11th MALAYSIA STATISTICS CONFERENCE 2024
Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara
19th September 2024

RAINFALL PREDICTION USING HYBRID BAYESIAN METHOD AND STATISTICAL DOWNSCALING BASED MACHINE LEARNING IN SELANGOR.

Noor Hamizah Mohamad Sani[1]; Shazlyn Milleana Shaharudin[2]; Muhammad Safwan Ibrahim[3]

[1] Universiti Pendidikan Sultan Idris, Tanjong Malim 35900, Perak, Malaysia; m2022100188@siswa.upsi.edu.my
[2] Universiti Pendidikan Sultan Idris, Tanjong Malim 35900, Perak, Malaysia; shazlyn@fsmt.upsi.edu.my
[3] Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan; msafwan@usim.edu.my

Abstract:

Rainfall projection is a critical aspect of weather forecasting and climate modelling in Selangor, Malaysia. This paper proposed the application of hybrid Bayesian method and statistical downscaling based-machine learning for rainfall prediction. Daily rainfall data (predictand) and atmospheric data (predictors) from 2008 to 2018 are the data used in this research involving 33 stations in Selangor. This study applied Principal Component Analysis (PCA) to capture high-dimensional data and select predictors. Then, Non-Homogeneous Markov Model (NHMM) and Random Forest (RF) to capture zero-inflated data and non-linearity data. The performance of the proposed models was analysed by using four performance metrics namely Root Mean Square Error (RMSE), Nash-Sutcliffe Efficiency (NSE), Mean Absolute Error (MAE) and Mean Forecast Error (MFE). The analysis of PCA indicates that five selected Principal Component's cut-off with eigenvalues at 0.535 and cumulative percentage of the total variance at 93.254%. Next, the analysis of NHMM indicates that three hidden state layers (K = 3) achieved perfect convergence at iteration 2000 with Bayesian Information Criterion (BIC) at 260,018.30, marking the optimal configuration for the model. The conclusion of this study is statistical downscaling-based RF-NHMM model consistently outperforms because it provides the most accurate and reliable rainfall predictions with minimal bias, as indicated by its superior performance in RMSE (2.298), NSE (0.752), and MAE (1.900), and its near-zero MFE (-0.049). The goal of this study is to enhance the accuracy of climate change projections by establishing the relationship between predictand variables and predictors using a hybridization approach.

Keywords:
Climate change; Homogeneous Markov Model; Random Forest; machine learning model

1. Introduction:

Rainfall is a crucial meteorological factor impacting human daily life and civilization development, as noted by Barrera-Animas et al. (2022) and Kanani et al. (2023). Altered rainfall patterns can lead to severe natural disasters such as floods, which pose significant risks to human safety and infrastructure (Usman et al., 2023). With climate change affecting the hydrological cycle, extreme weather events like floods and tsunamis are becoming more common. This underscores the need for accurate rainfall prediction models to mitigate risks and improve disaster preparedness by providing early warnings for extreme weather events (Prottasha et al., 2023). As climate change leads to increased variability in weather patterns, enhancing predictive modeling techniques has become increasingly critical (Sani et al., 2020).

Rainfall prediction models often need to handle large amounts of high-dimensional atmospheric data, which can complicate knowledge discovery and pattern classification due to redundant or irrelevant features (Zebari et al., 2020). Dimensionality reduction methods, such as Principal Component Analysis (PCA), are employed to simplify high-dimensional data without losing critical predictive features (Hasan & Abdulazeez, 2021). PCA is cost-efficient and effective for managing large datasets, reducing computational complexity and aiding in the accurate processing of atmospheric data (Abdulhammed et al., 2019).

Many datasets contain a significant number of zeros, which can lead to poor estimation and missed statistically significant findings if ignored (Gramosa et al., 2019). Specialized models that account for zero-bounded data are essential for realistic predictions and managing parameter uncertainties. Bayesian methods, which capture parameter uncertainties accurately, are particularly valuable in complex modeling scenarios like rainfall prediction (Cao et al., 2023). These methods enable robust decision-making by considering a range of potential outcomes and their probabilities (Bharadiya, 2023).

Non-linearity in rainfall data poses challenges for statistical models that assume linear relationships, particularly in statistical downscaling for climate projections (Zhang et al., 2023). To address this, machine learning-based statistical downscaling methods can handle complex, non-linear interactions more effectively (Putri et al., 2021). By integrating Bayesian approaches with machine learning techniques, these models can better capture non-linear relationships and enhance prediction accuracy. This combined approach addresses various modeling challenges, offering a sophisticated solution for improved rainfall prediction amidst climate variability. Despite advancements, integrating Bayesian methods with machine learning for rainfall prediction remains underexplored. This study aims to bridge this gap by developing a framework that combines these approaches, potentially advancing predictive accuracy and reliability for better decision-making in agriculture, water management, and disaster preparedness.

2. Methodology:

2.1 The Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is a statistical method used to model systems that transition between a set of hidden (unobserved) states over time. In HMM, the system is assumed to be a Markov process, meaning the future state depends only on the current state and not on the sequence of events that preceded it. Each hidden state generates observable outputs, and these observations provide indirect evidence about the underlying state. HMMs are widely used in time series analysis, speech recognition, and bioinformatics, where they help to uncover patterns and predict sequences of states based on observed data. The key components of an HMM include states, transition probabilities between states, emission probabilities (linking hidden states to observable outcomes), and initial state probabilities.

2.2 Random Forest

Random Forest (RF) by Breiman (2001) is a versatile machine learning technique applicable to both regression and classification tasks. In the RF approach, each decision tree is built using a bootstrap sample from the training set, meaning a random subset of the training data is used for growing each tree (Benmakhlouf et. al., 2023). To further prevent overfitting and enhance stability, only a subset of the features is considered at each node when determining the optimal split. This randomization process helps to decrease the correlation among individual trees, leading to a more diverse and robust ensemble.

3. Result:

Regression analysis is crucial for predictive modeling, such as forecasting rainfall, where machine learning-based statistical downscaling can enhance predictions (Mohammed et. al., 2020). The Random Forest (RF) method, utilizing 500 decision trees, was applied to forecast rainfall trends in Selangor. The data was split 70:30 for training and testing, as suggested by Patel et al. (2023) for optimal results. The RF model's predictions, visualized in a scatter plot (Figure 3.1), reveal clusters of lower and higher rainfall values, indicating potential underfitting and an inability to capture mid-range values. Performance metrics (Table 3.1) showed moderate accuracy, with RMSE and MAE indicating average error magnitudes and NSE reflecting decent predictive capability. Figure 4.5 displays the model's performance in predicting rainfall amounts over several days, with blue bars representing actual rainfall and orange bars showing predicted values. While the model captures the overall trend, there are noticeable discrepancies, especially during higher rainfall periods. These results confirm that the model approximates the general rainfall pattern but struggles with precise predictions, particularly underestimating higher rainfall amounts, as indicated by the smoother predicted values in the graph.

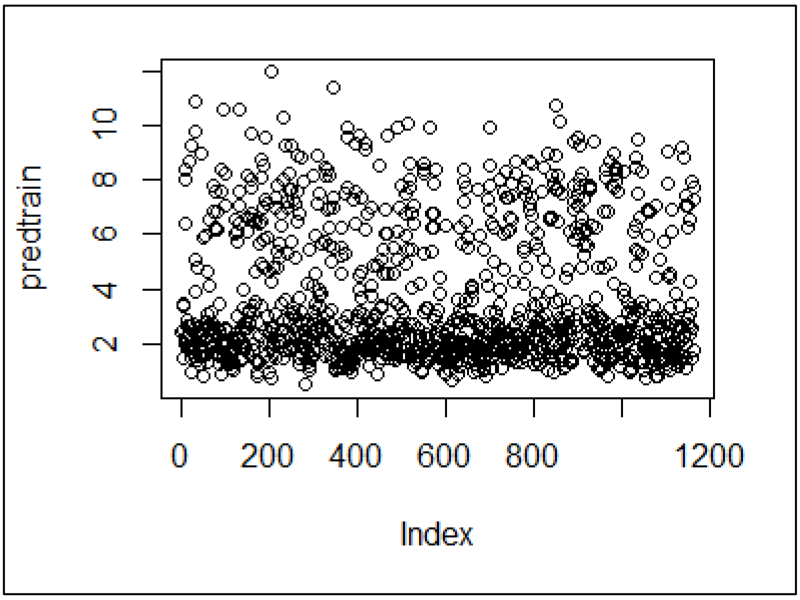

Figure 3.1. The scatter plot of calibration of predicted rainfall values

Calibration				Validation			
RMSE	**NSE**	**MAE**	**MFE**	**RMSE**	**NSE**	**MAE**	**MFE**
2.298	0.752	1.900	-0.049	4.937	0.006	3.851	0.867

Table 3.1. The performance of Statistical downscaling-based RF-NHMM model

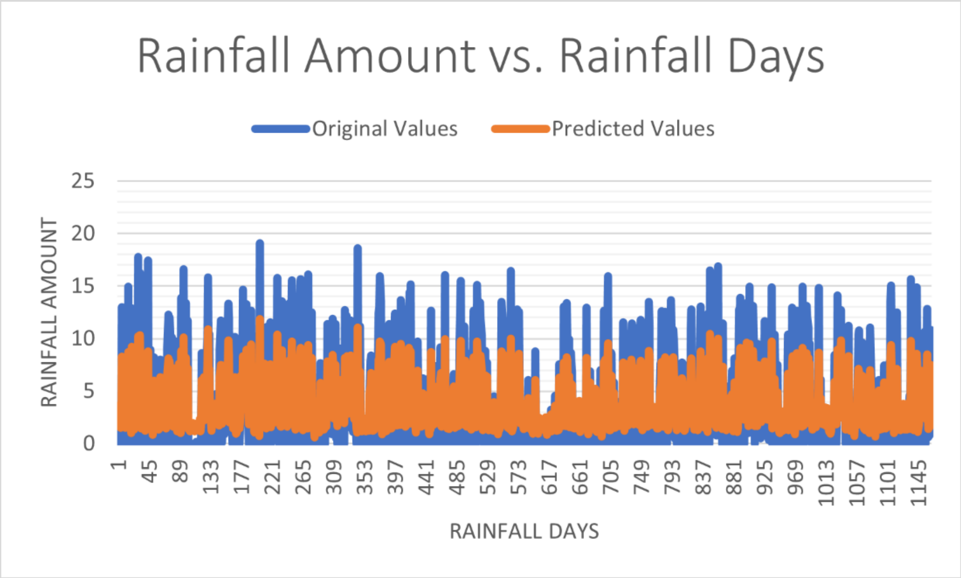

Figure 3.2. The performance of statistical downscaling-based RF-NHMM model in predicting rainfall amounts in calibration period

4. Discussion and Conclusion:

The main proposed of this study is to project daily rainfall data using hybrid NHMM and machine learning-based statistical downscaling in Selangor, Malaysia. The proposed model is statistical downscaling-based RF-NHMM model. Statistical downscaling-based RF-NHMM model performs the best across the majority of the metrics. It has a low RMSE and MAE, and a high NSE, indicating it makes the most accurate and reliable predictions with minimal bias. Statistical downscaling-based RF-NHMM model has demonstrated its capability to predict extreme values with high accuracy and reliability while maintaining minimal bias. This makes it a valuable tool for hydrologists and climatologists in analyzing environmental models and enhancing climate change assessments.

References:

Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., & Abuzneid, A. (2019). Features dimensionality reduction approaches for machine learning based network intrusion detection. *Electronics*, *8*(3), 322.

Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Machine Learning with Applications*, *7*, 100204.

Benmakhlouf, M., El Kharim, Y., Galindo Zaldívar, J., & Sahrane, R. (2023). Landslide susceptibility assessment in western external rif chain using machine learning methods.

Bharadiya, J. P. (2023). A review of Bayesian machine learning principles, methods, and applications. *International Journal of Innovative Science and Research Technology*, *8*(5), 2033-2038.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Cao, Q., Zhang, H., Lall, U., Holsclaw, T., & Shao, Q. (2024). The predictability of daily rainfall during rainy season over East Asia by a Bayesian nonhomogeneous hidden Markov model. *Journal of Flood Risk Management*, *17*(1), e12942.

Putri, C. D., Farikha, E. F., Hadi, A. F., Dewi, Y. S., Tirta, I. M., Ubaidillah, F., & Anggraeni, D. (2022, February). Projection Pursuit Regression on Statistical Downscaling Using Artificial Neural Network and Support Vector Regression for Rainfall Forecasting in Jember. In *International Conference on Mathematics, Geometry, Statistics, and Computation (IC-MaGeStiC 2021)* (pp. 204-210). Atlantis Press.

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, *2*(1), 20-30.

Kanani, S., Patel, S., Gupta, R. K., Jain, A., & Lin, J. C. W. (2023). An AI-enabled ensemble method for rainfall forecasting using long-short term memory.

Mohammed, M., Kolapalli, R., Golla, N., & Maturi, S. S. (2020). Prediction of rainfall using machine learning techniques. *International Journal of Scientific and Technology Research*, *9*(1), 3236-3240.

Patel, A., Keriwala, N., Soni, N., Goel, U., Bhoj, R., Adhyaru, Y., & Yadav, S. M. (2023). Rainfall Prediction using Machine Learning Techniques for Sabarmati River Basin, Gujarat, India. *Journal of Engineering Science & Technology Review*, *16*(1).

Prottasha, N. J., Tahabilder, A., Kowsher, M., Mia, M. S., & Kobra, K. T. (2023). Short-Term Rainfall Prediction Using Supervised Machine Learning. *Advance in Technology Innovation,* 111-120, 8(2).

Quadros Gramosa, A. H., Ferraz do Nascimento, F., & Castro Morales, F. E. (2020). A Bayesian approach to zero-inflated data in extremes. *Communications in Statistics-Theory and Methods*, *49*(17), 4150-4161.

Sani, N. S., Abd Rahman, A. H., Adam, A., Shlash, I., & Aliff, M. (2020). Ensemble learning for rainfall prediction. *International Journal of Advanced Computer Science and Applications*, *11*(11).

Usman, C. D., Widodo, A. P., Adi, K., & Gernowo, R. (2023). Rainfall prediction model in Semarang City using machine learning. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(2), 1224-1231.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(1), 56-70.

Zhang, X., Yin, Q., Liu, F., Li, H., & Qi, Y. (2023). Comparative study of rainfall prediction based on different decomposition methods of VMD. *Scientific Reports*, *13*(1), 20127.