MATEMATIKA, 2013, Volume 29, Number 1b, 77-87

Verification of Forecast Rainfall Anomalies

[1]Kho Pui Kim, [2]Fadhilah Yusof and [3]Zalina Mohd Daud
[1, 2]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
[3]RAZAK School of Engineering and Advanced Technology
UTM International Campus, Kuala Lumpur
e-mail: [1]cherylkho87@hotmail.com, [2]fadhilahy@utm.my

Abstract Statistical downscaling is used to relate the large scale climate information with the local variables that is to find the relationship between the National Center of Environmental Prediction (NCEP) data with the ground data. This study examines the verification of forecast rainfall anomalies during November-December-January-February (NDJF). The ground data used is the 30 years NDJF rainfall for 40 stations while the NCEP data is the 20 grids point Sea Level Pressure (SLP). In this paper, Canonical correlation analysis (CCA) is used to find the maximum correlated pattern between two variables. CCA model is verified using the mean square error skill score and anomaly correlation coefficient and used to simulate the current rainfall using the General Circulation Model (GCM) data as predictors. This is so called the validation method. Due to appearance of some biases, the anomaly correlation coefficient is considerably higher than the skill score. These biases may relate to the penalty associated with retaining the Sea Level Pressure (SLP) in the meteorological features when such features are not predictable.

Keywords Canonical Correlation Analysis (CCA); Mean Square Error Skill Score.

2010 Mathematics Subject Classification 62H06, 62H07

1 Introduction

Statistical downscaling is widely proposed as it can sufficiently be used to predict surface conditions from large-scale circulation under present-day climate conditions and less expensive. Numerous statistical downscaling methods have been developed include regression-based method, weather typing approaches and stochastic weather generators [1]. Downscaling using Canonical Correlation Analysis (CCA) method is used to find the relationship between the rainfall and the National Center of Environmental Prediction (NCEP) data. A common accuracy measurement for field forecasts, mean square error (MSE), is operated by spatially averaging the individual squared differences between the gridded forecast and observed fields [2].

One way of assessing the quality of forecast is using skill score which is aimed to equalize the effect of the intrinsic case or difficulty of different forecast situations. Verification of forecast gives information about the nature of forecast errors. Skill score is sensitive to biases and errors in variances. The verification of forecast is important to monitor forecast quality, improve forecast quality and compare the quality of different forecast systems. The skill score is in the range of $-\infty < \text{skill score} \leq 1$ with the perfect forecast skill score is equal to 1 [3].

Several studies have been conducted on downscaling using CCA method [4, 5, 6]. In Malaysia, CCA was used to analyze the maximum correlated coupled patterns between predictand and predictor matrices [7]. On the other hand, the skill scores and correlation coefficient as model verification was conducted on numerical weather prediction (NWP) models [8]. Murphy used MSE-based skill scores to assess the accuracy of forecasts [9, 10, 11]. Large improvements over the Southern Hemisphere as variance is large showed that improved skill for the anomaly correlation of geopotential heights at 500 hPa is better [12].

This paper is focussed on model verification using the skill scores of anomaly correlation obtained from CCA. The purposes of the paper are to present the decomposition of the skill scores based on the mean square error and used the decomposition to evaluate penalties relating to the reliability and bias of the forecasts since skill score is sensitive to biases and errors in variances.

2 Data Sets

The ground data used is the 30 years (1975-2004) historical November-December-January-February (NDJF) rainfall for 40 stations (as in Table 1). These data is obtained from Malaysian Meteorological Department and Malaysian Brainage and Irrigation Department. The NCEP data containing 26 variables but in this paper only the Sea Level Pressure (SLP) which accounted for 20 grids from 0°N to 7.5°N and 100°E to 105°E which has a resolution of 2.5°x2.5° is used. In this paper, 20 grids of NCEP are used since these grids cover the Peninsular Malaysia. Peninsular Malaysia is divided into 5 regions, namely center (C), East (E), North-west (NW), South-west (SW) and West (W).

Figure 1 Map of Malaysia with stations

Table 1 List of stations according to regions

Regions	District	Regions	District
C	S. K. Kg. Aur Gading	NW	SIK
C	Kg. Chebong	NW	Kolam Takongan Air Itam
C	Ldg. Sg. Sabaling	NW	Rumah Pam Bumbong Lima
C	Genting Sempah	NW	Ampang Pedu
C	Sg.Lui Halt	SW	Ibu Bekalan Kahang , Kluang
C	Rumah Pam Paya Kangsar	SW	Ldg. Getah Kukup , Pontian
C	Ldg Boh	SW	Pintu Kawalan Tampok Batu Pahat
C	Gua Musang	SW	Pintu Kawalan Separap Batu Pahat
E	Sek. Keb. Kemasek	SW	Senai
E	Sek. Keb. Kg. Jabi	SW	Kluang
E	Kg. Merang ,Setiu	SW	Tangkak
E	Endau	SW	Malacca
E	Rumah Pam Pahang Tua, Pekan	W	Ibu Bekalan Sg. Bernam
E	Kuantan	W	Kg. Sg. Tua
E	Kg. Menerong	W	Hospital Port Dickson
E	Kota Bharu	W	Petaling Jaya
NW	Ldg. Batu Kawan	W	Rumah Kerajaan JPS,Chui Chak
NW	Guar Nangka	W	Sitiawan
NW	Kodiang	W	Ipoh
NW	Pendang	W	Pusat Kesihatan Bt.Kurau

3 Method

3.1 Canonical Correlation Analysis (CCA)

A multivariate statistical analysis technique - Canonical Correlation Analysis (CCA) is used to find the maximum correlated pattern between two variables [13]. In other words, CCA is used to find the relationships between data of pairs of vectors *x* and *y* [2].
Let *x* and *y* are pairs of data vectors,

$$C'^{T} = \left[x'^{T}, y'^{T}\right] \tag{1}$$

The covariance matrix of C', $\left[S_c\right]$

$$\left[S_c\right] = \frac{1}{n-1}\left[C'\right]^{T}\left[C'\right] = \begin{bmatrix} \left[S_{xx}\right] & \left[S_{xy}\right] \\ \left[S_{yx}\right] & \left[S_{yy}\right] \end{bmatrix} \tag{2}$$

with the eigenvalues λ_m, eigenvectors e_m and f_m and can be computed.

$$e_m = \frac{x}{\sqrt{x'x}} \tag{3}$$

$$f_m = \frac{\left[S_{yy}\right]^{-\frac{1}{2}}\left[S_{yx}\right]\left[S_{xx}\right]^{-\frac{1}{2}} e_m}{\left\|\left[S_{yy}\right]^{-\frac{1}{2}}\left[S_{yx}\right]\left[S_{xx}\right]^{-\frac{1}{2}} e_m\right\|} \tag{4}$$

with the eigenvectors e_m and f_m, we can find the canonical vector a_m and b_m.

$$a_m = \left[S_{xx}\right]^{-\frac{1}{2}} e_m \tag{5}$$

and

$$b_m = \left[S_{yy}\right]^{-\frac{1}{2}} f_m \tag{6}$$

Lastly, the canonical variates, v_m and w_m can be formed.

$$v_m = a_m{}^T x' \tag{7}$$

and

$$w_m = b_m{}^T y' \tag{8}$$

with the canonical variates, a simple linear regressions model is constructed. The model is

$$v_m = \hat{\beta}_{0.m} + \hat{\beta}_{1,m} v_m, m = 1, \ldots, M. \tag{9}$$

3.2 Mean Squared Error Skill Score

Mean squared error (MSE) is the average squared difference between the gridded forecast, y_m and observed fields, o_m. MSE will be more sensitive to larger error and outliers due to the squaring of the forecast errors.

$$MSE = \frac{1}{M}\sum_{m=1}^{M}\left(y_m - o_m\right)^2 \tag{10}$$

In this paper, correlation coefficient is proposed as another useful accuracy9 measurement. Correlation is sensitive to outliers, but not sensitive to biases although it does reflect linear association between two variables. Hence, Murphy [7] manipulated

$$MSE = \left(\bar{y} - \bar{o}\right)^2 + s_y^2 + s_o^2 - 2 s_y s_o r_{yo}. \tag{11}$$

where

s_y = standard deviation of forecast fields

s_o = standard deviation of observations

r_{yo} = anomaly correlation coefficient between forecast field and observations

Skill score is the relative accuracy measurement. For the MSE using climatology as the control forecasts, the skill score is

$$SS = 1 - \frac{MSE}{MSE_{c\lim}} \tag{12}$$

where

$$MSE_{c\lim} = s_o^2 + \bar{o}^2 \tag{13}$$

Hence, $MSE_{c\lim}$ involves only the sample variance of the anomalies in the observations field and the square of the mean anomaly in this field. From Equation 12, substitute an expression for the Pearson product-moment correlation between the forecasts and observations, r_{yo} forms [2]

$$SS_{c\lim} = r_{yo}^2 - \left[r_{yo} - \frac{s_y}{s_o} \right]^2 - \left[\frac{\bar{y} - \bar{o}}{s_o} \right]^2. \tag{14}$$

where

$\bar{y}$ = mean of forecast fields
$\bar{o}$ = mean of observations

Equation 14 implies that skill score involves a contribution due to the correlation between the forecasts and observations, and penalties relating to the reliability and bias of the forecasts. The first term in skill score is the square of the anomaly correlation coefficient. In other words, it is the proportion measurement of variability in the observations that is accounted for the forecasts. The second term in skill score is the reliability measurement or conditional bias of the forecasts. The slope of regression model, *b*, is

$$b = \left[\frac{s_o}{s_y} \right] r_{yo}, \tag{15}$$

The second term vanishes only when $b = 1$, meaning that no conditional bias. The third term is the measurements of the unconditional bias in the forecasts. If the unconditional bias is small, the reduction in skill will be modest.

3.3 Significance Testing for Forecast Field

An easy approach to test the significance is the Fisher *Z* transformation,

$$Z = \frac{1}{2} \ln \left[\frac{1+r}{1-r} \right], \tag{16}$$

where *r* is the Pearson correlation.

The hypothesizes are

$$H_0 : r = 0 \text{ vs } H_1 : r \neq 0. \tag{17}$$

Meaning that when fail to reject H_0, there is no correlation between the forecasted field and observations. On the other hand, reject H_0 showed that there is correlation between the forecasted field and observation.

4 Results and Discussion

The historical rainfall data from period 1975 to 2004 for 40 stations is chosen as the predictands. Before running the CCA, the rainfall for November-December-January-February (NDJF) which is average daily rainfall for NDJF is prepared. Since anomaly correlation coefficient is needed, the mean have to be removed from the data as to obtain the anomalies. The same procedure is proposed for the SLP data. Using the CCA, the historical rainfall and SLP formed a new set of forecast of NDJF rainfall data. With the anomaly of the forecast rainfall and the historical rainfall, its anomaly correlation coefficient is calculated.

In this paper, only the first eigenvector is used since the first eigenvalue provides the largest variances. Hence, the CCA equation that have resulted from 40 stations and 20 grid points is

$$-0.1762\,y_1 - 0.1792\,y_2 - 0.2637\,y_3 - 0.1863\,y_4 - 0.1993\,y_5 - 0.1374\,y_6 - 0.1929\,y_7 - 0.1240\,y_8 - 0.1575\,y_9 - 0.1315\,y_{10} - 0.1616\,y_{11} - 0.0940\,y_{12} - 0.1174\,y_{13} - 0.1656\,y_{14} - 0.1437\,y_{15} - 0.0719\,y_{16} - 0.1967\,y_{17} - 0.1849\,y_{18} - 0.1106\,y_{19} - 0.1489\,y_{20} - 0.1130\,y_{21} - 0.0853\,y_{22} - 0.1905\,y_{23} - 0.0740\,y_{24} - 0.0896\,y_{25} - 0.1725\,y_{26} - 0.0513\,y_{27} - 0.1346\,y_{28} - 0.1379\,y_{29} - 0.1325\,y_{30} - 0.1856\,y_{31} - 0.1710\,y_{32} - 0.2119\,y_{33} - 0.1852\,y_{34} - 0.1812\,y_{35} - 0.1816\,y_{36} - 0.1287\,y_{37} - 0.1692\,y_{38} - 0.1869\,y_{39} - 0.1542\,y_{40} = 0.2225\,x_1 + 0.2224\,x_2 + 0.2235\,x_3 + 0.2236\,x_4 + 0.2231\,x_5 + 0.2227\,x_6 + 0.2237\,x_7 + 0.2250\,x_8 + 0.2243\,x_9 + 0.2239\,x_{10} + 0.2229\,x_{11} + 0.2247\,x_{12} + 0.2255\,x_{13} + 0.2245\,x_{14} + 0.2227\,x_{15} + 0.2216\,x_{16} + 0.2236\,x_{17} + 0.2247\,x_{18} + 0.2241\,x_{19} + 0.2221\,x_{20} - 451877.3217.$$

With the equation above, the forecasted rainfall can be estimated. The correlation coefficient between the forecasted and observation rainfall can also be calculated. The result shows that $\left[\frac{s_o}{s_y}\right] r \neq 1$, which means that the model is conditionally biased. The result for the third terms in Equation 14 shows the existence of unconditional biases in the forecasts. Since the bias is small as compared to the variance of observations, the reduction of the skill will be modest.

Table 2 correlation coefficient on regions

No	Regions	Correlation skill	No	Regions	Correlation skill
1	C	+	21	NW	+
2	C	-	22	NW	+
3	C	+	23	NW	+
4	C	+	24	NW	+
5	C	+	25	SW	+
6	C	-	26	SW	-
7	C	+	27	SW	+
8	C	-	28	SW	-
9	E	+	29	SW	-
10	E	+	30	SW	+
11	E	-	31	SW	--
12	E	+	32	SW	+
13	E	-	33	W	+

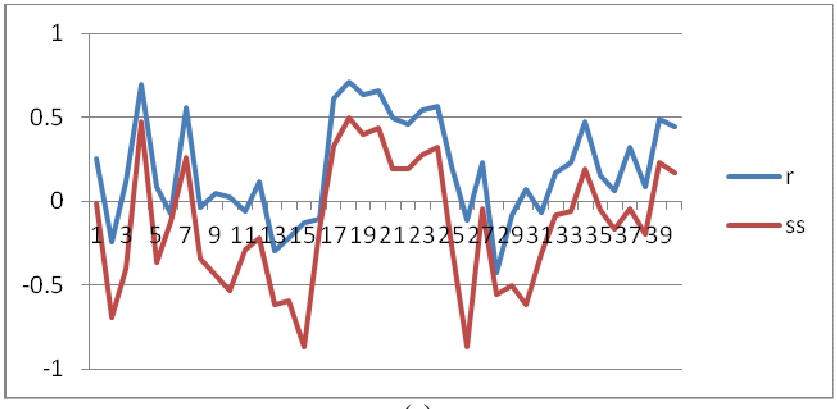
14	E	-	34	W	+
15	E	-	35	W	+
16	E	-	36	W	+
17	NW	+	37	W	+
18	NW	+	38	W	+
19	NW	+	39	W	+
20	NW	+	40	W	+

Correlation coefficients have different results on the regions of Malaysia. From Table 2, it is clearly shown that the North West and West regions have positive correlation coefficients. However, the center, the East and the South West regions have a mixture of negative and positive correlation coefficients.

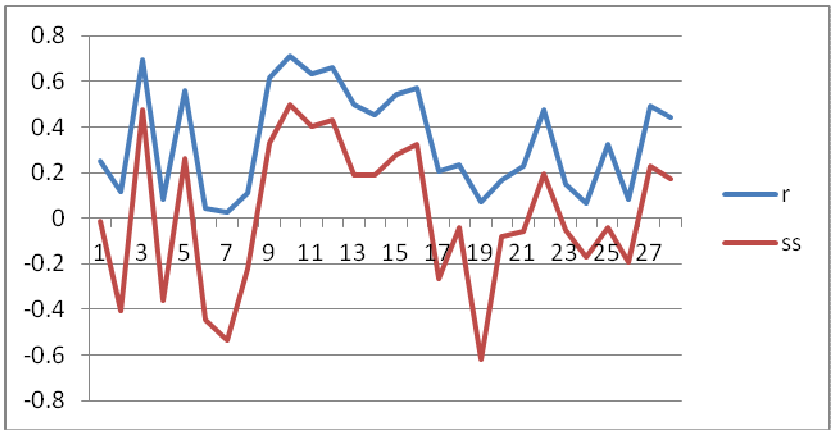
As can be seen, most of the negative correlation coefficients occurred on the eastern regions. This may due to the Northeast Monsoon which brings heavy rain particularly to the east coast regions from November to February. Positive correlations indicated that the observed and forecasted rainfalls behave in tandem and in the same direction while negative correlation showed that the direction is different.

Table 3 Root Mean Square Error (RMSE) of each region

Center	East	North West	South West	West
1.113869	1.364682	0.833659	1.015241	0.904081
1.621464	1.289265	0.835052	0.990412	0.946451
1.318869	1.156871	0.832569	0.99197	0.92195
1.072874	1.240648	0.841045	0.997896	0.90032
1.304533	1.205652	0.843234	0.988419	0.905586
1.109375	1.248473	0.830928	1.001482	0.944675
1.082227	1.375784	0.831459	1.000279	0.896265
1.617213	1.02753	0.829792	0.999208	0.907794



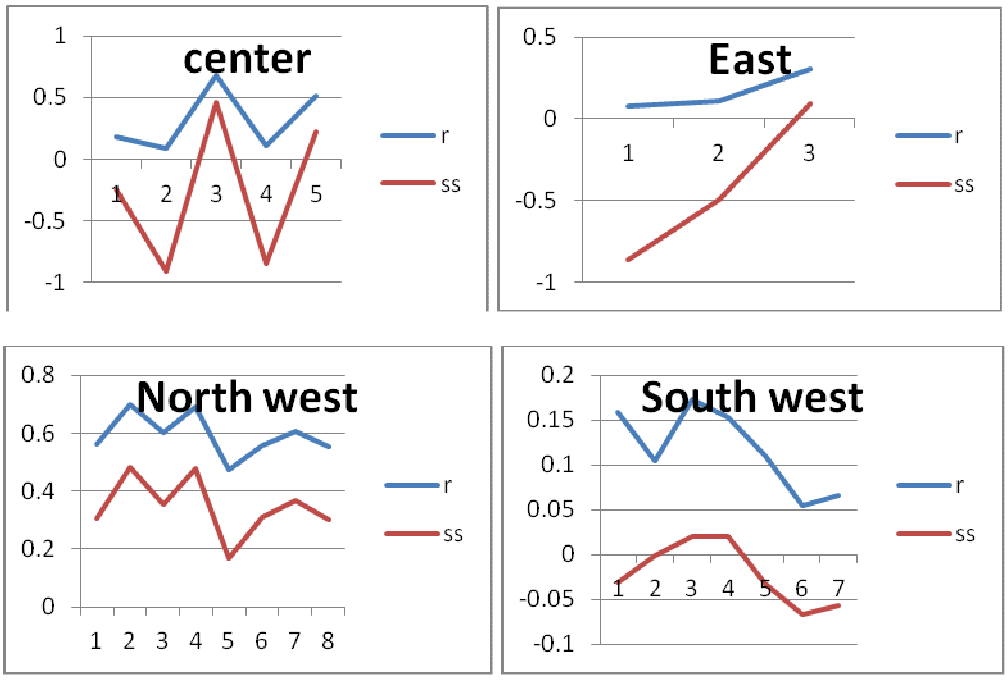
(a)



(b)

Figure 2 Plot of (a) all anomaly correlation coefficient and (b) positive anomaly correlation coefficient with the skill score for Peninsular Malaysia

Table 3 shows result of Root Mean Square Error (RMSE) calculated from the forecast. Small value of RMSE gives a good forecasting; hence, North West seems to have better forecasting than the others since the values of RMSE in this region are small. Figure 2 shows the plot of anomaly correlation coefficient (*r*) with the skill score (*ss*). Figure 2 (a) shows the overall pattern of *r* and *ss* for all 40 stations. From the figure, when *r* is negative they match poorly with the *ss* when compare to the positive correlation. The negative correlation gives inconsistence pattern of the plot especially at stations 13-16. Therefore, only positive correlation will be used in this study and hence only 28 stations will be retained from the overall 40 stations. From Figure 2 (b), the trend of the plot *r* follows the trend of *ss* even though they do not match equally. As can be seen, the anomaly correlation skill is higher than the skill score, meaning that some biases did appear in the forecast.



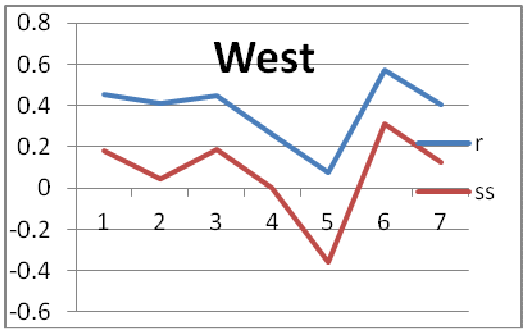


Figure 3 Plot of positive anomaly correlation coefficient with the skill score for Peninsular Malaysia according to regions

Figure 3 shows the plot of positive anomaly correlation coefficient with the skill score for Peninsular Malaysia according to regions. The trend for both *r* and *ss* are similar on North West and West regions. The center, east and South West show slightly different for both *r* and *ss*. The differences in pattern may due to the Northeast Monsoon. Hence, it indicates that North West and West regions match perfectly for the NDJF rainfall and SLP but other regions may be affected by the Northeast Monsoon. Significance testing for the forecasts and observations showed that *p*-value for the Fisher *Z* transformation is 0.0726. Therefore, there are correlations between the forecasts field and observations at 90% significance level.

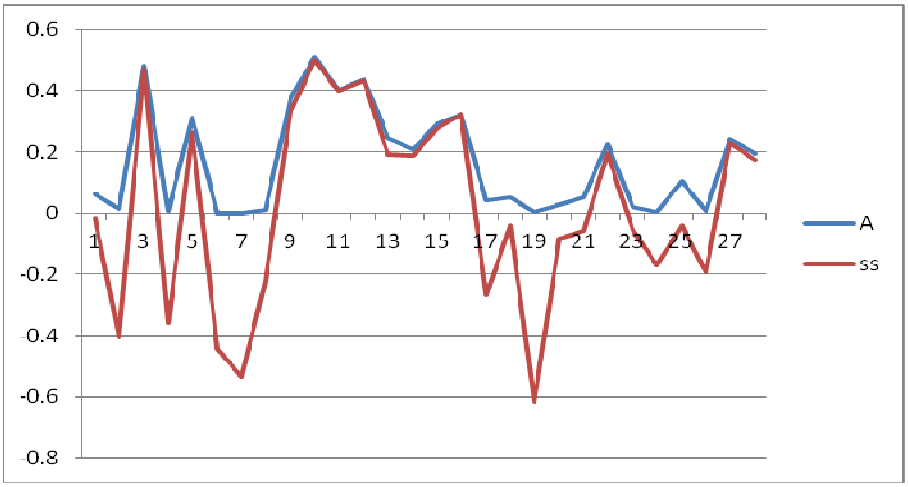


Figure 4 Plot of square anomaly correlation coefficient (A) with the skill score

As an improvement to the skill score, many calculations have been done. In this paper, the square of the anomaly correlation skill is calculated. Hence, it is clearly showed that the square of the anomaly correlation skill is better than the anomaly correlation skill (as in Figure 4). As can be seen, some of the skill scores match perfectly with the square anomaly correlation coefficient (A) especially stations of number 9 to 16. This may due to these stations are in the same region, the North West region, and do not influence by the monsoon. The matching of the plot indicates that square anomaly correlation coefficient can be used as an improvement to the skill score. From the MSE calculation, the model has conditional biases and that implies the square of the correlation coefficient will overestimate the score skill.

5 Conclusion

In this study, verification model for the CCA using NDJF rainfall and SLP is done. The anomaly correlation skills for North West and West regions have positive correlation. The plot of anomaly correlation coefficient and the skill score for Peninsular Malaysia are having similar patterns. However, when the Peninsular is divided into regions, the plot of anomaly correlation coefficient and skill score are perfectly matched for the North West and West regions. The model was also found to have conditional and unconditional biases and the correlation between observed and forecasted fields are significant up to the 90% level. . The biases exist may be related to the penalty associated with retaining the Sea Level Pressure (SLP) in the meteorological features when such features are not predictable. Hence, the square of the correlation coefficient will overestimate the skill. Therefore, the squared correlation is best regarded as measuring the potential skill. Since the bias is small as compared to the variance of the observations, the reduction of the skill will be modest. Only SLP is done in this study, hence, for future study, more atmospheric variables may be included in the model and would be useful to improve the forecast performance.

Acknowledgements

The authors would like to thank the Malaysian Meteorological Department and Malaysian Brainage and Irrigation Department for the data of the surface variables (rainfall and Universiti Teknologi Malaysia (UTM) for the Research University Grant (RUD) through Vot number 04J23.

References

[1] Hessami, M., Gachon, P., Ouarda, T. B.M.J. and St-Hilaire, A. Automated regression-based statistical downscaling tool. *Environmental Modelling & Software*. 2008. 23: 813-834.
[2] Wilks, D. S. *Statistical Methods in the Atmospheric Science, Second Edition*. United States: Elsevier. 2006.
[3] Roebber, P. J. and Bosart, L. F. The complex relationship between forecast skill and forecast value: A real-world analysis. *Weather and Forecasting*. 1996. 11: 544-559.
[4] Huth, R. Statistical downcsaling in central Europe: Evaluation of methods and potential predictor. *Climate Research*. 1999. 13: 91-101.
[5] Tatli, H., Dalfes, H. N. and Mentes, S. S. A statistical downscaling method for monthly total precipitation over Turkey. *Int. J. Climatol*. 2004. 24: 161-180.
[6] Busuioc, A., Tomozeiu, R. and Cacciamani, C. Statistical downscaling model based on canonical correlation analysis for winter extreme precipitation events in the Emilia-Romagna region. *Int. J. Climatol*. 2008. 28: 449-464.
[7] Juneng, L. and Tangang, F.T. Level and source of predictability of seasonal rainfall anomalies in Malaysia using canonical correlation analysis. *Int. J. of Climatol*. 2008. 28: 1255-1267.
[8] Murphy, A. H. and Epstein, E. S. Skill scores and correlation coefficients in model verification. *Monthly Weather Review*. 1988. 117: 572-581.
[9] Murphy, A. H. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*. 1988. 116: 2417-2424.
[10] Murphy, A. H. General decompositions of MSE-based skill scores: measures of some basic aspects of forecast quality. *Monthly Weather Review*. 1996. 124: 2353-2369.
[11] Tonani, M., Pinardi, N., Fratianni, C., Pistoia, J., Dobricic, S., Pensieri, S., de Alfonso, M. and Nittis, K. Mediterranean forecasting system: forecast and analysis assessment through skill scores. *Ocean Sci*. 2009. 5: 649-660.
[12] Krishnamurti, T. N., Rajendran, K., Vijaya Kumar, T. S. V., Lord, S., Toth, Z., Zou, X. L., Cocke, S., Ahlquist, J. E. and Navon, I. M. Improved skill for the anomaly correlation of geopotential heights at 500 hPa. *Monthly Weather Review*. 2001. 131: 1082-1102.

[13] Hsieh, W. Nonlinear canonical correlation analysis by neural networks. *Neural Networks*. 2000. 13:1095-1105.