A non-local model output statistics approach for the downscaling of CMIP5 GCMs for the projection of rainfall in Peninsular Malaysia

Muhammad Noor, Tarmizi bin Ismail, Shahid Ullah, Zafar Iqbal, Nadeem Nawaz and Kamal Ahmed

ABSTRACT

In this study, a non-local model output statistics (MOS) approach is proposed for the downscaling of daily rainfall of coupled model intercomparison project phase 5 (CMIP5) general circulation models (GCMs) for the projections of rainfall in Peninsular Malaysia for two representative concentration pathway (RCP) scenarios, RCP4.5 and RCP8.5. Projections of eight GCMs for both the mentioned RCPs were used for this purpose. The GCM simulations were downscaled at 19 observed stations distributed over Peninsular Malaysia. Random forest (RF) was used for the development of non-local regression-based MOS models. The results revealed a high accuracy of the models in downscaling rainfall at all the observed stations. The mean absolute error (MAE) of the models were found in the range of 0.8 – 0.39; normalized root mean square error (NRMSE) between 7.4 and 41.7, Percent Bias (PBIAS) between -0.3 and 10.1, Nash-Sutcliffe coefficient (NSE) between 0.81 and 0.99 and R² between 0.89 and 0.99. The increase in annual rainfall was in the range of 7.3 – 29.5%. The increase was higher for RCP8.5 compared to RCP4.5. The maximum increase was observed in the northern part of Peninsular Malaysia in the range of 20.7 – 29.5%, while the minimum in the south-west region was in the range of 7.6 – 15.2%.

Key words | radiative concentration pathways, rainfall projections, random forest, statistical downscaling

INTRODUCTION

Climate change is one of the most serious threats to the environment faced by mankind in the 21st century. Due to the increase in the concentration of carbon dioxide (CO₂) and other greenhouse gases (GHGs), the earth is warming to hazardous levels (Salman et al. 2018). The Intergovernmental Panel on Climate Change (IPCC) has reported an increase in global mean temperature by 0.56–0.92 °C between the years 1905 and 2005. It also projected an increase in global mean temperature by 2–4.5 °C by the end of this century (IPCC 2007). The consequences of global warming are a rise in sea level, variability in rainfall, higher temperature, severe and frequent storms (Khan et al. 2019). Due to the increase in temperature, the variation in rainfall is increasing throughout the globe (Shiru et al. 2018; Iqbal et al. 2019). According to the IPCC, in the 21st century extreme rainfall events with increasing frequency under various emission scenarios are expected to occur (IPCC 2007). In recent decades, more frequent extreme storm events with greater severity have been observed in urban cities resulting in a huge economic, environmental, infrastructural and human loss (Shahid et al. 2015). In many studies using the model simulations and climate scenarios, the impacts of changing climate on the design of urban water management infrastructure have been evaluated in various parts of the world. All of these studies have a consensus over the increased rainfall intensity and the
potential under-designing of the urban water management infrastructure.

There is a consensus that global temperature trends are increasing around the world, including Malaysia. The mean temperature in Malaysia is found to increase from 0.6 to 1.2 °C per 50 years (Khan et al. 2019). According to MMD (2009), remarkable variations have been found in intensity and frequency for both the floods and droughts in Malaysia. The future flood flow has an increasing range of 11–43%, and a decreasing range of 31–93% (San Liew et al. 2014). Increase in temperature and variations in rainfall patterns can definitely affect hydrological regimes as well as the quantity of runoff in Malaysia (Alias et al. 2016; Ismail et al. 2017). For developing an effective policy to adapt to these climatic variations, it is necessary to analyse and understand the present and possible future climate changes. One of the ways to reduce the vulnerability is to quantify and adapt to these impacts due to the changing climate.

General circulation models (GCMs) are one of the main tools used for the projection of present as well as future climate (Salman et al. 2018). However, the outputs of the GCMs are at larger resolution and cannot be directly used for understanding the dynamics of climate change on a local scale (Shiru et al. 2019). Downscaling methods are widely used to link the gap between larger resolution to the local scale (Sachindra et al. 2019). Typically, downscaling can be divided into two main groups: statistical downscaling and dynamic downscaling (Salman et al. 2018). The statistical downscaling technique is widely used in assessing the impacts of changing climate because of its simplicity and lower computational cost as compared to dynamic downscaling (Pour et al. 2014). The statistical downscaling methods are further classified into two main groups: model output statistics (MOS) and perfect prognosis (PP). MOS has received more attraction due to its capacity for model-inherent error and bias (Noor et al. 2018). In the MOS approach, a statistical relationship between predictand (observed precipitation) and predictors (GCM simulated precipitation) is established. The developed statistical relationship is then applied to project climate variables for future scenarios using GCM simulated predictors (Ahmed et al. 2018; Salman et al. 2018). In MOS downscaling the GCM simulation is corrected with respect to the observations. However, in many studies (Eden & Widmann 2014) event-wise MOS downscaling has been found more suitable. Using predictive information, the capacity of MOS can be improved by linking the local rainfall with spatial patterns in a spatial domain (Eden & Widmann 2014).

The climate of a region is influenced by the joint impact of large scale atmospheric variables in a large tempo-spatial zone (Pour et al. 2014). The atmospheric variables of a single predictor or measure at a single grid point are not adequate for downscaling and projection of the climate (Sachindra et al. 2019). For successful performance of the downscaling model, the identification of the spatial domain is of prime importance for large scale atmospheric variables (Ahmed et al. 2015). This can greatly improve the performance of the downscaling model. Therefore, the selection of a suitable climatic domain is necessary for better performance of the downscaling model (Pour et al. 2014).

The objective of the present study is to develop MOS downscaling models by linking local rainfall to spatial patterns of GCM hindcasts (precipitation) in a climatic domain. Usually, the relation between spatial patterns of GCM hindcasts and local rainfall is very complex. For modelling such types of complex non-linear relationships, data mining methods can effectively be used. Non-linear random forest (RF) has been found to be a more reliable data mining model for developing statistical downscaling models (Ahmed et al. 2018; Salman et al. 2018). Therefore, using the RF method, MOS downscaling models were developed for the projection of future rainfall in Peninsular Malaysia.

STUDY AREA AND DATA

Study area

The methodology adopted in this study was applied for assessing the variations in rainfall due to the changing climate in Peninsular Malaysia. Geographically, Peninsular Malaysia lies in the tropics between latitude 1.2° north to latitude 6.4° north, and longitude 99.35° east to longitude 104.20° east (Pour et al. 2014). The topography is composed of irregular mountainous forest regions slanting down to seashores. The climate of the study area is hot and humid and is very much influenced by the monsoon winds.
(Noor et al. 2019). The annual temperatures are found in the ranges of 21–32 °C (Khan et al. 2019). The annual aggregate of the rainfall in Peninsular Malaysia is about 2,000–4,000 mm, with 150–200 wet days per year (Noor et al. 2018). A map of the study area is shown in Figure 1.

Data

Thirty-five years (1971–2005) of observed rainfall data from 19 selected rain gauge stations distributed over Peninsular Malaysia were used for the study. The stations were selected based on the availability of long term quality data. The locations of the rainfall recording stations with station ID are presented in Figure 1.

A number of GCMs are available in the coupled model intercomparison project phase 5 (CMIP5) suit, however eight GCMs, namely BCC-CSM1.1, CSIRO-Mk5.6.0, HadGEM2-ES, IPSL-CM5A-MR, MIROC-ESM, CCSM4, NorESM1-M, MRI-CGCM3, under two representative concentration pathway (RCP) scenarios, namely RCP4.5 and RCP8.5, were employed in the present study for assessing the effects of changing climate on rainfall. Data of eight CMIP5 GCMs under two RCP scenarios (RCP4.5 and RCP8.5) for ensemble run r1p1i1 was used for the projection of rainfall at selected stations. The models simulated historical (1971–2005) and projected (2006–2099) daily rainfall data for the climatic domain of the study area which was downloaded from the IPCC data portal (www.ipccdata.org/sim/gcm_monthly/AR5/ReferenceArchive.html). The name of the GCMs with their resolution and respective centres used in the current study are presented in Table 1.

METHODOLOGY

The methodology adopted in this study for downscaling and projection of rainfall is outlined below:

1. The selected GCMs were re-gridded to a common resolution of 2° × 2°.
2. A supervised principal component analysis (PCA) was used to identify a set of predictors from 99 grids points (GCM simulated rainfall) surrounding the study area.
3. RF was used to establish a relationship between observed rainfall (station) and GCM simulated precipitation for each month separately.
4. The RF-based MOS models were used to generate future scenarios using two RCPs.

The details of the methods used in the present study are discussed in the following sections.

Figure 1 | Map of Malaysia and locations of the selected stations.
Selection of climate domain

The selection of proper climatic domain often improves the modelling results. Several authors, such as Sachindra et al. (2012), reported that climate domain should be large enough to capture the climatic phenomena of any area. Thus, in the present study, the climatic domain comprises of 99 grid points between latitudes 5.58°–13.37° N and longitudes 90°0’0”–110°0’0” E is used and is presented in Figure 2.

Supervised principal component analysis

Principal component analysis (PCA) is a multivariate statistical technique often employed in hydro climatological assessment studies for dimensionality reduction (Ahmed et al. 2018). The technique helps to obtain the finest estimates for high dimensional data series (Sa’adi et al. 2017). The PCA has the ability to capture the maximum variability in the input data; however, its efficiency is restricted to unsupervised issues (Barshan et al. 2011). Thus, in this study, a supervised version of PCA is used. The supervised version of PCA selects the PCs (predictors) based on univariate regression coefficients. The major advantage of univariate regression coefficients is it provides a fixed sequence of predictors (Ahmed et al. 2015). Furthermore, the supervised version of PCA helps to avoid the effects of noisy features on the prediction model (Bair et al. 2006). In the present study, unsupervised PCA was performed on the GCM simulated precipitation data obtained from 99 grid points to identify the model predictors. The downscaling models were developed using the first few PCs representing more than 95% variability of selected predictors.

Random forest

RF is a nonparametric statistical regression algorithm. RF uses regression trees for the construction of randomness in decision forests. All the trees are constructed using sample data of various bootstraps. Classification and regression
algorithm (CART) is applied in RF schemes. The following is the RF procedure:

1. A number of sample sets are drawn randomly from the observed data using the bootstrap method. Every sample is a bootstrap strap sample, and the components that are excluded in the bootstrap sample are named ‘out-of-bag’ (OOB) for that bootstrap sample.
2. The Classification and Regression Tree (CART) technique is applied for every bootstrap sample, to develop an unpruned regression tree with a modification that a random subset of variables, from which the best variables are split, is chosen at each node.
3. Prediction of new sample data is made by aggregating the predictions of all the trees.
4. For a new data set having data which was not included in the previous bootstrap, the above procedure is repeated. After each iteration, the error is estimated, and the random forest tree with the least error is considered as the best model for prediction.

The ‘Random Forest’ package written in R program was employed in this study for developing an MOS based RF downscaling model. The model was developed separately for each calendar month using observed rainfall as predictand and GCM simulated historical rainfall obtained after performing unsupervised PCA as model predictors. The developed models were then used to project future rainfall.

Performance evaluation of downscaling models

The performance of the downscaling models was evaluated using the following commonly used statistical indices:

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |x_{\text{obs},i} - x_{\text{sim},i}|
\]

(1)

\[
\text{NRMSE} = \frac{\left[ n \sum_{i=1}^{n} (x_{\text{sim},i} - x_{\text{obs},i})^2 \right]^{\frac{1}{2}}}{x_{\text{obs},i}}
\]

(2)

\[
\text{PBIAS} = 100 \times \frac{\sum_{i=1}^{n} (x_{\text{sim},i} - x_{\text{obs},i})}{\sum_{i=1}^{n} x_{\text{obs},i}}
\]

(3)

\[
R^2 = \frac{\sum_{i=1}^{n} (x_{\text{obs},i} - \bar{x}_{\text{obs}})(x_{\text{sim},i} - \bar{x}_{\text{sim}})}{\sqrt{\sum_{i=1}^{n} (x_{\text{obs},i} - \bar{x}_{\text{obs}})^2 \sum_{i=1}^{n} (x_{\text{sim},i} - \bar{x}_{\text{sim}})^2}}
\]

(4)

\[
\text{NSE} = 1 - \frac{\sum_{i=1}^{n} (x_{\text{sim},i} - x_{\text{obs},i})^2}{\sum_{i=1}^{n} (x_{\text{obs},i} - \bar{x}_{\text{obs},i})^2}
\]

(5)

where \(x_{\text{sim},i}\) and \(x_{\text{obs},i}\) are the \(i\)th simulated and observed data, and \(n\) is the number of the observations.

RESULTS AND DISCUSSION

Development and validation of downscaling model

Downscaling models were constructed through the calibration and validation processes for the historical period of 1971–2005. The performance of RF models mainly depends upon the number of trees. In this study, the number of trees was kept as 500 based on out-of-bag (OOB) error. The ability of downscaling models to downscale rainfall were assessed using scatter plots and various statistical indices for all the GCMs used.

Reconstruction of observed rainfall

The observed rainfall for the period 1971–2005 was compared with the GCMs downscaled rainfall to assess the performance of the downscaling model. In Figure 3, a scatter plot is used to show the comparison of monthly observed and downscaled rainfall at a station located in Kedah for all the GCMs used in this study. Usually, in many cases, the downscaling models cannot capture the rainfall perfectly (Ahmed et al. 2015). However, in the present study, the accuracy of RF models was found promising in most of the cases. The higher rainfall values were slightly over- and under-estimated in some cases. However, in most of the cases, data was found near the bisector line indicating the better-predicting capability of RF. This indicates the capability of RF-based MOS downscaling models to downscale daily rainfall in Peninsular Malaysia.

Performance evaluation using statistical indices

The performance of the RF model was evaluated by assessing the magnitude of error using the statistical indices MAE, NRMSE, PBIAS, \(R^2\) and NSE. The results of these statistical indices for the GCMs used in the study at a station located in Kedah are presented in Table 2 as an example. From the table it can be seen that all the models performed well. For all models, the statistical indices values were very
close to each other. Almost the same type of results were obtained for all the stations used in the study. The results revealed a high accuracy of the models in downscaling rainfall at all the observed stations. Overall, the mean absolute error (MAE) of the models was found to be in the range of 0.8–0.39; normalized root mean square error (NRMSE) between 7.4 and 41.7, Percent Bias (PBIAS) between −0.3 and 10.1, Nash–Sutcliffe coefficient (NSE) between 0.81 and 0.99 and $R^2$ between 0.89 and 0.99. The significance of $R^2$ was also assessed in the study and showed that all values are significant at 95% level of confidence. Furthermore, the results of PBIAS and NSE were found to be between −0.3 and 10.1 and 0.81 and 0.99, respectively, which are considered as very good, as suggested by Moriasi et al. (2007). It can be concluded that the RF based MOS downscaling model is useful in downscaling and projecting rainfall. Therefore, this model was used for the projection of rainfall under climate change scenarios.

**Rainfall projection for RCP scenarios**

Using RF-based MOS downscaling the rainfall was projected for the period 2006–2099 under RCP4.5 and RCP8.5 for all the models used. The magnitude of rainfall
projected under both the RCPs varied from station to station and model to model. The annual projected rainfall at a station located in Kedah is shown in Figure 4. For assessing the change in rainfall due to the changing climate, the model simulated averaged annual rainfall (2006–2099) was compared with the average annual observed rainfall (station rainfall) (1971–2005).

The % variation of model-simulated annual average rainfall was calculated separately for all the models. For all the stations all eight GCMs projected an increase in the percent annual average projected rainfall under both the RCPs. The percent variation of GCMs simulated annual average rainfall from the annual average observed rainfall under RCP4.5 and RCP8.5 are presented in Table 3. In Figures 5 and 6, Box and Whisker plots have been used to show the percent variations by GCMs simulations (of all eight GCMs) for RCP4.5 and RCP8.5 respectively for all the rain gauged stations used.

**DISCUSSION**

In this work, variations in rainfall due to changing climate was assessed using the projection of CMIP5 GCMs under
RCP4.5 and RCP8.5. An increase in annual future rainfall was projected in Peninsular Malaysia by all the models under both the RCPs. Many other studies also reported an increase in rainfall in Peninsular Malaysia. MMD (2013) found that the spatial distribution of rainfall will increase by the end of this century in Peninsular Malaysia under A1B scenario. Shahid et al. (2017) also reported an increase in the variability of rainfall in Peninsular Malaysia by the end of this century. However, the magnitude of change varied for all models under both the RCPs. The overall increase in the annual rainfall under both the RCPs was in the range of 7.3–29.5%. The magnitude of percent increase in annual average rainfall for RCP4.5 was between 7.3 and 27.7%. The minimum increase was observed for model NorESM1-M for station Perak 4409091 and the maximum increase was observed for model MRI-CGCM3 at station Johor 1437116. Under RCP8.5 the overall increase for all the GCMs used was between 7.6 and 29.5%. The minimum increase was observed for model CSIRO-Mk3.6 at station Negeri Sembilan 2725083 and the maximum increase was observed for model CSIRO-Mk3.6 at station Kelantan 6019004. The maximum increase in average annual projected rainfall was observed in the northern part of Peninsular Malaysia which was in the range of 20.7–29.5%, while the minimum increase was observed in the south-west region (7.6–15.2%). For other parts of the country the percent increase in annual average rainfall observed was as follows: western region 7.9–18.9%, eastern region 13.3–25.7%, northern region 20.7–29.5%, southern region 11.18–27.7%, north-west region 11.7–20.9%. The percent increase under RCP8.5 was found to be slightly higher than RCP4.5. Lee et al. (2014) also found that the frequency and variability of heavy rainfall events under RCP8.5 are projected to be more than RCP4.5. The expected increase in annual rainfall due to changing climate urges the need for adaptive measures in the design of water management infrastructure and long-term planning. Also, it improves the knowledge regarding expected changes in rainfall pattern and its intensity in future return periods. Therefore, it is suggested that for designing an urban water management infrastructure, climate change should be considered in order to avoid huge economic and life losses due to expected increases in the intensity of rainfall and decreases in the return periods.

**CONCLUSIONS**

The main objective of the study was to assess the variations in rainfall for selected locations of Peninsular Malaysia under the changing climate scenarios. For this purpose, projections of eight CMIP5 GCMs under RCP4.5 and RCP8.5 were used. For assessing the variation in rainfalls, the average annual projected rainfall was compared with the

---

**Table 2 | Calibration and validation of the downscaling model for various GCMs at station Kedah 5806066**

<table>
<thead>
<tr>
<th>Indices</th>
<th>Model</th>
<th>MAE</th>
<th>NRMSE %</th>
<th>PBIAS %</th>
<th>NSE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration</td>
<td>BCC-CSM1.1</td>
<td>0.51</td>
<td>12.1</td>
<td>1.1</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>CCSM4</td>
<td>0.27</td>
<td>12</td>
<td>-0.6</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>CSIRO-Mk3.6.0</td>
<td>0.34</td>
<td>13.1</td>
<td>2.1</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>HadGEM2-ES</td>
<td>0.36</td>
<td>14.6</td>
<td>-0.3</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>IPSL-CM5A-MR</td>
<td>0.21</td>
<td>12.1</td>
<td>3.1</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>MIROC-ESM</td>
<td>0.34</td>
<td>12.9</td>
<td>1.3</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>MRI-CGCM3</td>
<td>0.51</td>
<td>2.1</td>
<td>5.5</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>NorESM1-M</td>
<td>0.15</td>
<td>12.9</td>
<td>1.3</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>Validation</td>
<td>BCC-CSM1.1</td>
<td>0.45</td>
<td>20.6</td>
<td>2.1</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>CCSM4</td>
<td>0.54</td>
<td>21.3</td>
<td>1.9</td>
<td>0.95</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>CSIRO-Mk3.6.0</td>
<td>0.42</td>
<td>21</td>
<td>-0.1</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>HadGEM2-ES</td>
<td>0.4</td>
<td>17.2</td>
<td>4.1</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>IPSL-CM5A-MR</td>
<td>0.42</td>
<td>19</td>
<td>1.3</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>MIROC-ESM</td>
<td>0.42</td>
<td>17.9</td>
<td>1.1</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>MRI-CGCM3</td>
<td>0.42</td>
<td>19.2</td>
<td>1</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>NorESM1-M</td>
<td>0.45</td>
<td>17.8</td>
<td>1.9</td>
<td>0.93</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Figure 4  |  Annual projected rainfalls for various GCMs under RCP4.5 and RCP8.5 at station Kedah S806066. (Continued.)
Figure 4  |  Continued.
that there would be an increase in the range of 7.3–29.5% in the annual average rainfall by the end of this century due to the effects of increasing temperature and changing climate. Under RCP4.5, the increase was in the range of 7.3–27.7%. Under RCP8.5, the increase was in the range of 7.6–29.5%. The maximum variation in average annual projected rainfall was observed in the northern part of Peninsular Malaysia which was in the range of 20.7–29.5%, while the minimum variation was observed in the south-west region in the range of 7.6–15.2%. In this study, eight CMIP5 GCMs were used for downscaling and projection of precipitation. IPCC data portal has several other GCMs, which can be used in future studies to project rainfall more efficiently. Furthermore, various approaches can be used to assemble these GCMs together for convenient projection of rainfall for various climate change scenarios. Two representative concentration pathways (RCPs), namely RCP4.5 and RCP8.5, were used in the current study. Projection of rainfall under RCP6.0 and RCP4.5 can be useful in predicting the future rainfall in a more efficient way. In this study the variation was computed only for selected locations of rainfall gauged stations, estimations of these variations for ungagged sites can help in mitigation expected losses due to changing climate.

ACKNOWLEDGEMENT

The authors would like to acknowledge Universiti Teknologi Malaysia (UTM) for providing financial support for this research through RUG Grant No. Q.J130000.2522.18H94.

REFERENCES


MMD 2009 Climate Change Scenarios for Malaysia 2001–2090. Malaysian Meteorological Department, Malaysia.


San Liew, Y., Teo, F. Y. & Ghani, A. A. 2014 Assessment of the climate change impact on a dry detention pond at Kota Damansara, Malaysia. In: 13th International Conference on Urban Drainage, Sarawak, Malaysia.


First received 21 February 2019; accepted in revised form 26 July 2019. Available online 9 September 2019