Available online at https://journal.rescollacomm.com/index.php/ijqrm/index

International Journal of Quantitative Research and Modeling

e-ISSN 2721-477X
p-ISSN 2722-5046

Vol. 2, No. 3, pp. 139-146, 2021

Principal Component Regression in Statistical Downscaling with Missing Value for Daily Rainfall Forecasting

M Dika Saputra[1], Alfian Futuhul Hadi[2,*], Abduh Riski[2], Dian Anggraeni[2]

[1]*Master of Mathematics Department, Faculty of Mathematics and Natural Science, University of Jember, Indonesia*
[2]*Data Science Research Group, Department of Mathematics, University of Jember, Indonesia*

**Corresponding author email: afhadi@unej.ac.id*

Abstract

Drought is a serious problem that often arises during the dry season. Hydrometeorologically, drought is caused by reduced rainfall in a certain period. Therefore, it is necessary to take the latest actions that can overcome this problem. This research aims to predict the potential for a drought to occur again in the Kupang City, Indonesia by developing a rainfall forecasting model. Incomplete daily local climate data for Kupang City is an obstacle in this analysis of rainfall forecasting. Data correction was then carried out through imputed missing values using the Kalman Filter method with Arima State-Space model. The Kalman Filter and Arima State-Space model (2,1,1) produces the best missing data imputation with a Root Mean Square Error (RMSE) of 0.930. The rainfall forecasting process is carried out using Statistical Downscaling with the Principal Component Regression (PCR) model that considers global atmospheric circulation from the Global Circular Model (GCM). The results showed that the PCR model obtained was quite good with a Mean Absolute Percent Error (MAPE) value of 2.81%. This model is used to predict the daily rainfall of Kupang City by utilizing GCM data.

Keywords: Principal Component Regression, Statistical Downscaling, Missing Value, Rainfall Forecasting, Global Circular Model.

1. Introduction

Drought disaster in Indonesia is a problem that has a significant impact. As an agricultural country, drought can cause a decrease in food crop production which has an impact on decreasing the amount of national food and causing disruption of economic stability. Drought can be interpreted as a reduction in inventory water or moisture that is temporarily significantly below the normal or expected volume for a specified period of time (Murisidi and Sari, 2017; Field et al., 2004; Rozaki et al., 2021). The occurrence of natural disasters is also influenced by geographical conditions (Boesday et al., 2020; Titu-Eki and Kotta, 2021), one of which is the Kupang City. The rainy and dry seasons in Kupang City are closely related to the monsoon pattern so that it has an impact on reducing the formation of rain clouds in Kupang City.

Therefore, a number of handling efforts are needed to overcome the drought disaster in Kupang City. The latest effort that can be done is to develop a rainfall forecasting model. Accurate and precise rainfall forecasting models are an important part of providing rainfall information in the future. Based on previous research conducted by Estiningtyas and Wigena, it was explained that the best model of Principal Component Regression (PCR) was able to predict rainfall in El Nino conditions with an average RMSEP value of 95.22 and a correlation of 0.66 (Estiningtyas and Wigena, 2011). In addition, the several GCM models can approach the average monthly rainfall value with the largest correlation value of 0.497 at Bondan Station (Susandi et al., 2015). Based on the explanation above, it can be concluded that PCR and GCM models often have high accuracy values and low error values.

GCM data is a mathematical description of a large number of interactions of physics, chemistry, and dynamics of the Earth's atmosphere so that it produces a very large amount of data that can be used to make climate forecasts (Farikha et al., 2021). The resolution of the GCM data is too low to predict the local climate, so an alternative is needed in the use of GCM data, one of which is statistical downscaling (SD). Statistical Downscaling is a statistical model to describe the relationship between data on global-scale units and data on local-scale units within a certain time period (Sachindra et al., 2018). However, this application has some obstacles. One of the obstacles that occur is

the loss of local climate information from data provided by the Meteorology, Climatology and Geophysics Agency (BMKG).

The main objective of this study is; (1) how to overcome missing values in local climate data, (2) how to model statistical downscaling using the PCR method in daily rainfall forecasting, and (3) how the results of daily rainfall forecasting are for policy making in dealing with drought disasters.

2. Methodology

2.1. Data Description

In this study, two data were used, namely Global Circular Model (GCM) data and daily rainfall data in Kupang City. Both data were obtained from online publications issued by the BMKG through the website http://www.dataonlinebmkg.go.id. for daily rainfall data in Kupang City and GCM data obtained from the website https://cds.climate.copernicus.eu. Each data was obtained in the period January 1, 2019 to December 31, 2019. In the GCM data collection there are regional boundaries, it is latitude range of -13.25° SL to -7.25° SL and longitude range of 120.75° EL to 126.75° EL.

2.2. ARIMA(*p*, *d*, *q*) and Kalman Filter Imputation

The model state-space provides flexibility in extracting features from time series data. This model is generally used for the purpose of prediction, smoothing and likelihood assessment. The model state-space also provides a suitable framework for incorporating smoothing functions in various time series models to improve general predictions. In general, the model is state-space shown by equations (1) and (2)

$$y_t = H_t^T Y_t + \varepsilon_t \quad (1)$$
$$Y_t = Z_t Y_{t-1} + R_t \omega_t \quad (2)$$

where, for a state vector of length k, H_t is a vector of length k, ε_t $\tilde{\text{N}}\text{ID}\left(0, \sigma_{\varepsilon_t}^2\right)$, Z_t is a $k \times k$ matrix, R_t is a $k \times 1$ matrix dan $\omega_t \sim \text{NID}(0, W)$. Considering Equation (1), let $m = \max(p + d, q + 1)$. Then :

$$y_t = \theta_t y_{t-1} + \cdots + \theta_m y_{t-m} + a_t - \phi_1 a_{t-1} - \cdots - \phi_{m-1} a_{t-m+1} \quad (3)$$

Also, for $j < m$ and $\phi_0 = -1$, let:

$$\eta_t^{(m)} = \theta_t y_{t-1} - \phi_{m-1} a_t \quad (4)$$
$$\eta_t^{(j)} = \theta_j y_{t-1} + \eta_{t-1}^{(j+1)} - \phi_{j-1} a_t \quad (5)$$

Furthermore, the state vector $Y_t = \left(\eta_t^{(1)}, \ldots, \eta_t^{(m)}\right)^T$ satisfies

$$Y_t = \begin{bmatrix} \theta_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{m-1} & 0 & \ldots & 1 \\ \theta_m & 0 & \ldots & 0 \end{bmatrix} Y_{t-1} + \begin{bmatrix} 1 \\ -\phi_1 \\ \vdots \\ -\phi_{m-1} \end{bmatrix} \omega_t \quad (6)$$

The state-space form of the Arima model (*p*, *d*, *q*) has been found. Both have computational and conceptual advantages (Kumar and Goyal, 2011). This formulation ensures that the Arima model is responsive to the Kalman filter and smoothing for estimating model parameters and unobserved component extraction (Zulfi et al., 2018; Ananda and Wahyuni, 2021; Mehta and Sukmawaty, 2021). Estimation and updating of model parameters are part of the Kalman filter. These models are implemented using the "imputeTS" package, version 2.7, using the "StructTS" and "auto.arima" options of the "na.kalman" function in software version R. 4.1.1

2.3 Principal Component Regression

Rainfall data has very many variables and has high multicollinearity. Multicollinearity is the relationship of a condition where there is a correlation between independent variables or between independent variables that are not independent (Sahriman et al., 2014; Yu et al., 1997). The existence of multicollinearity in the multiple regression

model can cause the variance of the data set to enlarge so that the influence of each independent variable cannot be separated. Principle Component Regression (PCR) is an algorithm to reduce multicollinearity from a dataset (Nair et al., 2013). In addition, by usually regressing on only a subset of all the principal components, PCR can result in dimension reduction through substantially lowering the effective number of parameters characterizing the underlying model. This can be particularly useful in settings with high-dimensional covariates. Also, through appropriate selection of the principal components to be used for regression, PCR can lead to efficient prediction of the outcome based on the assumed model. In general, the PCR equation (7) is:

$$Y_{R(t)} = \beta_0 + \sum_{i=1}^{m} \beta_i C_i \quad (7)$$

where $Y_{R(t)}$ is the response variable data, β_0 is the intercept value, β_i is the coefficient for the i th component, and C_i is the i th principal component.

2.4 Method Evaluation

Root Mean Square Error (RMSE) is a measure of error or error between two corresponding values, in this case the predicted value and the actual value. In general, RMSE is formulated by equation (8).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t}^{n}(y_t - \hat{y}_t)^2} \quad (8)$$

Where y_t is the actual value, $\hat{y}_t$ is the predicted value, n is the number of data. In addition. to RMSE, the validation model that can be used is *Mean Absolute Percent Error* (MAPE). MAPE is used if the size of the forecasting variable is an important factor in evaluating the accuracy of the forecast. MAPE provides an indication of how big the forecast error is compared to the actual value of the *series*. The use of MAPE values has a *range of* values that can be used as measurement material regarding the ability of a forecasting model, the *range of* values can be seen in Table 1.

Table 1. Range of MAPE

Range MAPE	information
$<$ 10%	The ability of the forecasting model is very good
10 % - 20 %	The ability of the forecasting model is good
20% - 50%	The ability of the forecasting model is feasible
$<$ 50%	The ability of the forecasting model is poor

In general, MAPE is formulated in the equation (9).

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left(\frac{|y_t - \hat{y}_t|}{y_t}\right) \times 100 \quad (9)$$

With n is the amount of data, y_t is the actual value, $\hat{y}_t$ is the predictive value.

2.5 Model Development

This research will use NTT daily average rainfall data and Kupang's daily rainfall data at the Eltari observation station produced by BMKG from the website http://dataonline.bmkg.go.id/data_iklim. The probability of missing value at the daily rainfall of Kupang data does not depend on observed or unobserved data, these data are missing completely and randomly (MCAR) (Gill et al., 2007). NTT daily average rainfall data is used as a reference for the characteristics of the daily rainfall data for Kupang City to make the best estimation model. Algorithm performance is evaluated by various test scenarios. After that, build a forecasting model by adding GCM data. Use the best model to get long daily rainfall forecast results in Kupang City. For each test scenario, the following steps are performed.

1. Load the complete NTT daily average rainfall data (ts_complete).
2. Delete average NTT rainfall values based on missing values and unusual observation of daily rainfall for Kupang City and obtain time series with NA (ts_NA).
3. Apply the Imputation algorithm to ts_NA to get ts_Imputed.
4. Compare ts_complete and ts_Imputed using the appropriate of error size.
5. Get the smallest error measure. The smallest error size is the best model of the imputation algorithm.

6. Apply the best model of the imputation algorithm to the missing values and unusual observation of the daily rainfall of Kupang City.
7. Get GCM data (No-Missing-Value) through the website https://cds.climate.copernicus.eu/ with domain grid 3×3 to 12×12.
8. Statistical Machine Learning: divides data into two parts, training and testing. Build forecasting models using training data.
9. Validating and Evaluating The Learning: Test the forecasting model using data testing.
10. Get the best forecasting model from the best imputation method.
11. Do a daily rainfall forecasting for the future.

In addition, the variables in this study are the predictor variable (x) and the response variable (y). Predictor variables are GCM data with domain grid and response variables are Kupang City rainfall data. In this research, we use R Software 4.1.1 and steps in this study can be seen in Figure 1.

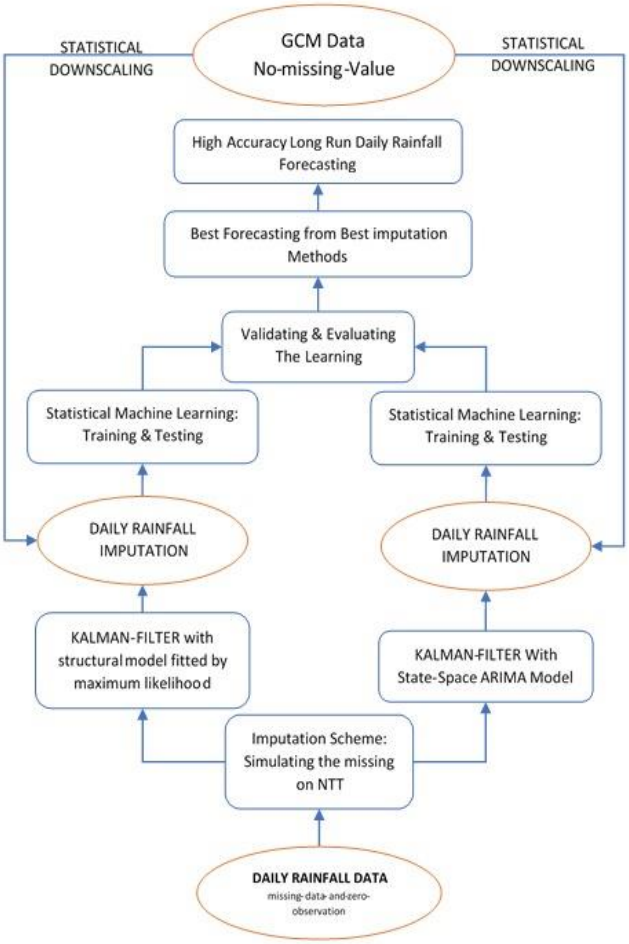

Figure 1. Flowchart of Research

3. Results and Discussion

3.1 Handling Missing Value

For solve missing values we use Arima method and Kalman Filter. Optimal accuracy in missing value imputation is assessed based on the given performance in selecting the best imputation method. Table 2. presents the performance results collected from the imputed missing values of the two methods used.

Table 2. Performance Indicators

NTT Daily Average Rainfall with Missing Value of Kupang		
Method	Parameter of Arima (p,d,q)	RMSE
Kalman Filter	-	0.979
Kalman with State-Space Model Arima	(2,1,1)	**0.930**
Kalman with State-Space Model Arima	(2,2,2)	0.938

The performance of the Kalman Filter with the Arima State-Space Model can be attributed to the relatively strong relationship between the missing values and the existing data. This best model is used for imputation of missing value in Kupang City. The graph of the imputed missing value in Kupang City is explicitly able to follow the trend (see Figure 2).

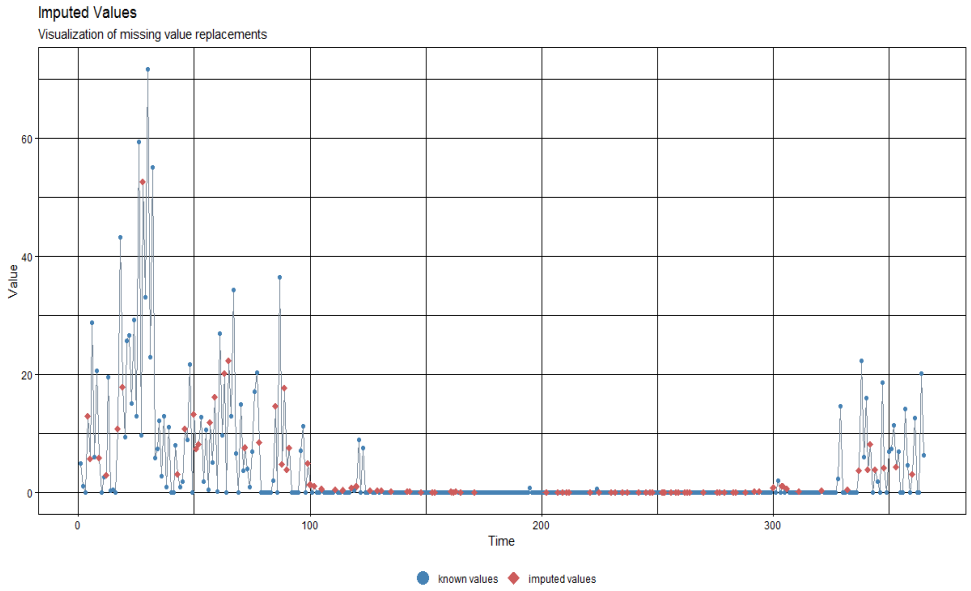

Figure 2. Plot Imputed Value

3.2 PCR Forecasting Models

Determination of the optimal domain produces an accurate daily rainfall forecasting model. In this study, the domain chosen for optimization consisted of 10 square grid sizes, 3 × 3 to 12 × 12. The ten grid sizes had the number of predictor variables 3, 16, 25, 36, 49, 64, 81, 100, 121, and 144 predictor variables. Therefore, in this study, PCR was used. Conceptually, PCR is based on PCA where there is a reduction in the dimensions of the predictor variables for each domain size. The new variable from the reduction was used to construct the PCR mode. The performance of the simulation model is evaluated by MAPE and compared with MAPE values in each grid size to get the best grid size with minimum MAPE. The simulation results using training data for each grid size are shown in Table 3. Table 3 shows that there is a small difference in RMSE between the 10 grid sizes but the 6 × 6 grid size produces a minimum MAPE. Grid size with minimum MAPE is the best domain and is suitable for daily rainfall forecasting in the future.

Table 3. Result Of Performance Statistics in Different Grid Sizes

Grid Size	Sum of PCs	Cumulative Variant Percentage (%)	MAPE
3 × 3	3	93.37%	1.51
4 × 4	6	94.34%	1.55
5 × 5	8	93.62%	2.67
6 × 6	**11**	**94.01%**	**1.30**
7 × 7	14	93.81%	1.33
8 × 8	17	93.44%	1.66
9 × 9	20	93.95%	1.71
10 × 10	25	93.74%	1.38

The grid size 6 × 6 has 36 predictor variables. In this study, the PCA method was used to extract orthogonal principal components (PCs). The PCs obtained were based on more than 93% variance. From Figure 3. The first 11

PCs have more than 93% information variance, so the number of new predictor variables used for this forecasting model consists of 11 PCs.

The SD results presented in Figure 4. are the results of the model test in the testing period. In this model, it can be seen that the prediction results are able to follow the actual data pattern. This model produces a MAPE value of 2.81% and an RMSE value of 10.81.

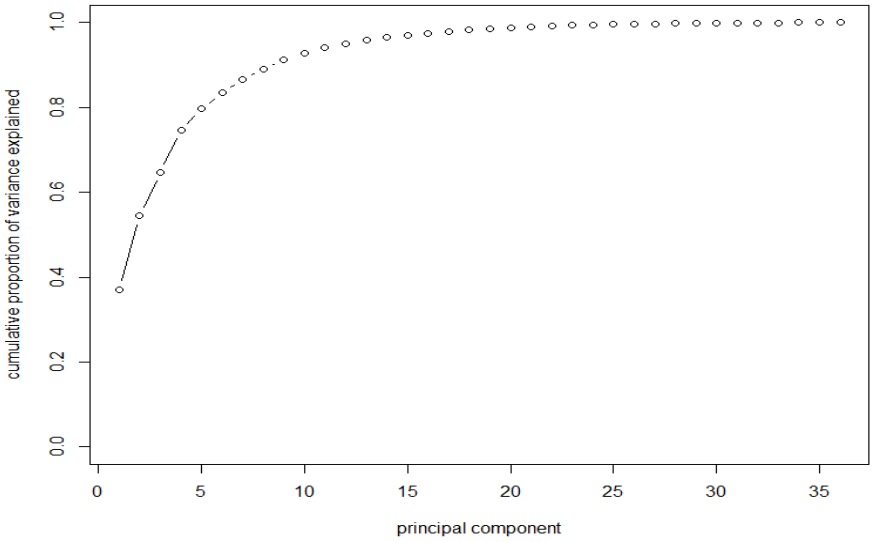

Figure 3. Cummulative Variance Plot

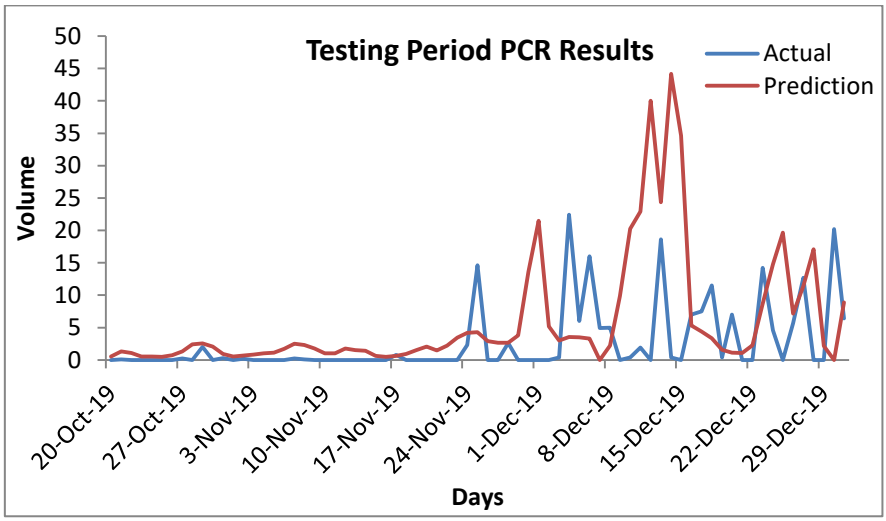

Figure 4. Daily Rainfall of Actual and Prediction by PCR

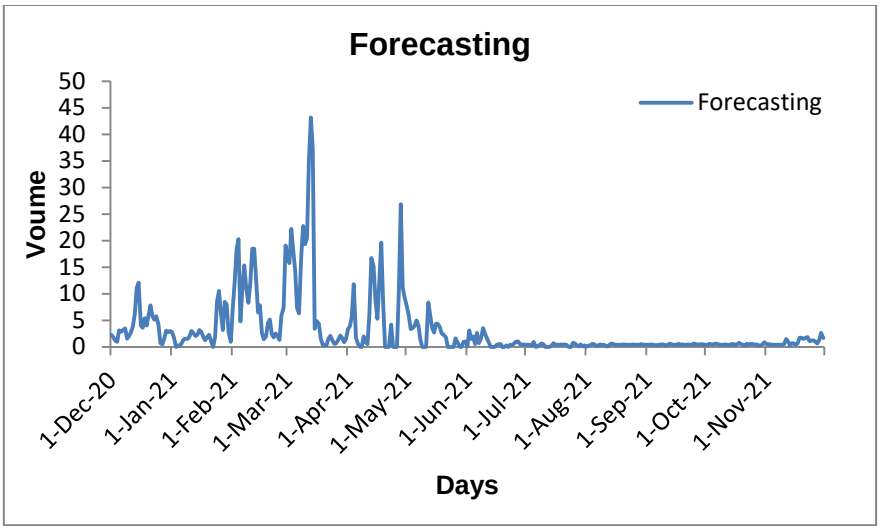

Figure 5. Daily Rainfall Forecasting by PCR

Therefore, the PCR model with main components (PCs) can be used for daily rainfall forecasting in the future period from December 1, 2020 to November 30, 2021. From Figure 5, the results can follow the same pattern from previous studies based on historical daily rainfall in Indonesia. research area. In the research area of daily rainfall from June to November tendency to lower and in December until May tendency to increase.

4. Conclusion

The Kalman Filter with the Arima State-Space Model is a technique imputation of missing values is more efficient and suitable to cope with missing values on climate data local. In this model there is an update of data when data is added or subtracted. The use of techniques statistical downscaling (SD)with the model Principal Component Regression (PCR) for daily rainfall forecasting in Kupang City has a very good ability based on the MAPE value range of 2.81 and the RMSE value of 10.81 in Table 1. The optimum PCR model is obtained at domain grid 6 × 6 with 11 selected principal components and a cumulative percentage of variance of 94.01%.

References

Ananda, E. Y. P., & Wahyuni, M. S. (2021, November). Rainfall Forecasting Model Using ARIMA and Kalman Filter in Makassar, Indonesia. In *Journal of Physics: Conference Series* (Vol. 2123, No. 1, p. 012044). IOP Publishing.

Boesday, O. J., Kaho, L. M. R., & Effendi, J. (2020, November). An Analysis of Disaster Mitigation Readiness in Coastal Areas of Kupang City. In *International Seminar on Sustainable Development in Country Border Areas* (Vol. 2, No. 1, pp. 71-81).

Estiningtyas, W., & Wigena, A. H. (2011). Teknik statistical downscaling dengan regresi komponen utama dan regresi kuadrat terkecil parsial untuk prediksi curah hujan pada kondisi el nino, la nina, dan normal. *Jurnal Meteorologi dan Geofisika*, *12*(1), 65-72.

Farikha, E. F., Hadi, A. F., Anggraeni, D., & Riski, A. (2021, May). Projection pursuit regression in statistical downscaling model using artificial neural network for rainfall prediction. In *Journal of Physics: Conference Series* (Vol. 1872, No. 1, p. 012021). IOP Publishing.

Field, R. D., Wang, Y., & Roswintiarti, O. (2004). A drought-based predictor of recent haze events in western Indonesia. *Atmospheric Environment*, *38*(13), 1869-1878.

Gill, M. K., Asefa, T., Kaheil, Y., & McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. *Water resources research*, *43*(7), 1-12.

Kumar, A., & Goyal, P. (2011). Forecasting of air quality in Delhi using principal component regression technique. *Atmospheric Pollution Research*, *2*(4), 436-444.

Mehta, W. A., & Sukmawaty, Y. (2021, November). Rainfall prediction climatological station of Banjarbaru using arima kalman filter. In *Journal of Physics: Conference Series* (Vol. 2106, No. 1, p. 012003). IOP Publishing.

Mursidi, A., & Sari, A. D. P. (2017). Management of drought disaster in Indonesia. *Jurnal Terapan Manajemen dan Bisnis*, *3*(2), 165-171.

Rozaki, Z., Wijaya, O., Rahmawati, N., & Rahayu, L. (2021). Farmers' disaster mitigation strategies in Indonesia. *Reviews in Agricultural Science*, *9*, 178-194.

Sachindra, D. A., Ahmed, K., Rashid, M. M., Shahid, S., & Perera, B. J. C. (2018). Statistical downscaling of precipitation using machine learning techniques. *Atmospheric research*, *212*, 240-258.

Sahriman, S., Djuraidah, A., & Wigena, A. H. (2014). Application of principal component regression with dummy variable in statistical downscaling to forecast rainfall. *Open Journal of Statistics*, *4*(09), 678.

Susandi, A., Tamamadin, M., Djamal, E., & Las, I. (2015, September). Information system of rice planting calendar based on ten-day (Dasarian) rainfall prediction. In *AIP Conference Proceedings* (Vol. 1677, No. 1, p. 100002). AIP Publishing LLC.

Titu-Eki, A., & Kotta, H. Z. (2021). Environmental geology assessment on the regional Pitay landfill site: a case study in Kupang, Indonesia. *SN Applied Sciences*, *3*(1), 1-13.

Yu, Z. P., Chu, P. S., & Schroeder, T. (1997). Predictive skills of seasonal to annual rainfall variations in the US affiliated Pacific islands: Canonical correlation analysis and multivariate principal component regression approaches. *Journal of Climate*, *10*(10), 2586-2599.

Zulfi, M., Hasan, M., & Purnomo, K. D. (2018, April). The development rainfall forecasting using kalman filter. In *Journal of Physics: Conference Series* (Vol. 1008, No. 1, p. 012006). IOP Publishing.