The potential of canonical correlation analysis in multivariable screening of climate model

The potential of canonical correlation analysis in multivariable screening of climate model

N N A Tukimat [1,2*], S Harun [3], M Y M Tadza [1]

[1]Faculty of Civil Engineering & Earth Resources, Universiti Malaysia Pahang, Malaysia
[2]Earth Resources & Sustainability Center (ERAS), Universiti Malaysia Pahang, Malaysia
[3]School of Civil Engineering, Universiti Teknologi Malaysia, Malaysia

*E-mail: nadrah@ump.edu.my

Abstract. The statistical downscaling model (SDSM) been used to analyse the potential changes of local climate trend in the long term. The difficulty of the SDSM model in selecting the best predictors group which having good association to the local climate. Even the SDSM provides screening process to analyse the predictor-rainfall relationship, however it has limited ability in analysing multiple variables from 26 predictors with 10 rainfall stations around Kedah state, Malaysia. In this regard, the Canonical Correlation Analysis (CCA) been used to analyse the multi predictor-rainfall relationships. The concept of canonical coefficient is sufficient to show the capability and reliability of the predictors based on the percentages of variance that can explained in the dependent variable using the independent variable. There were 10 predictors' group have been developed and one predictor's group was built based on the CCA result. The performances of these predictors groups were tested using statistical analyses. Results revealed that the predictors group selected by the CCA method has produced smaller values of MAE and MSE for all stations except at station of Ladang Tanjung Pauh. The box plot's results, which generated from one hundred simulated samples, indicated that the performance of CCA method was remarkable. The presence of discrepancies in the HadCM3-A2 and HadCM3-B2 scenario simulations were relatively small and considered acceptable.

1. Introduction

Today, downscaling has become the imperative model to bridge the spatial and temporal resolution of General Circulation Models (GCMs) in the direction of the local-scale surface weather. The GCMs' predictors providing coarse spatial resolution (50,000 km^2) and failed to resolve the important sub grid scale features. Therefore, an aid from downscaling such as dynamical and statistical downscaling will evaluate them becomes finer scale of meteorological variables around 10km [1] at a particular study area. Thus, statistical downscaling (SD) is the most largely used in the context of hydrologic impact studies due to the climate scenarios because it provides station-scale climate information from grid resolution GCM-scale using multiple regression techniques. These regressions emerged from the empirical statistical relationship between atmospheric circulation pattern (predictors) and local-scale parameters (predictands). Other advantages of using SDSM tools in the projection of future climate are computationally undemanding, low cost and simple assess, which make this model the most popular model among the researchers. The potential of SD, also studied by [2];[3];[4] proved the capability and the reliability of the SD simulation.

Yet, the accuracy of the climate simulation in SDSM refers to the predictors' selections that have better association with the particular local surface climate. The predictors typically derived from sea level pressure, geopotential height, wind fields, temperature variables, specific and relative humidity. The selection of predictors refers to the some criteria and behaviors. These predictors should reliably simulated by GCM, readily available from archives of GCM output, and strongly correlated with the surface variable of interest [1]. Based on the previous study, many methods including statistical analyses been applied in effort to achieve the closest calibration in measuring the potential of relationship among predictors and predictands. There were Multiple Linear Regression (MLR), Non-linear Programming (NLP) and Canonical Correlation Analysis (CCA) [5];[6];[7];[8];[9].

CCA method is used to define the input variables between historical seasonal average rainfall occurrence probabilities and GCM's simulated seasonal mean rainfall amount. However, most studies used CCA method to reduce the subspace between predictors and predictands linearity and focused on the seasonal climate at that particular area. However, the selection of the right predictors is still uncertain and suspicious. Therefore, the aim of this study is to evaluate and compare the performance of the predictors' selection based on the canonical coefficient for ten different locations of rainfall station with 10 other selected groups of predictors.

2. Material and Methods

The methodology of the study as shown in the Figure 1 meanwhile the list of the predictors groups as shown in the Table 1. In this study, the SDSM version 4.2 used to downscale the GCM output at regional scale and projection of temperature in the study area over the years 2010-2099. The SDSM is a hybrid tool which can predicts climate change at local scale by linking local climate variables with large-scale atmospheric variables using multiple regression technique [1]. Therefore, downscaling requires two types of data viz. predictand and predictor at the grid box of 28X x 33Y. The first type is the National Centre of Environmental Prediction (NCEP) reanalysis data set from 1961 until 1990 for calibration (1961-1975) and validation process (1976-1990). The second type is the GCMs predictors, namely Hadley Center General Circulation Model (HadCM3) of A2 and B2 scenarios (1961-2099) for the projection of climate scenarios.

There were 10 predictors' groups were selected based on the previous studies suggested by [10] and SDSM screening that had a better empirical statistical relationship with the precipitation. Many researchers agreed that the mean sea level pressure and geopotential height fulfill the criteria as precipitation predictors. Additionally, one more predictor group known as CCA Group selected based on the canonical coefficient analysis. The analysis calculated using data from the ten rainfall stations (multiple independent variables) and the 26 predictors from NCEP data (multiple dependent variables). Then, five predictors were group as a CCA group.

2.1 CCA Analysis

The choice of input variables (predictors) is non-trivial. The CCA is a multivariate statistical model, which measures the linear relationship and maximizes the relation between multiple dependent variables and multiple independent variables. The CCA is the combination of the principle component and factor analysis with MANOVA concept. It assists the researcher in measuring the interaction of data from two set of variables and concern with the variance among data set [11].

In this paper, the CCA applied to screen the potential of multiple variables from predictors and rainfall stations (known as predictand) and the association among them. The concept of the canonical coefficient was adequate to show the capability and reliability of predictors that can produce a better simulation based on the local surface climate (predictand) at multiple rainfall stations. The generated correlation coefficient value in CCA analysis shows the percentage of variance that can be explained in the form of multi dependent variable, by using the multi independent variable and also giving the criterion variables (product innovation variables) for each of them. The formula for the correlation matrix r_{xy} is:

$$Cov(xy) = \frac{1}{N}\sum(x_i y_i - \overline{xy}) \tag{1}$$

$$r_{xy} = \frac{Cov(xy)}{\sqrt{s_x^2 s_y^2}} \tag{2}$$

where x_i and y_i refer to the predictands and predictors data, $\overline{xy}$ is mean value of both variables, while s_x and s_y refer to their standard deviation. The capability among variables will be interpreted as values between -1 to 1 which shows the positive/negative association among them.

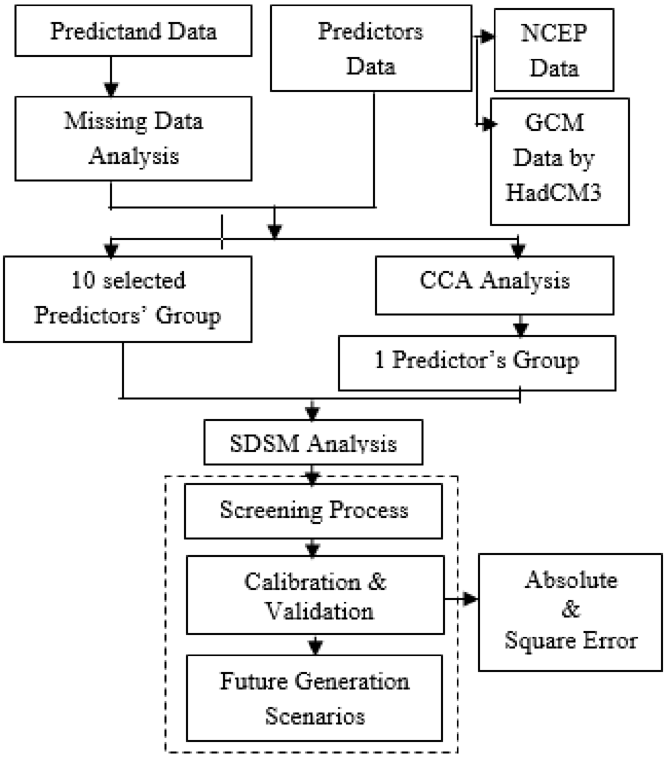

Figure 1. Methodology of the study

2.2 Study area

Kedah is located at north of peninsular Malaysia. About 97,000 hectares of its land covered by the largest double paddy cultivation area in Malaysia. Geographically, the area lies between 5°45'–6°30'N latitude and 100°10'-100°30'E longitude. The topography of the area is almost flat with a slope ranging from 1 in 5,000 to 1 in 10,000. The climate of the area like other parts of Malaysia can be classified into four seasons viz. south-west monsoon (May–Sept), north-east monsoon (Nov–Mar) and two inter-monsoon seasons. Dec–Feb and June–July are considered as warm seasons in the area, while Apr–May and Sept–Nov are considered as humid seasons.

The type of soil in the study area is heavy clayey in nature. The mean temperature varies between 27°C and 32°C. The relative humidity at this area is fluctuates between 54 % and 94 %. Thus, 10 rainfall stations that have been identified were considered from the quality of available rainfall records and its location in the Muda Irrigation Scheme area. The locations of stations, known as predictands as listed in Table 2.

Table 1. List of 10 predictors group

No	Predictor Group	Sources	No	Predictor Group	Sources
G1	Mean Sea Level Pressure	[7]	G2	Zonal Velocity	SDSM screening
	Surface Divergence			Meridional Velocity	
	500hPa Geopotential Height			850hPa Vorticity	
	850hPa Zonal Velocity			850hPa Geopotential Height	
	Specific Humidity			Specific Humidity	
G3	Airflow Strength	SDSM screening	G4	500hPa Zonal Velocity	Based on 500hPa
	Wind Direction			500hPa Meridional Velocity	
	850hPa Meridional Velocity			500hPa Geopotential Height	
	850hPa Surface Divergence			500hPa Vorticity	
	Relative Humidity			Specific Humidity	
G5	500hPa airflow Strength	Based on 500hPa	G6	850hPa Zonal Velocity	Based on 850hPa
	500hPa Wind Direction			850hPa Meridional Velocity	
	500hPa Surface Divergence			850hPa Vorticity	
	500hPa Relative Humidity			850hPa Geopotential Height	
	Specific Humidity			Specific Humidity	
G7	850hPa Airflow Strength	Based on 850hPa	G8	Mean Sea Level Pressure	SDSM screening
	850hPa Wind Direction			Zonal Velocity	
	850hPa Relative Humidity			500hPa Geopotential Height	
	850hPa Geopotential Height			Meridional Velocity	
	Specific Humidity			Specific Humidity	
G9	Mean Sea Level Pressure	SDSM screening	G10	Meridional Velocity	SDSM screening
	850hPa Surface Divergence			Surface Divergence	
	850hPa Vorticity			850hPa Zonal Velocity	
	850hPa Geopotential Height			850hPa Meridional Velocity	
	Relative Humidity			Relative Humidity	

Table 2. List of rainfall stations

	Location	Geographical coordinate		Station No
		Longitude (N)	Latitude (E)	
1	Gajah Mati	6^{o} 11’	100^{o} 32’	6105037
2	Ibu Bekalan Tupah	5^{o} 45’	100^{o} 27’	5704057
3	Kedah Peak	5^{o} 48’	100^{o} 25’	5704055
4	Keretapi Tokai	6^{o} 02’	100^{o} 25’	6004045
5	Kodiang	6^{o} 23’	100^{o} 18’	6302021
6	Kota Sarang Semut	5^{o} 59’	100^{o} 24’	5904051
7	Pendang	6^{o} 00’	100^{o} 29’	5904043
8	Sungai Limau	5^{o} 54’	100^{o} 23’	5803052
9	Telok Chengai	5^{o} 54’	100^{o} 23’	6004045
10	Ladang Tanjung Pauh	6^{o} 14’	100^{o} 26’	6204028

3. Result and Discussion

3.1 Performances of Simulated Results

The performances of the simulated results were analysed based on the comparison performances between historical and simulated in year period of 1961-1990. According to the CCA analysis, there were 5 predictors have been selected to form CCA group; 500hPa zonal velocity, airflow strength, 500hPa relative humidity, 850hPa meridional velocity and specific humidity. The selection made based on the canonical coefficient values from CCA analysis.

The performances of the simulated results produced by each groups were analysed based on the mean absolute error (MAE) and mean square error (MSE) as shown in Table 3. The result shows that the CCA predictor group was excellent and nearest to the historical data for each rainfall station except at stations of Ladang Tanjung Pauh and Keretapi Tokai. An error became higher which influenced by the frail association in the air-flow strength, meridional velocity and zonal velocity that produced smaller value in canonical coefficient. Nevertheless, the error at those locations was still low and acceptable.

The MAE and MSE were 0.607 and 0.520 respectively for Ladang Tanjung Pauh, and 1.68 and 4.50 for Keretapi Tokai, which still better compared to others 10 predictors groups. The results revealed that the values of MAE and MSE became lesser and getting better if the predictors chosen were based on the canonical coefficient. It also proved that the canonical coefficient is reliable in the context of predictor selection. The plots in Figure 2 also showed that the performance of the SDSM model using NCEP predictors for the simulation was reliable and unsuspicious.

Box plot in Figure 3 plotted for daily precipitation with CCA predictor group for one hundred simulations for each rainfall station. Box plots of the mean value present graphical plot to describe five summaries of numerical data: smallest, lower quantile, mean, upper quantile, and the largest simulated distribution.

Table 3. MAE and MSE results for monthly mean precipitation (mm)

Rainfall Station	**MAE**										
	CCA	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Gajah Mati	0.76	0.95	0.93	1.16	1.05	1.73	0.79	1.20	1.08	0.88	0.84
Ibu Bekalan Tupah	2.02	2.76	2.55	3.04	2.85	2.76	2.35	2.23	2.76	2.74	2.77
Kedah Peak	1.63	2.28	1.96	2.18	2.52	2.19	1.95	2.24	2.25	1.69	2.58
Keretapi Tokai	1.68	1.89	2.01	2.04	2.05	1.89	1.81	1.64	1.88	1.77	1.82
Kodiang	0.45	1.36	1.33	1.04	1.51	1.37	1.11	1.36	1.42	1.30	0.50
Kota Sarang Semut	0.69	0.93	1.01	1.70	1.07	1.48	1.42	1.28	0.96	1.03	0.86
Ladang Tanjung Pauh	0.61	0.63	0.47	1.50	0.89	1.26	0.60	1.18	0.63	0.65	0.81
Pendang	0.77	1.11	1.07	1.07	1.08	1.73	1.13	1.93	1.06	1.01	0.89
Sungai Limau	1.59	2.56	2.34	1.89	3.03	2.26	2.26	2.39	2.61	2.77	1.94
Telok Chengai	1.05	1.75	1.61	1.12	1.73	1.71	1.56	1.37	1.72	1.67	1.06

Rainfall Station	**MSE**										
	CCA	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Gajah Mati	0.87	1.47	1.60	2.14	1.73	4.67	1.24	2.85	1.86	1.17	1.12
Ibu Bekalan Tupah	5.37	9.74	8.41	11.05	10.18	11.00	7.30	6.43	9.63	9.52	9.70
Kedah Peak	4.09	8.04	5.45	6.47	8.60	6.56	6.13	6.80	7.55	4.90	8.90
Keretapi Tokai	4.50	5.68	5.87	6.01	5.84	5.48	5.45	5.06	5.41	5.25	5.47
Kodiang	0.38	2.30	2.29	1.74	2.97	3.74	1.59	2.68	2.53	2.21	0.46
Kota Sarang Semut	0.90	1.29	1.44	4.44	1.50	3.00	6.74	3.25	1.31	1.43	1.16
Ladang Tanjung Pauh	0.52	0.68	0.40	3.59	1.00	2.40	0.72	2.25	0.64	0.63	1.01
Pendang	0.92	2.00	1.82	1.95	1.86	3.59	2.04	4.59	1.87	1.80	1.28
Sungai Limau	3.38	8.09	6.84	4.86	10.35	7.14	6.15	7.87	8.45	9.17	4.69
Telok Chengai	1.62	4.09	3.63	2.11	3.97	4.71	3.60	3.08	4.01	4.01	1.64

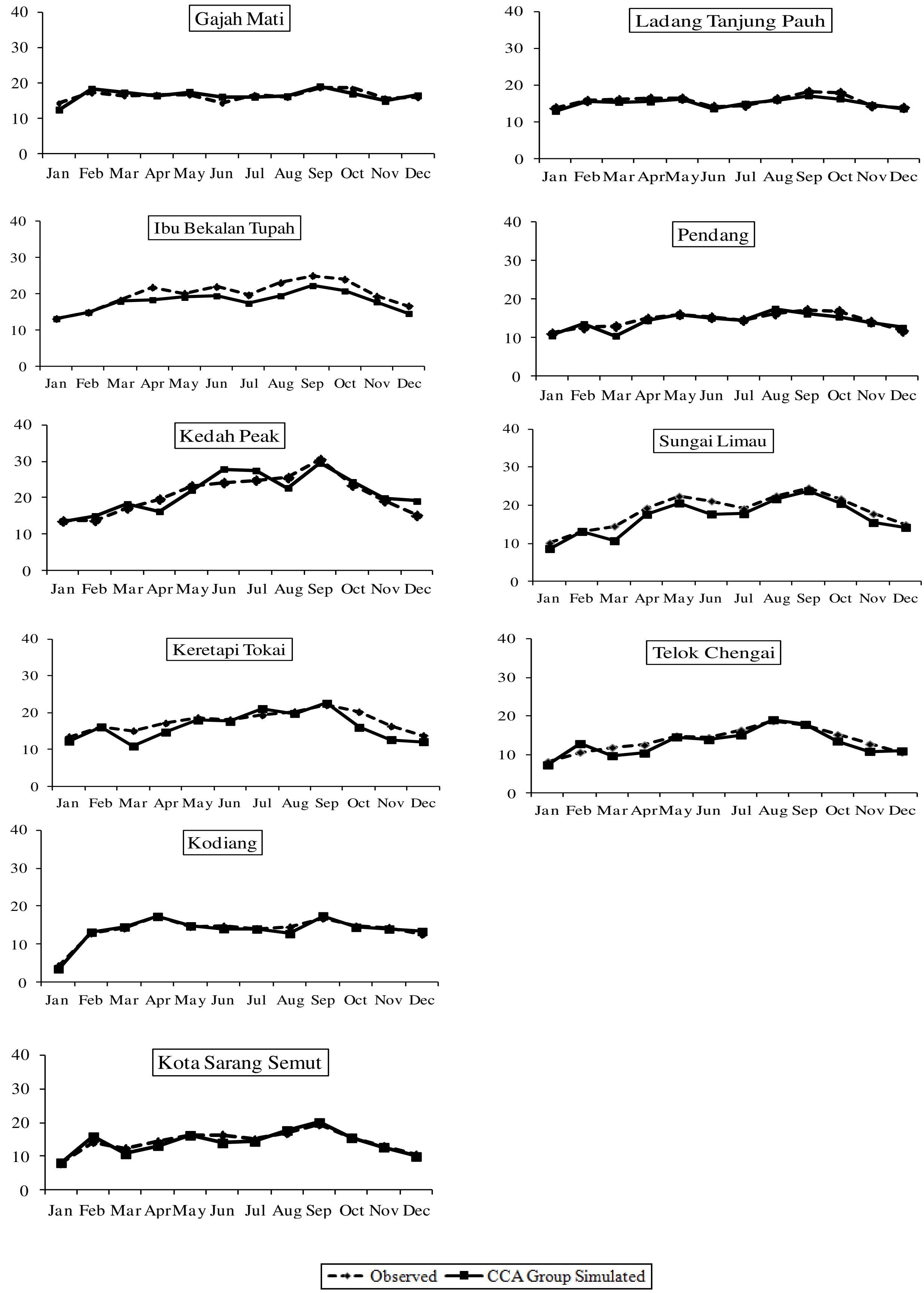

Figure 2. The comparison of validation process (1976-1990) between observed data with simulated value based on CCA predictor's group.

The function of the box plot was to assess the performance of CCA analysis based on the one hundred SDSM simulations. The box plots above clearly showed that the performance of the NCEP predictors were excellent for calibration and validation process in SDSM for each station starting from 1961 until 1990. Furthermore, the descripancies in interquartile range (IQR) for the validation had slight difference and can still be accepted. Most of the locations produce lesser change between 25 percentile of value and 75 percentile of value and the middle value of box was used as the average value for the simulation. This proved that the selected predictors based on the canonical coefficient were able to produce better simulation corresponding to the change in local predictand as well.

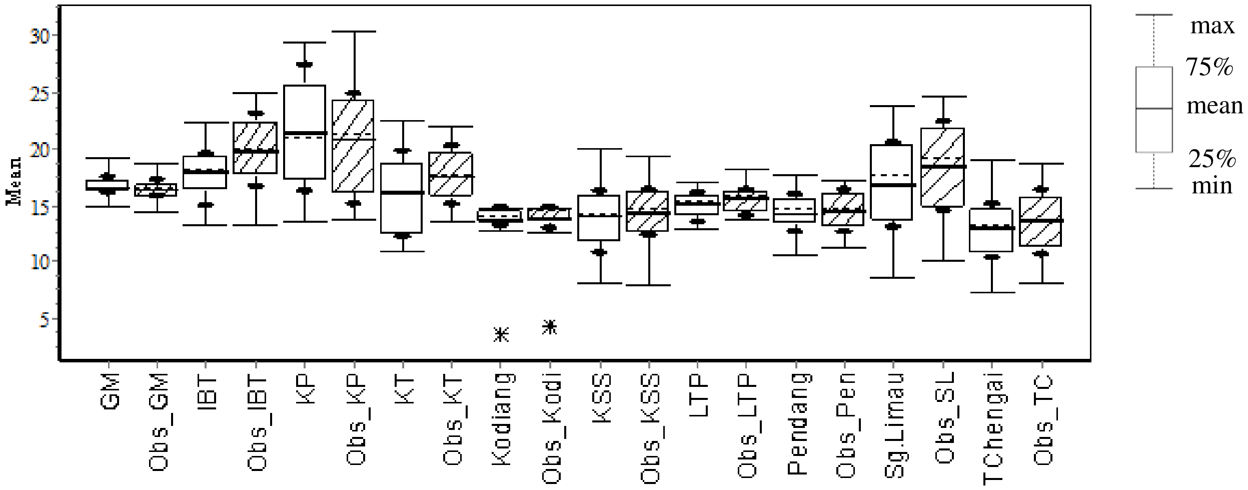

Figure 3. Box-plot of SDSM downscaling result for comparing between observed and CCA group simulated

4. Conclusion

CCA can implemented in measuring the linear relationship and maximizes the relation among multiple dependent variables and multiple independent variables. This study proved the performances of CCA predictors group based on the MAE, MSE and consistency of box plot. The concept of canonical coefficient is sufficient to show the capability and reliability of the predictors based on the percentages of variance that can explained in the dependent variable using the independent value. MAE and MSE values for CCA group produced a lesser value compared to other group except for Ladang Tanjung Pauh, which prefers predictors in G2. Moreover, box plots for each location provide a consistent value for one hundred simulations and this proved that the selected predictors based on the canonical coefficient were able to produce better simulation which corresponds to the change in local predictand as well.

References

[1] Wilby, R. L., & Dawson, C. W. 2007, SDSM 4.2 — A decision support tool for the assessment of regional climate change impacts.

[2] Chen, S.-T., Yu, P.-S., & Tang, Y.-H. 2010, Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology, 385*(1-4), 13-22. DOI: 10.1016/j.jhydrol. 2010. 01.021

[3] Sharma, M., Coulibaly, P., Dibike, Y., 2010, Assessing the need for Downscaling RCM Data for Hydrologic Impact Study. *Journal of Hydrologic Eng.* 16, 534(2011) DOI:10.1061/(ASCE)HE.1943-5584

[4] Ethan D. G., Roy M. R., Changhai L., Kyoko I., David j. G., Martyn P. C., Jimy D., Gregory, T., A., 2011, Comparison of Statistical and Dynamical Downscaling of Winter Precipitation Over Complex Terrain. Journal of Climate, American Meterological Society. ISSN: 1520-0442.DOI: 10.1175/2011JCLI4109.1

[5] Gangopadhyay, S., Clark, M. P., Rajagopalan, B. 2002, Statistical Downscaling: A Comparison of Multiple Linear Regression and k-Nearest Neighbor Approaches. *American Geophysical Union*, The Smithsonian/NASA Astrophysics Data System

[6] Busuioc A, Giorgi F, Bi X, Ionita M. 2006, Comparison of Regional Climate Model and Statistical Downscaling Simulations of Different Winter Precipitation Change Scenarios over Romania. *Theoretical and Applied Climatology 86:* 101–123.

[7] Benestad R.F. 2002, Empirically downscaled multimodel ensemble temperature and precipitation. Scenarios for Norway. *Journal of Climate 15: 3008–3027*

[8] Chen, D., & Chen, Y. 2003, Association between winter temperature in China and upper air circulation over East Asia revealed by canonical correlation analysis. *Global and Planetary Change, 37*(3-4), 315-325. DOI:10.1016/S0921-8181(02)00206-0

[9] Huth R. 2004, Sensitivity of Local Daily Temperature Change Estimates to the Selection of Downscaling Models and Predictors. *Journal of Climate 17: 640–652*. DOI: 10.1175/1520-0442(2004)017

[10] Dibike, Y. B., Gachon, P., St-Hilaire, A., Ouarda, T. B. M. J., & Nguyen, V. T. 2008, Uncertainty analysis of statistically downscaled temperature and precipitation regimes in northern canada. *Theoretical and Applied Climatology, 91*(149-170), DOI: 149-170. 10.1007/s00704-007-0299-z

[11] Weiss D.J. 1972, Canonical correlation analysis in counselling psychology research. *Journal of Counsel. Psychol., 19*, 241–252. DOI:10.1037/h0032675