Long-term forecasting of climatic parameters using parametric and non-parametric stochastic modelling

Min Yan Chia[1], *Yuk Feng* Huang[1*], *Chai Hoon* Koo[1] and *Zheng Rong* Chong[1]

[1]Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Malaysia.

Abstract. Climatic parameters fluctuate dynamically and their turbulences become more significant as the influence of the climate change increases. A robust model that is able to factor in the recent climate change for long-term climatic parameters forecasting is desired to strategically plan for future anthropogenic activities. In this study, two stochastic time series model, namely the seasonal auto-regressive integrated moving average (SARIMA) model and the artificial neural network (ANN) model are used to predict monthly mean temperature (T_{mean}), relative humidity (RH), wind speed (u) and pan evaporation (E_{pan}) up to 12 months ahead. This study is conducted using data collected from three meteorological stations in the northern region of the Peninsular Malaysia. The stochastic models forecasted the T_{mean} with the highest accuracy, followed by RH, u and E_{pan}. Besides, despite the increasing time step (from 1 to 12 months), the accuracy of the models remain consistent. However, both of the models are susceptible to the occurrence of extreme climates. In general, the SARIMA model performs better than the ANN model, probably attributed to its ability to consider the seasonality of the climatic data rather than depending solely on black-box computation.

1 Introduction

The effect of climate change is becoming increasingly prominent over the last few decades. The consequences of climate change are witnessed in many aspects of the natural systems, including the weather, agriculture, ecosystem and hydrology [1]. The Intergovernmental Panel on Climate Change (IPCC) had observed that from the year 2002 to 2017, the global mean temperature experienced an increase of 0.5 °C, with the increase in land surface temperature exceeded this mean value by another 0.5 °C [2]. In the same report, the IPCC stressed that the degree of climate change is highly correlated with the anthropogenic activities through the emission of greenhouse gases (GHG) as well as land use. It is now clear that in the near future, the human activities have to be planned carefully and strategically as an effort to curb the rate of climate change. Numerous models and simulations known as global climate models (GCMs) have been proposed to project the future climate as a cautious

* Corresponding author: huangyf@utar.edu.my

note for the aggressive development. These models are widely used in forecasting dry spell, precipitation and temperature [3-5].

Although the GCMs have wide spatial applications, experts have claimed that there still exist inevitable uncertainties when performing such modelling or simulation works. These uncertainties arise from different sources, including the downscaling, natural variability and the model itself [6]. Besides, the use of the GCMs requires the users to arbitrarily assume the representative concentration pathway (RCP) which illustrates the trajectory of greenhouse gases up to year 2100 [7]. While the prediction of the true pathway is not possible, simulating GCMs over all the pathways could be computationally expensive and impractical, not to mention the variety of GCMs that have distinct performances under different conditions. A robust and simple approach is needed.

In this study, parametric and non-parametric stochastic models are used to perform long-term forecasting of climatic parameters in the Peninsular Malaysia. The used of stochastic models do not require the input of RCP and merely produce future projections based on a historical time series. The auto-regressive integrated moving average (ARIMA) is a traditional time series model that combines regression analysis with the moving average. It includes a differencing term to transform the non-stationary time series into a stationary time series model before performing subsequent analysis [8]. The parametric ARIMA model weights the historical time series differently, whereby newer data are given higher weightage. Nonetheless, the seasonal ARIMA (SARIMA) is claimed to be having higher efficiency in forecasting climatic parameters due to its nature that is able to deal with seasonal fluctuations [9]. The SARIMA model has proved its suitability in many applications, including temperature and wind speed [10, 11].

On contrary, the representative for non-parametric stochastic model in the artificial neural network (ANN). The ANN is classified as a non-parametric model due to black-box operation that gives random weightage to the historical data. The non-linearity in ANN allows it to better adapt to complex processes and problems such as the climatic parameters [12]. Time series modelling involving climatic processes using ANN is common [13]. The originality of this research work is to compare the performances of parametric (represented by SARIMA) and non-parametric (represented by ANN) stochastic modelling in forecasting multiple climatic parameters using univariate time series data in a region with tropical climate. The output of this research work could provide a robust and simple forecasting strategy for the decision makers that eliminate the need of simulating the complex and uncertain GCMs.

2 Methods

2.1 Study Area and Data

Three meteorological stations in the northern region of the Peninsular Malaysia are selected to be included in this study. The three stations share similar characteristic whereby all of them are located at coastal areas. The three stations are Station 48600 (Pulau Langkawi), Station 48601 (Bayan Lepas) and Station 48615 (Kota Bharu). The details of the stations are provided in Table 1, whereas the exact locations of the stations are shown in Fig. 1. Four types of monthly climatic data, including the mean temperature (T_{mean}), relative humidity (RH), wind speed (u) and pan evaporation (E_{pan}) are obtained from the Malaysia Meteorological Department for the period of year 2002 to 2017. In order to assess the long-term forecasting ability of the SARIMA and ANN models, data from the year 2002 to 2016 are used for training and modelling, whereas the data in the year 2017 are used for comparison with the forecasted climatic data. In other words, the SARIMA and ANN models developed

in this study should be able to predict values of climatic parameters 12 months ahead of current time.

Table 1. Details of meteorological stations.

Station ID	Station Name	Latitude (°N)	Longitude (°E)	Elevation (m)
48600	Pulau Langkawi	6.33	99.73	6.4
48601	Bayan Lepas	5.13	100.27	2.5
48615	Kota Bharu	6.17	102.30	4.4

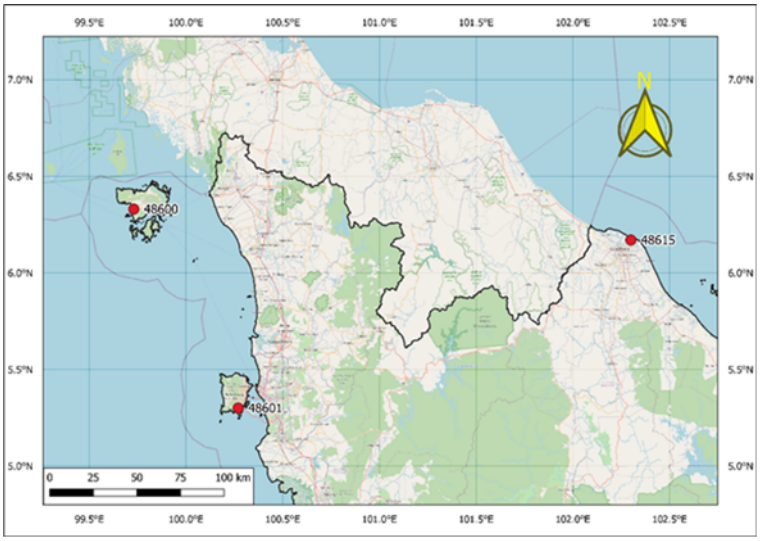

Fig. 1. Locations of the studied meteorological station.

2.2 Seasonal Auto-Regressive Integrated Moving Average

The parameters of the SARIMA model are tuned by using the trial and error method. In general, six parameters have to be decided for a SARIMA model to operate. The design of SARIMA model is split into two parts, where the lag order, differencing order as well as moving average order of the seasonal component (P,D,Q) and ARIMA component (p,d,q) have to be selected. Moreover, an intermittent time (ω) needs to be specified for the SARIMA model. The parameters of the ARIMA component is determined by plotting the auto-correlation function (ACF) and partial ACF (PACF). Whereas, the parameters of the seasonal component is tuned by minimising the Akaike Information Criterion (AIC) of the SARIMA model. In this study, the SARIMA model will be referred in the form of SARIMA(p,d,q)(P,D,Q)$_\omega$. The SARIMA model can be represented mathematically by Equation 1.

$$\Phi_P(B^\omega)\phi_p(B)\nabla_\omega^D\nabla^d X_t = \theta_q(B)\Theta_Q(B^\omega)\varepsilon_t \quad (1)$$

where X_t is the stochastic climatic parameter, ε_t is the normal random variable, B is the regressive operator, Φ is the seasonal autoregressive operator, ϕ is the non-seasonal auto-regressive operator, ∇_ω^D is the seasonal differencing operator, ∇^d is the non-seasonal

differencing operator, Θ is the seasonal moving average operator and θ is the non-seasonal moving average operator.

2.3 Artificial Neural Network

The non-parametric ANN, in the form of multilayer perceptron (MLP) works on the principle that different weightage is given to every input into the network. The input is then transformed using tangent-sigmoid activation function to determine the strength of the output of the hidden neurons. The final output of the ANN is the summation of the values output by the hidden neurons. In this study, the ANN is trained using the Levenberg-Marquardt algorithm and the optimum number of hidden neurons is determined by the grid search method with the minimisation of the prediction error. Time lags of six months is used to ensure the meaningful time series is fed into the model. The mathematical expression of the ANN is shown in Equation 2.

$$y = f(\sum_{i=1}^{n} w_i x_i + b) \tag{2}$$

where f is the activation function, w is the weight term, b is the bias term, n is the number of input, x is the input and y is the output. The general structure of the developed ANN models can be referred to other literatures [12].

2.4 Performance Evaluation

In order to compare the performances of the SARIMA and the ANN model, three performance evaluation metrics are used. The mean absolute percentage error (MAPE) measures the degree of deviation of the predicted values from the actual values. The root mean square error (RMSE) is used to detect the presence of large errors. The coefficient of determination (R^2) evaluates the goodness of fit of the model as well as the scattering of the predicted values from the actual values. The equations used to calculate MAPE, RMSE and R^2 are shown in Equation 3, 4 and 5, respectively.

$$MAPE = \frac{1}{N}\frac{\sum_{i=1}^{N}|y_p - y_a|}{y_m} \times 100\ \% \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(y_p - y_a)^2} \tag{4}$$

$$R^2 = 1 - \frac{(y_p - y_a)^2}{(y_a - \bar{y})^2} \tag{5}$$

where N is the number of observations, y_p is the predicted value, y_a is the actual value and $\bar{y}$ is the mean of the actual values.

3 Results and Discussion

3.1 Tuning of SARIMA Models

The parameters of the SARIMA models are determined for different climatic parameters at different meteorological stations. The tuned SARIMA models and their performances are compiled in Table 2. As shown in Table 2, most of the parameters for ARIMA and seasonal components are (2,0,0) and (0,1,1), respectively. In other words, the time series of the climatic parameters at different stations is auto-regressive with respect to the second order

(only requires two data points backwards). There are two exception alternatives for the parameters of the ARIMA component, namely (2,0,2) and (2,1,2). The former is used for *u* at Station 48601 (Bayan Lepas) whereas the latter is used for *RH* at Station 48601 (Bayan Lepas) and T_{mean} at Station 48615 (Kota Bharu). This means that some sort differencing is required for the mentioned climatic parameters in order to ensure stationarity in the time series.

When the time series trends of the climatic parameters are analysed with a seasonal return period of 12-time steps, seasonal component of (0,1,1) is obtained. At all the studied stations, the seasonal trends of the climatic parameters require first order differencing to make the trends are stationary. On the other hand, the seasonal trends are not auto-regressive except for the T_{mean} at Station 48601 (Bayan Lepas). By combining the seasonal and ARIMA components, it is discovered that none of the time series of the investigated climatic parameters is stationary and all of them exhibit seasonality. The analysed time series can be used for the forecasting purpose.

Referring Table 2, it is observed that the SARIMA model forecasted the T_{mean} of the year 2017 with the highest accuracy in terms of MAPE and RMSE, followed by *RH, u* and E_{pan}. This finding is reasonable as the temperature fluctuation is a simple process that can be easily predicted by univariate analysis. On contrary, moving from *RH* to *u* and then to E_{pan}, the processes become increasingly complex which involve other environmental conditions. For instance, the magnitude of E_{pan} is dictated by multiple factors such as temperature, humidity as well as surface moisture.

Table 2. Performance of tuned SARIMA models for climatic parameters at different stations.

Station	Climatic Parameter	SARIMA Model	MAPE (%)	RMSE	R^2
Station 48600 (Pulau Langkawi)	T_{mean} (°C)	SARIMA(2,0,0)$(0,1,1)_{12}$	1.62	0.52	0.76
	RH (%)	SARIMA(2,0,0)$(0,1,1)_{12}$	3.50	3.53	0.51
	u (m/s)	SARIMA(2,0,0)$(0,1,1)_{12}$	7.44	0.20	0.80
	E_{pan} (mm/day)	SARIMA(2,0,2)$(0,1,1)_{12}$	12.39	0.66	0.29
Station 48601 (Bayan Lepas)	T_{mean} (°C)	SARIMA(2,0,0)$(0,1,1)_{12}$	1.67	0.52	0.20
	RH (%)	SARIMA(2,1,2)$(0,1,1)_{12}$	3.05	3.07	0.26
	u (m/s)	SARIMA(2,0,2)$(0,1,1)_{12}$	9.87	0.22	0.52
	E_{pan} (mm/day)	SARIMA(2,0,0)$(0,1,1)_{12}$	13.91	0.73	0.09
Station 48615 (Kota Bharu)	T_{mean} (°C)	SARIMA(2,1,2)$(0,1,1)_{12}$	2.25	0.67	0.88
	RH (%)	SARIMA(2,0,0)$(0,1,1)_{12}$	3.89	3.66	0.71
	u (m/s)	SARIMA(2,0,0)$(0,1,1)_{12}$	29.36	0.81	0.50
	E_{pan} (mm/day)	SARIMA(2,0,0)$(0,1,1)_{12}$	8.72	0.37	0.75

3.2 Optimum ANN Models

The performances of the ANN models in forecasting climatic parameters are shown in Table 3. The forecasting plot for the ANN models are shown in Figure 3. For different climatic

parameters at different stations, the optimum number of hidden neurons varies, but generally, minimum number of hidden neurons is five. Similar to the SARIMA models, the ANN models could forecast T_{mean} more accurately, with *RH*, *u* and the E_{pan} follow suit. However, as compared to the SARIMA models, the accuracy of the ANN models is slightly lower as shown in the MAPE and RMSE values. There are some results with high errors with high values of R^2. This means that the forecasting time series is less accurate as compared to the mean value, and naively using the mean value as the forecasted value is better.

The forecast time series of the SARIMA and the ANN models at Station 48600 (Pulau Langkawi) are compared in Fig. 2. As shown by the figure, it can be seen that both models are unable to adapt themselves towards extreme values. When the values of the climatic parameter is extremely high or extremely low, the models would tend to underestimate and overestimate, respectively. Besides, the distances between the forecast plots and the actual plots of the SARIMA models are smaller than that of the ANN models, suggesting that the SARIMA models should be the more suitable approach for forecasting climatic parameters in the Peninsular Malaysia with tropical climate.

Moreover, the time series plots show that the accuracies of the model predictions are not affected by the length of the time window. That is to say, regardless of the one month or twelve months ahead, the accuracies of the models are similar. This means that theoretically, the SARIMA and the ANN models can forecast the climatic parameters for prediction horizon of more than twelve months. With consistent update of the model structure, the SARIMA and the ANN models can be deemed as robust in terms of the temporal context.

Table 3. Performance of optimum ANN models for climatic parameters at different stations.

Station	Climatic Parameter	Number of Hidden Neurons	MAPE (%)	RMSE	R^2
Station 48600 (Pulau Langkawi)	T_{mean} (°C)	7	1.93	0.62	0.01
	RH (%)	5	4.06	4.09	0.41
	u (m/s)	8	10.77	0.30	0.71
	E_{pan} (mm/day)	6	8.05	0.46	0.11
Station 48601 (Bayan Lepas)	T_{mean} (°C)	9	1.76	0.58	0.28
	RH (%)	8	3.63	3.67	0.01
	u (m/s)	9	12.59	0.31	0.11
	E_{pan} (mm/day)	8	18.29	0.92	0.14
Station 48615 (Kota Bharu)	T_{mean} (°C)	5	1.34	0.41	0.76
	RH (%)	6	2.78	2.12	0.46
	u (m/s)	5	8.07	0.41	0.48
	E_{pan} (mm/day)	9	12.32	0.56	0.59

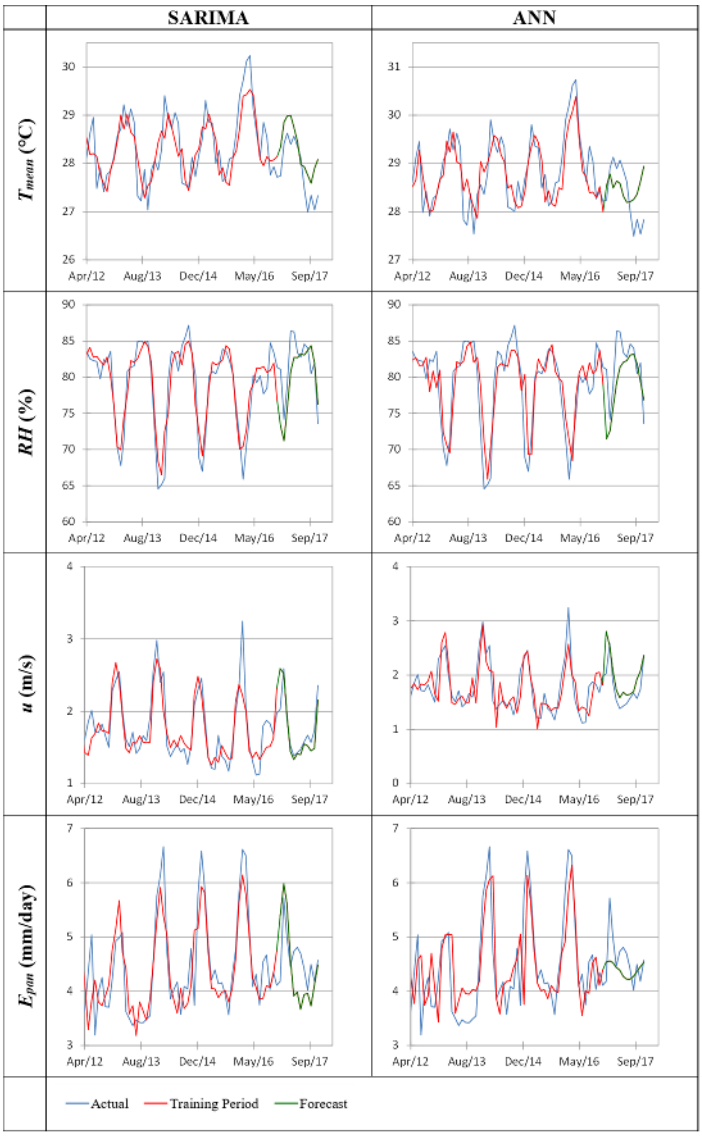

Fig. 2. Time series plot for the SARIMA and the ANN models at Station 48600 (Pulau Langkawi).

4 Conclusion

The SARIMA and ANN models are used to forecast T_{mean}, RH, u and E_{pan} for the northern region of Peninsular Malaysia up to 12 months ahead. It is concluded that the SARIMA models is more suitable for this task due to the lower MAPE and RMSE achieved. This is attributed to the nature of SARIMA model in considering the non-stationarity and seasonal trend of the climatic parameters. Both SARIMA and ANN models could predict T_{mean} with the highest accuracy, followed by RH, u and E_{pan} due to the difference in the complexity of the processes. However, the SARIMA and ANN models could not predict extreme values accurately and thus need more study for the mitigation of this issue. The prediction of the models are not affected by the size of the forecasting window, suggesting that the models can forecast the climatic parameters for longer time ahead. Although the performance of the non-

parametric ANN lags slightly behind the parametric SARIMA model, the authors strongly believe that the performance of ANN can be boosted through various hybridisation techniques to optimise the model.

Acknowledgement: This research was funded by Universiti Tunku Abdul Rahman (UTAR), Malaysia through the Universiti Tunku Abdul Rahman Research Fund under project number IPSR/RMC/UTARRF/2018-C2/K03.

References

1. Y. Dinpashoh, S. Jahanbakhsh-Asl, A. A. Rasouli, M. Foroughi, V. P. Singh, Theor. Appl. Climatol, **136** (1-2), pp. 185-201 (2019)
2. Intergovernmental Panel on Climate Change, Global warming of 1.5°C (2019)
3. K. Sreelatha, P. AnandRaj, ISH J Hydraul Eng, **28**(1) pp. 1-14 (2020)
4. H. Moon, L. Gudmundsson, S. I. Seneviratne, J Geophys Res Atmos, **123** (7), pp. 3483-3496 (2018)
5. L. B. Díaz, R. I. Saurral, C. S. Vera, Int J Climatol, **41**(S1), pp. 1-19 (2020)
6. A. Wootten, A. Terando, B. J. Reich, R. P. Boyles, F. Semazzi, J Appl Meteorol Climatol, **56**(12), pp. 3245-62 (2017)
7. R. Nakamura, T. Shibayama, M. Esteban, T. Iwamoto, S. Nishizaki, Coast. Eng. J, **62**(1), pp. 101-27 (2020)
8. Y. Zhang, H. Yang, H. Cui, Q. Chen, Nat. Resour. Res, **29**(2), pp. 1447-1464 (2019)
9. M. Farsi, D. Hosahalli, B. R. Manjunatha, I. Gad, E-S. Atlam, A. Ahmed, Alex. Eng. J, **60**(1), pp. 1299-316 (2021)
10. P, Aghelpour, B. Mohammadi, S. M. Biazar, Theor. Appl. Climatol, **138**(3-4), pp. 1471-80 (2019)
11. D. B. Alencar, C. M. Affonso, R. C. L. Oliveira, J. C. R. Filho, IEEE Access, **6**, pp. 55986-94 (2018)
12. M. Y. Chia, Y. F. Huang, C. H. Koo, K. F. Fung, Agronomy, **10**(1), pp. 101 (2020)
13. G. Papacharalampous, H. Tyralis, D. Koutsoyiannis, Water Resour. Manag, **32**(15), pp. 5207-5239 (2018)