Stochastic downscaling of rainfall for use in hydrologic studies

R. Mehrotra, **Ashish Sharma and Ian Cordery**

School of Civil and Environmental Engineering, University of New South Wales, Australia

Abstract: During the past decade downscaling has become an established field to relate atmospheric circulation to surface variables for forecast and prediction of the regional climate. This article is a part of an ongoing study to develop statistical downscaling methods to enable downscaling of rainfall at hydrologically relevant spatial scales. This paper reviews the development and recent advances in stochastic downscaling techniques. Emphasis is given on application, advantages and limitations of currently used approaches. Results from an application of a non-parametric downscaling technique to select predictors (based on leave-one-out cross validation) for downscaling of point daily rainfall occurrence from area-average rainfall occurrence and temperature from ten locations near Sydney are presented. Results of the study indicate that the area-average temperature range and area-average rainfall occurrence on the current day are good predictors of point rainfall occurrence, for the Sydney region.

Keywords: *downscaling, atmospheric circulation, leave-one-out cross validation and rainfall prediction*

1. INTRODUCTION

It is well known that there is a strong sensible physical linkage between climate on the large scale and weather on the local scale. Downscaling has emerged as a potential tool to relate atmospheric circulation patterns to surface variables for generation of series, for forecasting and for predicting the regional climate from large-scale circulation data. The general limitations, theory and practice of downscaling are well described in the literature (Wilby, 1994; 1997; Wilby and Wigley, 1997; Yarnal et al., 2001). A common classification describes downscaling techniques using two categories: i) Dynamical downscaling (based on physical dynamics and more commonly known as regional climate modelling) and ii) statistical or empirical downscaling.

The list of large-scale variables used in downscaling is very broad and comprise precipitation, temperature, sea level pressure (SLP) and geo-potential heights at 500-hPa level, 800/850 hPa and 700 hPa. Useful summaries of downscaling techniques and the predictors used are presented in Wilby et al. (1998), Wilby and Wigley (1997) and Yarnal et al. (2001). However, outside of passing references in many studies to the effect that a range of predictors were evaluated, there is little systematic work that has explicitly evaluated the relevant skill of different atmospheric predictors (Winkler et al., 1997). The present study provides a useful step in this direction.

This paper is organised as follows. The next section reviews empirical downscaling approaches. Limitations and challenges in downscaling and future research scope are discussed in sections 3 and 4. Section 5 gives a preliminary application of a non-parametric downscaling model to select the best predictors by downscaling the point rainfall occurrence at multiple stations and comparing the results. The paper concludes with a discussion of the approaches to be investigated and formulated in later stages of this work.

2. OVERVIEW OF EXISTING EMPIRICAL DOWNSCALING APPROACHES

Empirical downscaling involves developing a relationship between large-scale atmospheric predictor variables and circulation characteristics (e.g. area average precipitation or temperature, mean-sea level pressure, and vorticity) with local-scale meteorological variables (e.g. precipitation, temperature, and evaporation). It relies on the principle that there is a correspondence between predictor variables and the local-scale climate. The concept of downscaling does not imply that the local climate would be solely determined by the large-scale processes. Rather, the local climate is seen as a random process conditioned upon a driving large-scale climate regime.

There are many ways to categorize empirical downscaling techniques. von Storch (1999); Yarnal et al. (2001) and Prudhomme et al. (2002) provide an excellent review and discussions of various downscaling techniques. A common approach classifies these techniques using two broad categories: i) deterministic (regression analysis based) approaches and ii) stochastic (weather generator based) approaches.

2.1. Deterministic approaches

Deterministic downscaling approaches involve linear/non-linear relations either between variables at local and large scales, or between atmospheric patterns and the local-scale variables of interest.

In the simplest form, large scale and local scale variables are the same (Sailor and Li, 1999). More sophisticated techniques consider area averages or modelling of means and covariances (Bürger, 1996) and Artificial Neural Network (ANN) based approaches (Cavazos, 1997; Wilby et al., 1998; Cavazos, 1999). Another category includes approaches relating a weather pattern or classification scheme to station or region-average meteorological data using a correlation approach (Conway and Jones, 1998; Cavazos, 1999).

2.2. Stochastic approaches

These approaches involve selection of a classification scheme for large-scale atmospheric variables (which are termed "weather patterns"), and defining the relationship(s) between local-scale variables and weather patterns in a stochastic manner (Bardossy and Plate, 1992; Hughes and Guttorp, 1994; Wilby et al., 1994; Lattenmaier, 1995; Wilks, 1999; Charles et al., 1999). Hughes and Guttorp (1994) described a non-homogeneous hidden Markov model (NHMM) to relate local-precipitation occurrences to large-scale variables by introducing the hidden weather state as a link between the two scales. The model was applied at 30 raingauge stations in southwestern Australia (Hughes, 1999). Wilks (1999), Zortia and von Storch (1999) and Semenov et al. (1998) compare these approaches.

Nearest neighbor resampling considering Markovian dependence has been used to generate local scale variables in a physically consistent manner (Lall and Sharma, 1996; Harrold at el., 2002). Rajagopalan and Lall (1999) compared nearest-neighbor resampling with a parametric time series model due to Richardson (1981). Buishand and Brandsma (2001) used nearest neighbor resampling for multisite generation of daily precipitation and temperature conditioned on atmospheric variables, at 25 stations in the German part of the Rhine basin.

3. ASSUMPTIONS, LIMITATIONS AND CHALLENGES IN EMPIRICAL DOWNSCALING

Empirical downscaling focuses on statistical descriptions of the relationships between predictor and local-scale variables, paying little attention to physical linkages between the atmosphere and the surface environment. Some major assumptions include stability of the statistical relationships over time, integrity of the GCM output, and application of these downscaling techniques calibrated for the present climate, to future climate (Wilby, 1997; Wilby and Wigley, 1997; Yarnal et al., 2001).

The relationship between the atmosphere and surface variables can be unstable, since short-term relationships are conditional on long-term variations in the climate system. The surface variable under consideration may also be dependent on additional atmospheric variables. It is also assumed that GCMs adequately represent the large-scale features of the atmosphere. However, it has been observed that errors in empirical downscaling can be a result of improper simulations of long wave amplitudes, pressure fields and geopotential heights (Yarnal et al., 2001).

4. FUTURE RESEARCH SCOPE

Empirical downscaling has emerged as a potential tool to relate atmospheric circulation patterns to surface variables. With the increase in computing power, the possibility of greater use and refinement of empirical downscaling methods and introduction of new techniques has also increased. One possible research area is better understanding of the physical processes that govern the relationship between the atmosphere and surface variables and incorporation of this knowledge in the downscaling methodology. Other possible areas may include: the temporal and spatial scales at which these relationships remain stable, comparison of differences in model performance among various empirical downscaling approaches, development of better predictor selection criteria and, more flexible descriptions of the relationships between the atmosphere and the surface environment, in both space and time.

5. APPLICATION OF A NON-PARAMETRIC MODEL

Results of a preliminary application of a non-parametric downscaling model are presented here. In order to test the utility of the identified predictors in downscaling the rainfall occurrence state, the different model configurations (predictor variables) are evaluated by downscaling daily rainfall occurrence at multiple locations in a leave-one-out cross validation mode. This approach allows the identified predictors to be tested on data points not used in model development, hence allowing one to assess the performance of these predictors for downscaling. The model used in the study is based on the nearest neighbour approach (Lall and

Sharma, 1996; Rajagopalan and Lall, 1999; Buishand and Brandsma, 2001, Harrold at el., 2002) and is designed to provide a good representation of both the rainfall occurrence process and the seasonal variation of rainfall within a year. The temporal variability is achieved through the use of predictor conditioning variables while the seasonality is represented using a moving window approach. As re-sampling is done jointly at all stations, spatial correlations among the stations are automatically preserved. The model is non-parametric, i.e., assumptions about the probability distribution and the nature of dependence between variables are minimised.

The model is applied using 26 years of historical daily rainfall and temperature data from ten stations located around Sydney, Australia. None of these stations contain more than 1 percent of missing records. These missing values are filled in using the available records of adjacent stations. Figure 1 shows the location of these stations while Table 1 provides the latitude, longitude, elevation above MSL, average annual rainfall and average maximum and minimum temperature details at these stations.

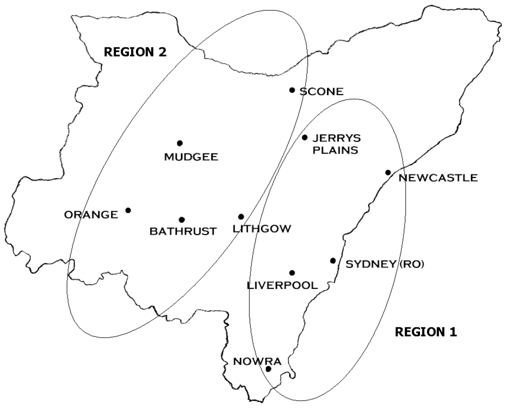


Figure 1: Location of stations used in the study.

The set of predictors used in our model includes combinations of area average rainfall occurrence and area average temperature range for the region. The utility of the selected predictors is evaluated by downscaling the daily rainfall occurrence in a leave-one-out cross validation mode at all stations and comparing the results with the null case i.e. downscaling the rainfall occurrence without any predictor.

5.1. The model

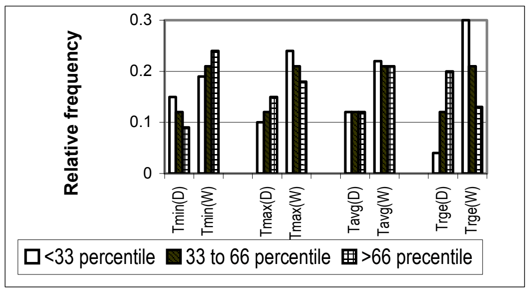
Formation of a stochastic downscaling model for a time series $\{x_1, x_2,..., x_t,...\}$ involves specification of the conditional probability distribution $f(x_t|z_t)$, where z_t is a vector of predictors formed from the large-scale predictor variables, which are indicative of the present state of x_t. The prediction of x_t can be made using the conditional probability distribution $f(x_t|z_t)$, which is defined as:

$$\mathrm{f}(\mathrm{x_t}|\mathrm{z_t}) = \frac{f(x_t, z_t)}{\int f(x_t, z_t)dx_t} \qquad (1)$$

Resampling models using nearest neighbour methods have been frequently used for estimation of $f(x_t| z_t)$. Lall and Sharma (1996) proposed that the conditional distribution in (1) can be represented as a function of the k-nearest neighbours of the conditioning vector z_t, with more probability $p(i)$ being assigned to the neighbours that lie closer to the conditioning vector, and less probability to the more distant neighbours:

$$p(i) = \frac{1/i}{\sum_{j=1}^{k} 1/j} \qquad (2)$$

Where $p(i)$ is the probability that the i^{th} nearest neighbour will be resampled, and k is the number of neighbours considered.



Tmin - Daily minimum temperature (˚C); Tmax - Daily maximum temperature (˚C)
Tavg - Average of daily minimum and daily maximum temperature (˚C)
Trge - Range of daily minimum and daily maximum Temperature (˚C)
D - Dry state; W - Wet state

Figure 2: Frequency of occurrence of rainfall states in different temperature classes.

In our application, x_t is a vector of daily rainfall occurrence at multiple stations and z_t is a vector formed from the combinations of the area-average temperature and rainfall occurrence on the current and the previous days. We analysed the sensitivity of the area-average values of minimum temperature, maximum temperature, average temperature and temperature range of the previous and the current days' to the current day's rainfall occurrence, and found the temperature range of the current day to be a good predictor of the rainfall state (Figure 2 and Table 2). We also considered the division of the study area into two homogeneous regions based on the elevation of the stations (Figure 1). The number of combinations (models) considered in the predictor set formulation is presented in Table 3.

The moving window approach provides an attractive alternative for modelling seasonality (Rajagopalan et al., 1996; Sharma and Lall, 1999, Harrold et al., 2002). A window of specified length is centred at the current calendar day, and all days falling within the moving window (from all historical years) form the local subset of data used in the model for the current day. These windows naturally represent the seasonal variability present in the historical record.

After analysing the sensitivity of the model to different choices of width of moving window l, a value of l = 15 days was chosen for use in our application. We obtained k, the number of nearest neighbors, by trialling a range of possible values, and finally selecting the value of k as ten for all models. The null case in our model corresponds to z_t being a null (empty) vector.

Table 1: Name, location, elevation and other details of the stations used in the study.

Station name	Code	Latitude (S)	Longitude (E)	Elevation (m)	Average Annual Rainfall (mm)	Average annual daily maximum temperature (˚C)	Average annual daily minimum temperature (˚C)
Jerrys plains (Jerrys Plains P.O.)	61086	-32.4983	150.9083	90	640.2	25.2	10.5
Liverpool (Whitlan Centre)	67035	-33.9272	150.9128	20	869.5	23.2	11.6
Mudgee Post Office	62021	-32.5955	149.5956	454	675.7	23.0	8.3
Newcastle (Nobbys Signal Station)	61055	-32.92	151.7978	33	1141.9	21.8	14.2
Nowra Ran Air Station	68076	-34.9506	150.5358	109	1134.7	21.3	11.3
Orange (Orange Airport)	63231	-33.3828	149.1217	948	922.2	17.4	6.1
Scone (Scone ScS)	61089	-32.0632	150.9272	216	654.8	23.9	11.0
Sydney Regional Office	66062	-33.8522	151.2039	42	1221.9	21.6	13.7
Bathurst Agri. Research Station	63005	-33.4289	149.5559	713	634.0	19.7	6.8
Lithgow Composite	63224	-33.4922	150.1483	950	869.5	18.2	6.3

Table 2. Overall *MSE* and *LR* values for different cases of the area-average values of temperature.

Sr. No.	Description	Number of variables in Z_t vector	*MSE*	*LR*
1	Maximum temperature of the previous day	1	0.214	1.049
2	Minimum temperature of the previous day	1	0.204	1.095
3	Maximum and minimum temperatures of the previous day	2	0.185	1.172
4	Previous day temperature range	1	0.190	1.160
5	Previous day average of maximum and minimum temperature	1	0.221	1.02
6	Maximum temperature of the current day	1	0.198	1.119
7	Minimum temperature of the current day	1	0.201	1.114
8	Maximum and minimum temperatures of the current day	2	0.164	1.268
9	Current day temperature range	1	0.164	1.272
10	Current day average of maximum and minimum temperature	1	0.218	1.031
11	Temperature range of the current and the previous days	2	0.157	1.300

Table 3. Overall *MSE* and *LR* values for different models.

Model No.	Description	Number of variables in Z_t vector	*MSE*	*LR*
1	Null case	0	0.193	1.000
2	Rainfall state of the previous day	1	0.177	1.112
3	Rainfall state of the current day	1	0.085	1.590
4	Rainfall state of the current and the previous days	2	0.091	1.591
5	Rainfall state of the current day and temperature range of the previous day	2	0.093	1.592
6	Rainfall state and temperature range of the current day	2	0.093	1.594
7	Temperature range of the current day and rainfall state of the previous day	2	0.158	1.294
8	Temperature range of the current day of region 1 and region 2	2	0.155	1.316
9	Rainfall state of the current day of region 1 and region 2	2	0.059	1.731
10	Rainfall state and the temperature range of the current day of region 1 and region 2	4	0.064	1.713
11	Rainfall state of the current day and temperature range of the current and previous days	3	0.154	1.297

Downscaling of rainfall occurrence proceeds by defining the vector z_i for every day i in the seasonal subset of the historical record formed by the moving window centred at the Julian day corresponding to t, which is the current day in the predicted sequence. The k nearest neighbours of z_t have the k minimum Euclidean distances between

z_i and z_t. These Euclidean distances E_i are calculated as:

$$E_i = \sqrt{\sum_{j=1}^{m} (z_{j,i} - z_{j,t})^2} \qquad (3)$$

where m is the number of variables included in the z vector (Table 2).

The downscaled rainfall occurrence $p(x_t \mid z_t)$ is given by:

$$p(x_t | z_t) = \sum_{i=1}^{k} x_i \times p(i) \qquad (4)$$

Since the occurrence data is binary, with 0 representing a dry day and 1 representing a wet day, $p(x_t \mid z_t)$ is the probability of occurrence of a wet day.

The performance of a particular model is evaluated by comparing the model predicted rainfall occurrence of each day to the historical rainfall state, using the following two criteria:

(i) Mean square error (*MSE*): This is a measure of deviation of predicted rainfall state from the observed state and is calculated as:

$$MSE_l = \frac{1}{N} \sum_{i=1}^{N} (x_{l,i} - p(x_{l,i} | z_i))^2 \qquad (5)$$

where N is the number of days of records considered, and MSE_l is average *MSE* at a station (l). $x_{l,i}$ is observed and $p(x_{l,i}/z_i)$ is the expected or average predicted rainfall state on the i^{th} day at the l^{th} station. The average of MSE_l for all stations is reported as the overall *MSE* for a particular model.

(ii) Likelihood ratio (*LR*): This ratio is given by:

$$LR_l = \frac{1}{N} \sum_{i=1}^{N} \frac{p(x_{l,i} | z_i)}{p(x_{l,i})} \qquad (6)$$

where LR_l is the likelihood ratio for station l, and $p(x_{l,i})$ is the marginal probability of the observed rainfall occurrence for day i at station l, computed using the moving window. For null case, this ratio is equal to one while for other cases we expect it to be greater than one. The average of LR_l for all stations is reported as the overall *LR* for a particular model.

Table 3 provides the overall *MSE* and *LR* values for different models. These results suggest that the inclusion of the area-average rainfall state of the current day as a large-scale predictor provides good results. Also, the division of the study area on the basis of elevation of the stations further improves the downscaled results. It should be noted that the above results are ascertained using a leave-one-out cross-validation formulation; hence one should expect similar results when the model is applied on new data.

6. CONCLUSIONS

This paper is a part of an ongoing study to compare and develop statistical downscaling methods for downscaling of rainfall at hydrologically relevant spatial scales. A review of recent literature highlights the need for development of more general and flexible circulation classification schemes and investigating the differences in model performances among various empirical downscaling approaches. There is little systematic work that has explicitly evaluated the relevant skill of different atmospheric predictors.

Results of a preliminary application of a non-parametric downscaling model for evaluating a configuration of predictor variables by downscaling the daily rainfall occurrence (in a leave-one-out cross validation mode) at multiple stations indicate the temperature range to be a good predictor in comparison to the average temperature and the maximum and minimum temperature of a day. The area-average rainfall state of the current day considered as a large-scale predictor provides good results and the division of the study area into the sub-areas on the basis of elevation further improves the results.

We are also working on NHMM based downscaling approach, results of which would form a basis of comparison of non-parametric and parametric techniques, and would help in developing better relationships between atmosphere and the surface environment.

7. ACKNOWLEDGEMENTS

We thank the constructive comments of an anonymous reviewer that helped improve the quality of the final paper.

8. REFERENCES

Bardossy, A., and E.J. Plate, Space-time model for daily rainfall using atmospheric circulation patterns, *Water Resources Research*, *28*, 1247-1259, 1992.

Buishand, T.A., and T. Brandsma, Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resources Research*, *37* (11), 2761-2776, 2001.

Bürger, G., Expanded downscaling for generating local weather scenarios, *Climate Research*, *7*, 111-128, 1996.

Cavazos, T., Downscaling large-scale circulation to local rainfall in North-Eastern Mexico, *International Journal of Climatology*, *17*, 1069-1082, 1997.

Cavazos, T., Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas, *Journal of Climate*, *12*, 1506-1523, 1999.

Charles, S.P., B.C. Bates, P.H. Whetton, and J.P. Hughes, Validation of downscaling models for changed climate conditions: case study of southwestern Australia, *Climate Research*, *12* (1), 1-14, 1999.

Conway, D., and P.D. Jones, The use of weather types and air flow indices for GCM downscaling, *Journal of Hydrology*, *212/213*, 348-361, 1998.

Harrold, T.I. Stochastic generation of daily rainfall for catchment water management studies. PhD Thesis, School of Civil Engineering, University of New South Wales, 2002. http://adt.caul.edu.au.

Harrold, T. I., A. Sharma and S. J. Sheather, Representation of long-term variability in daily rainfall generation. Hydrology and Water Resources Symposium, Institution of Engineers, Australia, 2002.

Hewiston, B.C., and R.G. Crane, Climate downscaling: techniques and application, *Climate Research*, *7*, 85-95, 1996.

Hughes, J.P., and P. Guttorp, Incorporating Spatial Dependence and Atmospheric Data in a Model of Precipitation, *Journal of Applied Meteorology*, *33* (12), 1503-1515, 1994.

Hughes, J.P., Guttorp, P. and Charles, S.P., A non-homogeneous hidden Markov model for precipitation occurrence, *Journal of Applied Statistics*, *48*, 15-30, 1999.

Lall, U., and A. Sharma, A nearest neighbor bootstrap for time series resampling, *Water Resources Research*, *32* (3), 679-693, 1996.

Lattenmaier, D., Stochastic modeling of precipitation with applications to climate model downscaling, in *Analysis of Climate variability: Applications of statistical techniques*, edited by H.v. Storch, and A. Navarra, Springer Verlag, 1995.

Prudhomme, C., N. Reynard, and S. Crooks, Downscaling of global climate models for flood frequency analysis: where are we now? Hydrological Processes, 16, 1137-1150, 2002.

Rajagopalan, B., and U. Lall, A k-nearest neighbour simulator for daily precipitation and other weather variables, *Water Resources Research*, *35*, 1999.

Richardson, C., Stochastic simulation of daily precipitation, temperature and solar radiation, *Water Resources Research, 17*, 182-190, 1981.

Sailor, D., and X. Li, A semiempirical downscaling approach for predicting regional temperature impacts with climate change, *Journal of Climate*, *12*, 103-114, 1999.

Semenov, M., R.J. Brooks, E.M. Barrow, and C. Richardson, Comparison of the WGEN and LARS-WG stochastic weather generators in diverse climates, *Climate Research*, *10*, 95-107, 1998.

von Storch, H., Zorita, E. and Cubasch, U., Downscaling of global climate change estimates to regional scale: an application to Iberian rainfall in winter time, *Journal of Climate*, *6*, 1161-71, 1993.

Wilby, R., Stochastic weather type simulation for regional climate change impact assessment, *Water Resources Research*, *30*, 3395-3403, 1994.

Wilby, R., Non-stationarity in daily precipitation series: implications for GCM downscaling using atmospheric circulation indices, *International Journal of Climatology, 17*, 439-454, 1997.

Wilby, R., B. Greenfield, and C. Glenny, A coupled synoptic-hydrological model for climate change impact assessment, *Journal of Hydrology*, *153*, 265-290, 1994.

Wilby, R., and T.M.L. Wigley, Downscaling general circulation model output: a review of methods and limitations, *Progress in Physical Geography*, *21* (4), 530-548, 1997.

Wilby, R., T.M.L. Wigley, D. Conway, P.D. Jones, B.C. Hewiston, J. Main, and D.S. Wilks, Statistical downscaling of general circulation model output: A comparison of methods, *Water Resources Research*, *34*, 2995-3008, 1998.

Wilks, D.S., Multisite downscaling of daily precipitation with a stochastic weather generator, *Climate Research*, *11*, 125-136, 1999.

Yarnal, B., A.C. Comrie, B. Frakes, and D.P. Brown, Developments and prospects in Synoptic Climatology, *Int. Journal of Climatology*, *21*, 1923-1950, 2001.

Zortia, E., and H. von Storch, The analog method as a simple statistical downscaling technique: Comparison with more complicated methods, *Journal of Climate*, *12*, 2474-2489, 1999.