multiple linear regression to estimate missing rainfall records with the added benefit that predictions could still be made when the control data was also missing. However extensive computation is required for these calculations and they could be negatively affected by changing climatological relationships. Spatial kriging of rainfall data has also been tested (Seo, 1996) but these are computationally heavy methods and often result in only small increases in accuracy (Teegavarapua and Chandramoulia, 2005). Neural networks have been applied to patching of missing rainfall (e.g. French et al., 1992). Such methods require training the data based on past relationships, may result in overfitting, are computationally expensive and difficult to interpret.

In contrast to the patching of missing rainfall, for accumulation periods there is extra information available in the record: the total rainfall over that period (assuming losses from the gauge are insignificant). This information can be used to reduce the errors involved in disaccumulation methods and yield less error than direct patching of rainfall.

3. METHOD

Simple rainfall patching techniques were tested to determine their accuracy. The following methods were included:

Nearest Neighbour by Distance (ND): selecting the closest gauge with data.

Nearest Neighbour by Correlation (NC): selecting the neighbouring gauge that has the highest correlation with the gauge to be patched.

Inverse Distance Weighted (IDW): using multiple neighbouring gauges weighted by distance (1).

$$P_c = \frac{\sum_i P_i d_{ci}^{-k}}{\sum_i d_{ci}^{-k}}, \qquad (1)$$

Where P_c is the rainfall for the gauge to be patched, P_i is a neighbouring gauge, d_{ci} is the distance between the gauges and k is a weight known as the friction distance that ranges from 1.0 - 6.0. (Vieux, 2001) The most commonly used value for k is 2 (Teegavarapua and Chandramoulia, 2005) which is applied here.

Average of Gauges Selected by Correlation (A): All gauges with a correlation larger than 0.7 are averaged.

Weighted Average of Gauges Selected by Correlation (WA): All gauges with a correlation larger than 0.7 are averaged weighted by the correlation level (2).

$$P_c = \frac{\sum_i P_i r_{ci}}{\sum_i r_{ci}}, \qquad (2)$$

where r_{ci} is the Pearson correlation coefficient for the two gauges found by:

$$r_{ci} = \frac{1}{n s_c s_i} \sum_k (P_{i,k} - \overline{P}_i)(P_{c,k} - \overline{P}_c), \qquad (3)$$

where $\overline{P}$ is the average rainfall for the time series, s is the standard deviation and n is the length of the time series.

Because rainfall is closely related to topographic features, neighbouring rain gauges may receive different magnitudes of rainfall, especially in environments with sparsely located rainfall gauges. This can cause significant problems since the overall magnitude of rainfall at neighbouring stations can vary significantly. One method to overcome this is to scale the rainfall based on average rainfall. The scaling process weights the data based on long term means at the gauges (4). Thus

$$P_c = \left(\frac{\overline{P}_c}{\overline{P}_i} \right) P_i, \qquad (4)$$

where $\overline{P}$ is the mean rainfall calculated for a period of time (long term monthly means were used in this study.) The scaling process can be seen as a simple implementation of the spatial interpolation used in kriging or smoothing splines, in this case just using the data available at gauges rather than fitting a complete surface.

Tests were completed for each rainfall patching method, both with and without scaling. Disaccumulation tests were also completed on the same data as the patching tests. The following tests were completed:

Allocating equally (AE): the total rainfall accumulated is equally allocated over the days in the accumulation period.

Distance (D): Observing rainfall distribution at a neighbouring station selected by distance during the accumulation period and using this information to allocate accumulated rainfall.

Correlation (C): Observing rainfall distribution at a neighbouring station selected by long-term correlation during the accumulation period and using this information to allocate accumulated rainfall.

Tests were completed for both patching and disaccumulation using the R software environment, an open source programming language providing data processing, statistical and graphical methods, making it ideal choice for these tasks (R Development Core Team, 2007).

The tests were completed by completely recalculating selected records on days when there was complete (i.e. non-accumulated) data available. The constructed records were then compared against the actual records to evaluate the performance of the different techniques.

The disaccumulation tests were completed by using the total rainfall over periods of 2-5 days as the accumulated rainfall, which was then disaccumulated to completely reconstruct a station's record.

The performance of all methods was compared using three error statistics: root mean square error (RMS), square root transformed root mean square error (RMSSQRT) and bias. These three commonly used measures allow analysis of errors biased towards higher and lower flows as well as determination of any over or underestimation.

4. STUDY AREA

The study area is focused on the Eurobodalla Shire on the south east coast of NSW. This area includes the Clyde (number 216), Moruya (217) and Tuross (218) River catchments. Data from 114 rainfall gauges were available between the periods of 1848 and 2006 with relatively sparse coverage over some areas (Figure 1.)

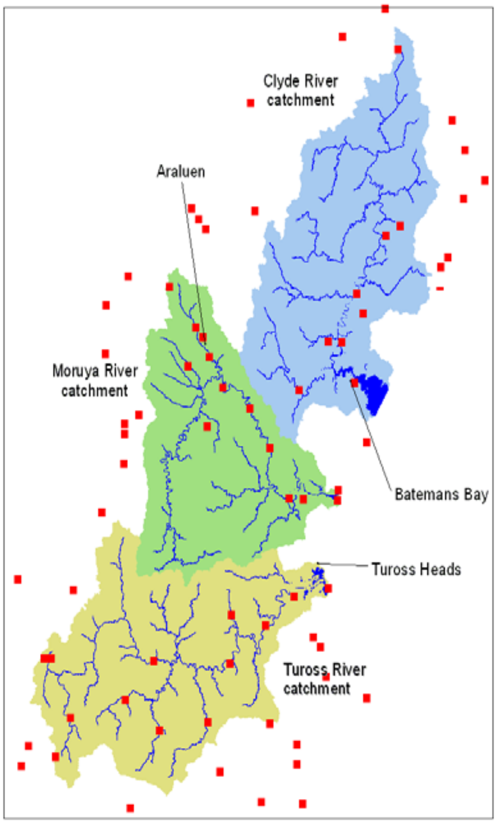

Figure 1. Map of study area and rain gauges (squares)

5. RESULTS

5.1. Rainfall Patching Results

The results from the patching tests are shown for four example gauges. The results presented in Tables 1 and 2 indicate that no method performed best on all of the gauges. The NC method had the smallest error (RMS and RMSSQRT) for gauge 69099 while the weighted average of highly correlated stations had the smallest error for all other gauges. The fact that the correlation methods performed best indicates that it is a good decision to select stations on metrics other than distance. Bias results varied for all of the gauges tested however it remained small for all of the methods.

Overall the results for scaled data had a larger RMS error and did not support the use of this method. The fact that just using the raw station data achieves a lower error is a surprising result as it would be expected that gauges with higher average rainfall should on average have higher event rainfall. The cause of this problem could possibly be traced to the difference between the long term means that were used to scale the data and the short term relationship between the gauges. (e.g. Figure 2).

Table 1. Error results for scaled tests

	RMS error (mm)			
	69099	**68079**	**69052**	**68021**
ND	3.91	11.12	8.26	8.73
NC	3.16	11.12	7.23	8.73
IDW	4.10	11.15	6.21	5.85
A	3.82	11.14	5.51	5.75
WA	3.76	10.99	5.49	5.75
	RMSSQRT error (mm)			
	69099	**68079**	**69052**	**68021**
ND	0.83	1.48	1.12	1.05
NC	0.76	1.48	1.09	1.05
IDW	0.83	1.50	0.88	0.92
A	0.81	1.51	0.85	0.82
WA	0.81	1.49	0.83	0.82
	Bias (mm)			
	69099	**68079**	**69052**	**68021**
ND	0.26	-0.62	0.35	-0.15
NC	0.4	-0.62	0.32	-0.15
IDW	0.25	0.21	-0.05	-0.03
A	0.37	-0.54	-0.06	-0.04
WA	0.3	-0.63	0.07	-0.04

A further effect of patching rainfall was identified as significant changes to the number of wet days within a rainfall records. This occurs when the patching process result in the estimation of small rainfall events on days when there was not any rainfall. Figure 3 details these results for gauge 69052 where rainfall distribution curves have been calculated for records constructed using each of the methods. Overall it can be seen that the average methods (IDW, A and WA) perform significantly worse in this respect due to the multiple gauges used for each patching process.

Table 2. Error results for unscaled tests

	RMS error (mm)			
	69099	**68079**	**69052**	**68021**
ND	3.46	10.62	7.58	7.7
NC	3.41	10.62	7.13	7.7
IDW	4.12	10.65	6.15	6.00
A	3.88	10.65	5.34	5.62
WA	3.72	10.44	5.2	5.62
	RMSSQRT error			
	69099	**68079**	**69052**	**68021**
ND	0.81	1.43	1.10	1.03
NC	0.79	1.43	1.08	1.03
IDW	0.83	1.44	0.88	0.91
A	0.83	1.45	0.83	0.82
WA	0.81	1.42	0.84	0.82
	Bias (mm)			
	69099	**68079**	**69052**	**68021**
ND	0.43	0.15	-0.37	0.05
NC	0.2	0.15	-0.24	0.05
IDW	0.25	0.21	-0.05	-0.06
A	0.28	0.25	0	-0.11
WA	0.31	0.29	-0.02	-0.11

Although the significant change in number of wet days has only a small effect on the overall RMS error it could cause more significant effects in modelling processes, such as rainfall-runoff modelling that uses catchment moisture deficit accounting (e.g. Croke and Jakeman, 2004).

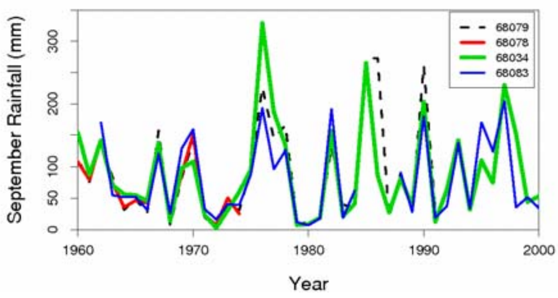

Figure 2. September rainfall results for gauge 68079 and neighbouring gauges. (long term relationship: 68079 > 68034 > 68083 > 68078)

5.2. Rainfall Disaccumulation Results

The error results for the disaccumulation tests are shown in Tables 3, 4 and 5. From the results it can be seen that the correlation method had the lowest overall error with the same results as the distance method on some gauges and slightly better on other gauges.

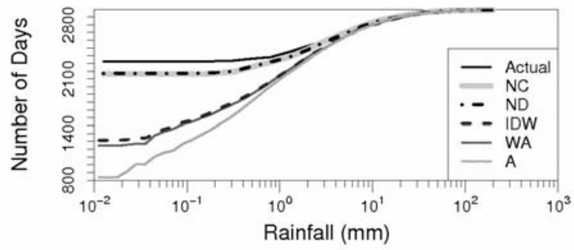

Figure 3. Rainfall distribution curve results for patching tests on 69052

Generally, results using a five-day period had a larger error than tests using a two-day period. But there is not a consistent increase with error as the number of days is increased. Figure 4 shows the results for the correlation tests performed on gauge 69099. This demonstrates that the disaccumulation method performs quite well for larger rainfall events, but there can be significant variance at lower rainfall levels. The bias results (Table 5) indicate a very low bias for all of the disaccumulation methods tested.

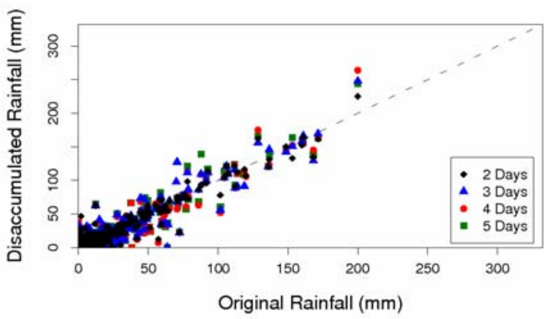

Figure 4. Rainfall disaccumulation results (69099)

Table 3. RMS error (mm) results for disaccumulation tests on four gauges

	Days	69099	68079	69052	68021
AE	2	3.97	7.57	6.41	6.14
	3	4.85	8.9	7.71	7.5
	4	5.27	9.72	8.47	8.18
	5	5.48	9.96	8.18	8.67
D	2	3.12	6.36	5.82	5.56
	3	3.32	7.1	5.95	6.25
	4	3.26	6.48	6.03	5.84
	5	3.45	7.07	6.15	5.84
C	2	3.10	6.36	4.59	5.56
	3	3.31	7.1	5.18	6.25
	4	3.22	6.48	5.06	5.84
	5	3.44	7.07	5.05	5.84

When comparing the correlation test results against the patching test results from the previous section, it can be seen that the disaccumulation test achieves a lower RMS error for most of the time periods selected. For a few gauges the errors are slightly larger than direct patching. It could be generally recommended, however, to use the disaccumulation methods when the extra information is available.

Table 4. RMSSQRT error results for disaccumulation tests on four gauges

	Days	69099	68079	69052	68021
AE	2	0.78	1.05	0.91	0.90
	3	0.96	1.29	1.12	1.12
	4	1.07	1.42	1.25	1.24
	5	1.12	1.51	1.34	1.33
D	2	0.73	1.02	0.88	0.90
	3	0.94	1.06	1.03	0.98
	4	0.87	1.04	0.97	0.97
	5	1.05	1.07	1.05	0.97
C	2	0.71	1.02	0.79	0.90
	3	0.92	1.06	0.85	0.98
	4	0.86	1.04	0.84	0.97
	5	1.03	1.07	0.86	0.97

Table 5. Bias results (mm) for disaccumulation tests on four gauges

	Days	69099	68079	69052	68021
AE	2	0.00	-0.01	-0.02	-0.01
	3	-0.01	-0.02	-0.02	-0.02
	4	-0.01	-0.02	-0.03	-0.02
	5	-0.02	-0.03	-0.03	-0.03
D	2	0	0	0	0
	3	0	0	0	0
	4	0	0	0	0
	5	0	0	0	0
C	2	0	0	0	0
	3	0	0	0	0
	4	0	0	0	0
	5	0	0	0	0

6. DISCUSSION

The methods tested in this study were conceptually simple ones that allowed the quick computation of rainfall patching and disaccumulation and the associated errors. The errors calculated also give an indication of the magnitude of error for the areas between rain gauges when areal estimates are generated.

The results indicate that significant errors could be present in the rainfall records between gauges. RMS errors were observed to be within 3 -12 mm for rainfall records which could add significant errors to rainfall records. A reduction in RMS and the RMSSQRT error was observed for some gauges when using correlation as a station selection variable instead of distance. The relationship between distance and correlation for all gauges is shown in Figure 5. Although a direct relationship between distance and correlation can be seen there is also significant variance where some gauges further apart have a higher correlation and close gauges have lower correlation.

The scaling method often used in rainfall patching was observed to increase errors in patching for most gauges. Variance between short term and long term rainfall ratios between gauges were observed to be significant, indicating that the scaling method may not be ideal for variable Australian climate. There could be potential, however, in using short term means for scaling data with the possible use of rainfall surfaces to overcome the significant effect of missing data on short term means.

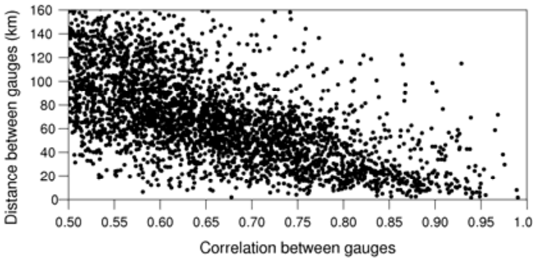

Figure 5. Relationship between correlation and distance for all gauges

Further consideration also needs to be applied to the types of errors generated by different rainfall patching methods. Methods that utilise multiple stations tend to have smaller maximal errors but may suffer from more errors at low rainfalls. Depending on the modelling application desired, different patching methods may perform better. For example, reducing high rainfall errors may be desirable for water quality modelling while low rainfall errors could potentially be of greater interest for water allocation and supply investigations.

Identification of accumulation periods and use of the total rainfall during the disaccumulation was seen to reduce the error as compared to patching processes. Once again using correlation as a selection variable reduced the error for some gauges and it is a viable method to use instead of distance-based methods. Only periods between two and five days were tested as it is likely that for longer periods of accumulation the value of rainfall measured in the gauge would be of lesser reliability. These time periods are also the typical length over which accumulation occurs in BoM gauges.

The results presented are generally applicable to all of the gauges within the sub catchments. Using correlation as a station selector variable could potentially reduce the error; however for other gauges the performance was similar to using distance. A possible extension to the process would be to use seasonal correlations to account for changing climatic patterns by season. Scaling by long term means did not necessarily reduce the error of the patching results and could potentially have the opposite effect.

Further testing will be required to estimate the errors in rainfall estimates and fully compare the various methods available. Possible options could include investigating the difference between the simple and more complicated techniques such as kriging and neural networks. Another possible approach would be to remove some of the gauges from the analysis and determine the resulting effect on the rainfall estimate. Other possible estimation techniques that could be investigated include using thin plate smoothing splines, which could be used to determine longer term relationships between gauges (Hutchinson, 1995) offering an alternative that is in-between simple distance weighted and complex kriging techniques.

7. CONCLUSION

In this study different rainfall patching and disaccumulation methods were tested on a catchment on the southeast coast of Australia. Of the techniques trialled, correlation was observed to be the best station selection variable for patching processes, with the lowest error achieved using a correlation weighted average of neighbouring gauges. Scaling of data using long-term means is not recommended, particularly under conditions when the rainfall is highly variable as observed in this catchment. Further studies with short term means need to be completed to see if the method is a viable alternative. Using the total rainfall over an accumulation, coupled with the rainfall distribution from a neighbouring station selected by correlation, was found to achieve the lowest error for disaccumulation operations.

8. REFERENCES

Croke, B.F.W and Jakeman, A.J. (2001), Predictions in catchment hydrology: an Australian perspective, Australian Journal of *Marine and Freshwater Research*, 52, 65-79.

Croke, B.F.W. and Jakeman, A.J. (2004), A catchment moisture deficit module for the IHACRES rainfall-runoff model, *Environmental Modelling and Software*, 19, 1-5.

French, M.N, Krajewski, W.F. and Cukyendal, R.R. (1992), Rainfall forecasting in space and time using a neural network,. *Journal of Hydrology*, 137, 1- 37.

Hall, A.J. and P.A. Barclay (1975), Methods of determining areal rainfall from observed data, paper presented at A National Symposium on Hydrology, Australian Academy of Science, Canberra, Australia, November 25–27.

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

Makhuva, T., Pegram, G., Sparks, R. and Zuchinni, W. (1997), Patching rainfall using regression methods. 1. Best subset selection, EM and psuedo-EM methods: theory, *Journal of Hydrology*, 8, 289-707.

McCuen R. (1998), Hydrologic Analysis and Design, Prentice-Hall, Eaglewood Cliffs, NJ.

R Development Core Team, (2007), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna Austria.

Seo, D. (1996), Nonlinear estimation of spatial distribution of rainfall - an indicator cokriging approach, *Stochastic Hydrology and Hydraulics*, 10, 127-150.

Simanton, J.R. and Osborn, H.B. (1980), Reciprocal-distance estimation of point rainfall, *Journal of Hydraulic Engineering Division*, 1980, 106HY7.

Teegavarapua R. and Chandramoulia, V. (2005). Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records, *Journal of Hydrology*, 312: 191-206.

Wei, T.C. and McGuiness, J.L. (1973), Reciprocal distance squared, a computer technique for estimating area precipitation, Technical Report ARS-Nc-8, US Agricultural Research Service, North Central Region, Ohio.

Vieux, B.E. (2001), Distributed Hydrologic Modeling using GIS, Water Science and Technology Library, Kluwer Academic Publishers, Dordrecht.