Medium Term Forecasting of Rainfall using Artificial Neural Networks

[1]Iseri, Y. [2]G. C. Dandy, [2]H. R. Maier, [3]A. Kawamura and [1]K. Jinno

[1]Institute of Environmental Systems, Kyushu University, [2] Centre for Applied Modelling in Water Engineering (CAMWE), Department of Civil and Environmental Engineering, University of Adelaide, [3]Department of Civil Engineering, Tokyo Metropolitan University,
E-Mail: gdandy@civeng.adelaide.edu.au

Keywords: *forecasting of rainfall; artificial neural networks; partial mutual information; climate indices; sea surface temperatures*

EXTENDED ABSTRACT

The state of the atmosphere and ocean can be characterized by climate indices. One of the well known indices is the Southern Oscillation Index (SOI). SOI measures the sea level pressure difference between Tahiti and Darwin, indicating the occurrence of the El Niño phenomenon in the Central Pacific region. The Pacific Decadal Oscillation Index (PDOI) represents decadal scale atmosphere-ocean oscillation in the Pacific Ocean while the North Pacific Index (NPI) measures the intensity of the Aleutian low pressure cell (Kawamura *et al*. 2003).

A number of researchers have studied the possibility of forecasting rainfall several months in advance using climate indices such as SOI, PDOI and NPI (e.g. Silverman and Dracup, 2000). Furthermore, the existence of substantial databases of sea surface temperature anomalies (SST) opens the possibility of using these data to forecast rainfall several months in advance. Most of the research carried out in this area has used traditional statistical methods such as linear correlation or time series methods to identify the significant variables. These methods test for a linear relationship between the independent variables and rainfall, whereas the relationships are more likely to be non-linear as the underlying processes are themselves non-linear.

This paper describes the use of partial mutual information (PMI) to identify the significant inputs for medium term rainfall forecasting in Japan. In particular, a study is made of monthly rainfall in the City of Fukuoka. Fukuoka, which is located in the northern part of Kyushu Island, is vulnerable to drought. In fact, the city was affected by devastating droughts in 1978 and 1996 (Kawamura and Jinno, 1996). Therefore a more successful rainfall prediction model would be of great benefit to the city.

The possible inputs considered include the SOI, NPI and PDOI as well as SST in selected locations from a 5°x 5° grid in the Pacific Ocean. The selected inputs are used to develop artificial neural network models (ANNs) to forecast rainfall in Fukuoka several months in advance.

Six distinctive scenarios are considered in this study. Three of the scenarios use input data with lags between 1 month and 12 months and the other three scenarios use data with lags between 3 months and 12 months in order to investigate the possibility of forecasting more than 3 months in advance. The three scenarios considered for the two different ranges of lags are as follows:

(1) use only SST as candidate predictors
(2) use only climate indices as candidate predictors
(3) use both SST and climate indices as candidate predictors

One of the objectives of this study is the identification of a possible relationship between rainfall in Fukuoka and hydro-climatic variables such as SST and climate indices, using partial mutual information. The other objective is to verify the forecasts produced using the predictors identified with partial mutual information and investigate whether the inclusion of SST in addition to climate indices improves the prediction accuracy.

It is found that the North Pacific Index (NPI) lagged by 6 months has a strong relationship with August rainfall in Fukuoka. Some improvement in forecasts can be achieved by including sea surface temperature anomalies as additional inputs.

1. INTRODUCTION

Climatic variability and its effect on human activity have been discussed many times in the literature. One of the most crucial issues of global climatic variability is its effect on water resources. If more accurate predictions of rainfall were possible, this would enable more efficient utilization of water resources. However, long-term rainfall prediction models are still unsatisfactory, whereas short-term rainfall prediction models have undergone significant development. The probable reasons for the difficulties in conducting long-term rainfall prediction are the complexity of atmosphere-ocean interactions and the uncertainty of the relationship between rainfall and hydro-meteorological variables.

So far, long-term climate prediction using numerical models has not demonstrated useful performance, and statistical models have shown better performance than numerical models (Zwiers and Von Storch, 2004). Consequently, in this study Artificial Neural Networks and linear regression models have been applied to nonlinear and linear statistical rainfall prediction. Moreover, Partial Mutual Information (PMI) is used to identify nonlinear relationships between rainfall and hydro-climatic variables. The PMI method was first applied to water resources variables by Sharma (2000) and Sharma *et al.* (2000) in order to detect nonlinear relationships between them. In this study, the hydro-climatic variables considered are sea surface temperatures (SST) and climatic variability indices such as Southern Oscillation index (SOI), Pacific Decadal Oscillation Index (PDOI) and North Pacific Index (NPI).

Monthly rainfall data for Fukuoka, which is located in the northern part of Kyushu Island, is predicted in this study. Fukuoka is vulnerable to drought having been affected by devastating droughts in 1978 and 1996 (Kawamura and Jinno, 1996). Therefore, a better rainfall prediction model would be beneficial for the city.

This paper consists of two sections. Firstly, the partial mutual information between August Rainfall in Fukuoka and hydro-climatic variables were computed in order to identify the predictors. Secondly, forecasting of August rainfall using the identified inputs was conducted using Artificial Neural Networks.

2. DATA

The Southern Oscillation Index (SOI), Pacific Decadal Oscillation Index (PDOI) and North Pacific Index (NPI) were used to investigate the relationship between global scale climatic variability and the precipitation in Fukuoka. Similarly, data from a 5°×5° grid of Sea Surface Temperature anomalies in the Pacific ocean were used in order to detect the possible effect of regional SST on precipitation. All of the data used in this study are monthly values.

2.1 Precipitation in Fukuoka

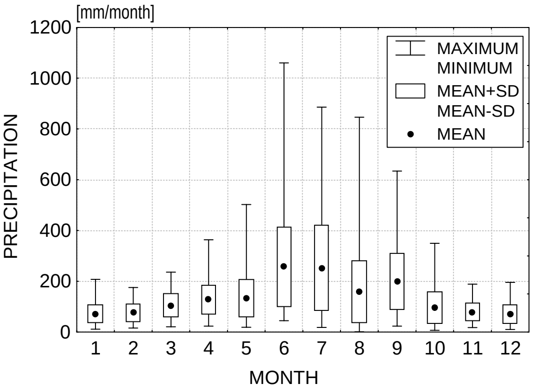

Figure 1. Mean, standard deviation, maximum, and minimum monthly precipitation in Fukuoka, Japan (1=January, 12=December).

Precipitation in Fukuoka has been recorded since January 1890. Figure 1 shows the average, standard deviation, maximum, and minimum monthly precipitation (January – December) for this period. From Figure 1, it can seen that June, July, August and September have high average precipitation. Precipitation from June to September is therefore of critical importance in order to maintain a reliable water supply. Preliminary analyses indicated that August rainfall has comparatively high correlation with the three climatic indices (SOI, PDOI, NPI), therefore August rainfall has been selected as the predicted variable in this study.

Monthly precipitation in Fukuoka is not normally distributed but is positively skewed. Consequently, a cubic root transformation was carried out in order to normalize the data. The normalized monthly precipitation data were standardized to a mean of zero and a standard deviation of one, by subtracting the normalized monthly mean and dividing by the normalized standard deviation for the base period 1901 – 2002. These normalized and standardized precipitation data are used in this study.

2.2 SOI

A well-known atmospheric phenomenon is the Southern Oscillation (SO). The SO is an atmospheric see-saw process in the tropical Pacific sea level pressure between the eastern and western hemispheres associated with the El Niño and La

Niña oceanographic features. The oscillation can be characterized by a simple index, the Southern Oscillation Index (SOI). (Kawamura *et al*., 1998). This index was used by NOAA (The National Oceanic and Atmospheric Administration) to evaluate when El Niño and La Niña are occurring (Japanese Study Group for Climate Impact & Application, 1999). The feature is known as the El Niño Southern Oscillation (ENSO) phenomenon.

The SOI was derived from monthly mean sea level pressure differences between Papeete, Tahiti (149.6°W, 17.5°S) and Darwin, Australia (130.9°E, 12.4°S). The database for the calculation of the SOI in the present study consists of 137 years of monthly mean sea level pressure data at Tahiti and Darwin from January 1866 to December 2002. The data were obtained from Ropelewski and Jones (1987) and Allan *et al*. (1991), who carefully infilled all missing values by correlation with data from other observation stations. The data from before 1920 are somewhat less reliable than the later values (Kawamura *et al*., 1998). For the details of statistical and long-term characteristics of SO, SOI and their barometric pressure data refer to Kawamura *et al*. (2002) and Jin *et al*. (2003).

2.2. PDOI

The Pacific Decadal Oscillation (PDO) is described as a long-lived pattern of Pacific climatic variability somewhat like El Niño. PDO has two phases (the warm and cool phases), and each phase persisted for 20 to 30 years in the 20th century. The fingerprints of PDO are most visible in the North Pacific/North American region. Several studies found evidence for just two full PDO cycles in the past century: cool phases occurred during the periods 1890-1924 and 1947-1976, while warm phases prevailed during the periods of 1925-1946 and 1977 through the mid-1990s (Mantua *et al*., 1997).

PDOI is the leading principal component of monthly sea surface temperature (SST) anomalies in the North Pacific Ocean north of 20°N (Zhang *et al*., 1997; Mantua *et al*., 1997). The PDOI data since 1900, which are used in this study, were obtained from the website of the Joint Institute for the Study of the Atmosphere and Ocean [http://tao.atmos. washington.edu/main.html].

2.3. NPI

Trenberth and Hurrell (1994) have defined the North Pacific Index (NPI) as the area-weighted sea level pressure over the region 30°N to 65°N, 160°E to 140°W to measure the decadal variations of atmosphere and ocean in the north Pacific. They found that this index is highly correlated with the leading principal component of the 500 hPa geopotential height. NPI is also a good index for the intensity of the Aleutian Low pressure cell. NPI data since 1899 were obtained from the website of the University Corporation for Atmospheric Research [http://www.ucar. edu/ucar/index.html].

2.4 Sea Surface Temperature Anomalies

In this study, Kaplan sea surface temperature anomalies were used. These are global sea surface temperature anomalies using monthly data on a 5° ×5° grid (Kaplan *et al*. 1998; Parker *et al*. 1994; Reynolds *et al*. 1994). The data were provided on the website of the International Research Institute for climate prediction [http://iri.columbia.edu/]. The available sea surface temperature anomalies in the Pacific Ocean (42.5S-32.5S, 117.5E-242.5E, 27.5S-7.5N, 117.5E-287.5E and 12.5N-62.5N, 117.5E-242.5E) for the period of January 1856 to December 2002 were used for computation in this study.

3. METHODS

The procedures crucial for developing the prediction model are the identification of predictors and the determination of which prediction model to employ. As the first step of this study, PMI scores between candidate inputs and the desired output (i.e. the August rainfall which is transformed and standardized as described above) were computed for six different scenarios in order to detect suitable inputs for forecasting. After the input identification process, the selected inputs were utilised for forecasting using Artificial Neural Networks models. It is expected that the non-linear relationships captured by the PMI algorithm will best be represented in the predictions using ANNs.

3.1 Partial Mutual Information

Determination of the inputs for forecasting is one of the most important steps in the model development process. Cross-correlation is widely used for selecting appropriate predictors, however it is only able to detect linear relationships between predictors and outputs. Hence, non-linear relationships between potential inputs and the output might not be detected. Therefore, in identifying suitable inputs for the prediction, the stepwise Partial Mutual Information (PMI) algorithm was used in this study. This algorithm was proposed by Sharma (2000) as a method to capture both linear and non-linear relationships between model inputs and output and modified by Bowden *et al*. (2005).

The PMI algorithm applied in this study is as follows:

1. Identify the set of variables that are likely to be useful predictors of the system being modelled. Denote this variable set as the vector $\mathbf{z}_{in}$. Denote the vector that will store the final predictors of the system as $\mathbf{z}$. This is a null vector at the start of the algorithm.
2. Estimate the PMI between the dependent variable y and each of the plausible new predictors in $\mathbf{z}_{in}$, conditional on the pre-existing predictor set $\mathbf{z}$.
3. Identify the variable in $\mathbf{z}_{in}$ having the highest PMI score in step 2.
4. Use the bootstrapping method to estimate the 99th percentile sample PMI score for the variable identified in step 3.
5. If the PMI score for the identified variable is higher than 99th percentile randomised sample PMI score of step 4, include the variable in the predictor set $\mathbf{z}$, and remove it from $\mathbf{z}_{in}$. If the dependence is not significant, go to step 7.
6. Repeat steps 2-5 as many times as are needed.
7. This step will be reached only when all the significant predictors have been identified.

PMI scores between August rainfall and the following 6 sets of inputs were computed:

(a) SSTa for lag 1 to 12 months

(b) Four climate indices (SOI, PDOI, NPI) for lags 1 to 12 months

(c) The data which showed significant PMI score in the PMI computation for (a) and (b),

(d) SSTa for lags 3 to 12 months

(e) Four climate indices (SOI, PDOI, NPI) for lags 3 to 12 months

(f) The data which showed significant PMI score in the PMI computation for (d) and (e)

After the computation of PMI scores, Artificial Neural Networks models were developed for each of the above cases.

3.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are used as prediction models in this study. Although several dynamic models have been developed for prediction of meteorological variables, statistical models such as ANNs have played a significant role. Since ANNs have the ability to represent non-linear relationships between inputs and output, it is expected that the non-linear relationships captured by the PMI algorithm will be well represented using ANNs.

4. RESULTS AND DISCUSSION

4.1 Input Identification by PMI Scores

The results of the PMI computations for the input sets given in (a) to (f) above are summarised in Table 1. It can be seen that when SSTa in the Pacific Ocean is included as an input for PMI calculation, January (i.e. lag 7) SSTa at the grid location of 27.5°N 132.5°W has the highest PMI score among all inputs. However, when SSTa are used exclusively as candidate inputs, which corresponds to cases (a) and (d), the PMI score for the second predictor and its 99 percentile value were nearly the same. The results in Table 1 suggest NPI in February is one of the best predictors for August rainfall.

4.2 Development of ANN Models

The identified inputs shown in Table 1 were used to develop a prediction model using artificial neural networks with 15 inputs for model (a), 1 input for model (b), 7 inputs for model (c), 10 inputs for model (d), 1 input for model (e) (This model is the same as the model (b)) and 2 inputs for model (f). August rainfall in Fukuoka city is the dependent variable for all models.

The data from 1901 to 1997 (97 years) were used for testing, training and validation. The SOM data division method (Bowden et al., 2002) was used to divide the data for model (c) into training, testing and validation sets of sizes 64, 22 and 11 (respectively). This model contains the most detailed information on the atmosphere and ocean and is expected to have the best performance of the 5 models (models (b) and (e) are the same). The data for the other 4 models were also divided in the same way, namely, 64 observations for training, 22 data for testing and 11 for validation.

Table 1. PMI scores and the locations of identified inputs. (When the total number of identified inputs is greater than six, the six variables with the highest PMI are shown)

Variable	Lead time (months)	Location	PMI	99th percentile PMI
(a) SSTa in the Pacific ocean for lead times 1 to 12 months (total of 6816 possible inputs)				
SSTa	7	27.5°N, 132,5°W	0.18454	0.13741
SSTa	1	17.5°N, 117.5°W	0.14791	0.13566
SSTa	3	7.5°N, 77.5°W	0.16217	0.13406
SSTa	6	12.5°S, 157.5°E	0.17563	0.13380
SSTa	11	42.5°S, 157.5°E	0.14809	0.13236
SSTa	2	22.5°N, 112.5°W	0.16867	0.13236
	Total number of identified inputs		15 inputs	
(b) SOI, PDOI, NPI for lead times 1 to 12 months (total of 36 possible inputs)				
NPI	6		0.17075	0.12412
	Total number of identified inputs		1 input	
(c) the identified inputs in (a) and (b) combined together (total of 16 possible inputs)				
SSTa	7	27.5°N, 132.5°W	0.18454	0.13741
NPI	6		0.19192	0.13002
SSTa	8	22.5°N, 137.5E	0.15564	0.13124
SSTa	6	12.5°S, 157.5E	0.16343	0.13238
SSTa	2	27.5°N, 107.5°W	0.15173	0.13080
SSTa	1	12.5°N, 117.5°W	0.16504	0.13080
	Total number of identified inputs		7 inputs	
(d) SSTa in the Pacific ocean for lead times 3 to 12 months (total of 5860 possible inputs)				
SSTa	7	27.5°N, 132.5°W	0.18454	0.13741
SSTa	7	22.5°N, 157.5°E	0.14113	0.13428
SSTa	7	32.5°N, 127.5°W	0.15747	0.13752
SSTa	9	57.5°N, 142.5°W	0.15954	0.13632
SSTa	11	42.5°S, 157.5°E	0.14863	0.13682
SSTa	6	12.5°S, 157.5E	0.14992	0.13682
	Total number of identified inputs		6 inputs	
(e) SOI, PDOI, NPI for lead times 3 to 12 months (total of 30 possible inputs)				
NPI	6		0.17075	0.12412
	Total number of identified inputs		1 input	
(f) the identified inputs in (d) and (e) combined together (total of 7 possible inputs)				
SSTa	7	27.5°N, 132.5°W	0.18454	0.13741
NPI	6		0.19192	0.13002
	Total number of identified inputs		2 inputs	

A constructive approach is employed in order to determine the structure of ANNs used in this study. The approach begins from an ANN structure with no hidden nodes (Maier and Dandy, 2000), and calculates the root mean square error (RMSE) for the training set. After computation of the RMSE for the structure with no hidden nodes, the number of hidden layers is fixed at one and the number of hidden nodes increased by one at a time while computing the RMSE for each structure. When the reduction in the training RMSE becomes reasonably small, the number of hidden nodes is not increased any further and the structure is assumed to be optimal.

After the determination of the optimal ANN model, cross-validation with the validation set is

Table 2. RMSE and R^2 between observed and predicted rainfall of training, testing and validation set for models (a), (b), (c), (d) and (f)

MODEL	RMSE			R^2		
	training	testing	validation	training	testing	validation
(a)	0.611	0.877	0.784	0.667	0.243	0.147
(b) and (e)	0.920	1.091	0.653	0.307	0.022	0.309
(c)	0.465	0.743	0.633	0.808	0.514	0.366
(d)	0.823	0.794	0.621	0.397	0.375	0.307
(f)	0.520	0.898	0.691	0.759	0.213	0.210

employed for each model in order to assess their generalization ability.

The RMSE and coefficient of determination (denoted as R^2) between the observed and the predicted rainfall data for the training, testing and validation sets for models (a), (b), (c), (d) and (f) are given in Table 2**.** From this table, it can be seen that model (c) showed the best performance of the 5 models, although this was only slightly better than models (b) and (d) for the validation data. The lower RMSE for model (c) compared to models (b) and (d) indicates the value of using both SSTa data and NPI as predictors of August rainfall.

The results for model (f) are not as conclusive as it has a lower RMSE than models (d) and (e) for the training set but a higher value for the validation set. Overall, model (b) that uses a single value of NPI with a lag of 6 months as the input variable gives reasonable results.

This approach needs to be applied to forecasting rainfall in other parts of the world in order to validate its generality.

5. CONCLUSIONS

The medium term forecasting of August rainfall in Fukuoka city was conducted in this study. In order to identify the adequate predictors, the partial mutual information was used for the candidate predictors, which are sea surface temperature anomalies in the Pacific Ocean and three climate indices.

When data with lead times between 1 and 12 months were used to forecast August rainfall, it was found that a model with the North Pacific index and selected SSTa as inputs performed reasonably well.

If lead times of greater than 3 months are required, the North pacific Index for February gave the best results.

6. REFERENCES

Allan, R.J., Nicholls, N., Jones, P.D. and Butterworth, I.J. (1991), further extension of the Tahiti-Darwin SOI, Early ENSO events and Darwin pressure." *Journal of Climate* 4: 743-749.

Bowden, G.J., Maier, H.R. and Dandy, G.C. (2002), Optimal division of data for neural network models in water resources applications, Water Resources Research 38 (2): 2.1-2.11.

Bowden, G.J., Dandy, G.C. and Maier, H.R. (2005), Input determination for neural network models in water resources applications. Part 1-background and methodology, *Journal of Hydrology* 301 (1-4): 75-92.

Japanese Study Group for Climate Impact & Application (1999), *El Niño & Global Environment*, Seizando, Japan (in Japanese).

Jin, Y.H., Kawamura, A., Jinno, K. and Iseri, Y. (2003) On the long-term variability of Southern Oscillation Index, *Proc. of 2003 Korea Water Resources Association*, 151-158.

Kaplan, A., Cane, Y., Kuhsnir, A., Clement, A., Blument, M. and Rajagopalan, B. (1998), Analyses of global sea surface temperature 1856-1991, *Journal of Geophysical Research*, 103: 18,567- 18,589.

Kawamura, A., McKerchar, A.I., Spigel, R. H. and Jinno, K. (1998), Chaotic characteristics of the Southern Oscillation Index time series, *Journal of Hydrology*. 204: 168-181.

Kawamura, A., Eguchi, S., Jinno, K. and McKerchar, A. (2002), Statistical characteristics of Southern Oscillation Index and its barometric pressure data. *Journal of Hydroscience and Hydraulic Engineering*, Vol. 20, No. 2: 41-49.

Kawamura, A., Iseri, Y., Jin Y.H., and Jinno, K. (2003), Relationship between atmospheric-oceanic indices and precipitation in Fukuoka, Japan, *Proc. of Int'l Conf. on Managing Water Resources under Climate Extremes and Natural Disasters*, 21-30.

Kawamura A. and Jinno, K. (1996). "Integrated water resources management in Fukuoka Metropolitan Area." *Environmental Research Forum* 3&4: 97-109.

Maier, H.R. and Dandy, G.C. (2000), Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications, *Environmental modeling & software* 15: 101-124

Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J.M. and Francis, R.C. (1997), A Pacific interdecadal climate oscillation with impacts on salmon production, *Bulletin of the American Meteorological Society* 78: 1069-1079.

Parker, D.E., Jones, P.D., Folland, C.K. and Bevan, A., (1994), Interdecadal changes of surface temperature since the late nineteenth contury, *Journal of Geophysical Research* 99: 14,373-14,399

Reynolds, R.W. and Smith, T.M. (1994), Improved global sea surface temperature analysis using optimum interpolation, *Journal of Climate*, 7: 929-948

Ropelewski, C.F. and Jones, P.D. (1987), An extension of the Tahiti-Darwin Southern Oscillation index, *Monthly Weather Review*, Vol, 115: 2161-2165.

Sharma, A. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1- A strategy for system predictor identification., *Journal of Hydrology* 239: 232-239.

Sharma, A., Luk, K.C., Cordery, I., Lall, U. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 2 – Predictor identification of quarterly rainfall using ocean-atmosphere information, *Journal of Hydrology* 239: 240-248

Silverman, D. and Dracup, J.A. (2000), Artificial Neural networks and long-lead precipitation prediction in California, *Journal of applied meteorology* 39: 57- 66

Trenberth, K.E., Hurrel, J.W. (1994), Decadal atmosphere-ocean variations in the Pacific. *Climate Dynamics* 9: 303-319

Zhang, Y., Wallace, J. M., Battisi, D. S. (1997). ENSO-like interdecadal variability 1900-1993, *Journal of Climate* 10: 1004-1020.

Zwiers F.W. and Von Storch, J. (2004), On the role of statistics in climate research, *Int. J. Climatology* 24: 665-680

.