Handling Missing Values and Unusual Observations in Statistical Downscaling Using Kalman Filter

M D Saputra¹, A F Hadi²*, A Riski², D Anggraeni²

¹ Department of Mathematics, University of Jember, Jember, 68121, Indonesia
² Data Science Research Group, Department of Mathematics, University of Jember, Jember, 68121, Indonesia

*E-mail: afhadi@unej.ac.id

Abstract. Rainfall forecasting model using data Global Circular Model (GCM) with Statistical Downscaling technique has a fairly high accuracy. However, missing local climate information poses a constraint in data analysis and forecasting. Missing value imputation is one solution that can be used. Kalman Filter Imputation and State Space Model Arima are imputation methods that operate recursively where there is an update of prediction values when data updates occur. This study aimed to find the best model to use for missing value imputation with small imputation errors. The results of the missing value imputation were used to obtain the best statistical downscaling model on a 3 × 3 to 12 × 12 grid. The research was conducted on the daily rainfall data of Kupang City with 17% missing values and 8% unusual data at the Eltari Meteorological observation station, Kupang city. The average daily rainfall data in East Nusa Tenggara Province were utilized as a reference for the characteristics of rainfall data at the Kupang City observation station. The best missing value imputation was obtained by using the Arima State-Space Model (2,1,1) with a Root Mean Square Error (RMSE) of 0.930 and the model was statistical downscaling best obtained on a grid 6 × 6 with a Mean Absolute Percentage Error (MAPE) of 1.3 % and the number of PCs 11.

1. Introduction

The missing values problems often occur in several types of time series data used in rainfall forecasting. This problem can be overcome with the downscaling method. Downscaling is a method that can explain the relationship of information or data with large-scale atmospheric variables. Data with large-scale atmospheric variables were obtained from the Global Circular Model (GCM). The data can be reduced to a finer, spatial, and temporal scale [1]. In statistics it is called Statistical Downscaling.

In the application of statistical downscaling, local climate information is needed to find the relationship between parameters and global scale climate [2]. However, this application has several 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). According to Cyer, if the missing value lies in the middle of the data, it is necessary for estimate to fill in the missing value [3]. Kalman Filter Imputation and State Space Model Arima is a model formulation that provides a very powerful tool for recursive estimation in a dynamic system [4, 5, 6]

Based on the previous research conducted by Muflihah and Pahlawan, it is explained that the Kalman filter produces the highest correlation value of 0.78-0.86 compared to other interpolation methods [7]. Then research conducted by Albinsson and Gillsbro shows that the Kalman filter has the best results
with an average error of 0.67 compared to TS-EM which has an average error of 0.69 [8]. Based on the explanation above, it can be concluded that Kalman filters and state space models often have high accuracy values and low RMSE values. This is evidenced by many researchers discussing the imputation of missing values using kalman filter. In this study, the Kalman filter and state space model Arima will be used for imputation of missing values and test the results of imputation on statistical downscaling for rainfall forecasting using the principal component regression method.

This research was conducted on the daily rainfall data of Kupang City from BMKG data. From the data obtained, there are 17% missing values and 8% unusual observation at the Eltari Meteorological observation station, Kupang City. The average daily rainfall data for East Nusa Tenggara Province was used as a reference for the characteristics of daily rainfall data at the Eltari Meteorological observation station in Kupang City to obtain the best Arima model. The best imputation results were used as local data to obtain models in statistical downscaling.

2. Material and Method

2.1. Global Circular Model (GCM)
The Global Circular Model is a computer-based climate model that uses numerical and deterministic equations that follow the physical rules used to predict climate and weather, and able to understand climate change and weather [9]. GCM data have high spatial and temporal resolution [10].

2.2. ARIMA(p, d, q) in State-Space forms, Kalman Filter and Smoothing
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 [11]. 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)

\[ y_t = H_t Y_t + \varepsilon_t \]  
\[ Y_t = Z_t Y_{t-1} + R_t \omega_t \]  

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

\[ y_t = \theta_0 y_{t-1} + \cdots + \theta_m y_{t-m} + a_t - \phi_1 a_{t-1} - \cdots - \phi_m a_{t-m+1} \]  

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

\[ \eta_t^{(m)} = \theta_0 y_{t-1} - \phi_m a_t \]  
\[ \eta_t^{(j)} = \theta_j y_{t-1} + \eta_{t-1}^{(j+1)} - \phi_j a_t \]  

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

\[ Y_t = \begin{bmatrix} \theta_0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{m-1} & 0 & \cdots & 1 \\ \theta_m & 0 & \cdots & 0 \end{bmatrix} Y_{t-1} + \begin{bmatrix} 1 \\ -\phi_1 \\ \vdots \\ -\phi_{m-1} \end{bmatrix} \omega_t \]  

The state-space form of the Arima model \( (p, d, q) \) has been found. Both have computational and conceptual advantages [12, 13]. This formulation ensures that the Arima model is responsive to the kalman filter and smoothing for estimating model parameters and unobserved component extraction [14]. 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. 4.0.2 [15, 16].
2.3. Statistical Downscaling (SD)
Statistical downscaling is defined as a connecting technique between global scale variables (explanatory variables) and local scale variables [17]. The general equation for Statistical Downscaling is as follows [18].

\[ y = f(x) + \varepsilon \]  

(7)

Variable \( y \) is the response variable with dimension \((t \times 1)\), \( x \) is the predictor variable which has dimension \((t \times p)\) where \( t \) is time, \( p \) is the number of grids, and \( \varepsilon \) the residual value.

2.4. Principal Component Regression (PCR)
PCR is a method that can be used to solve multi-collinear problems [19]. Rainfall data has so many variables that can lead to high multicollinearity. This method will produce uncorrelated main components. In general, the PCR equation is:

\[ Y_{R(t)} = \beta_0 + \sum_{i=1}^{m} \beta_i C_i \]  

(8)

where \( Y_{R(t)} \) is the response variable data, \( \beta_0 \) is the intercept value, \( \beta_i \) is the coefficient for the \( i \) th component, and \( C_i \) is the \( i \) th principal component.

3. 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) [20]. NTT daily average rainfall data is used as a reference for the characteristics of the daily rainfall data for Kupang to make the best estimation model. Algorithm performance is evaluated by various test scenarios. For each test scenario, the following steps are performed [21, 22, 23]:

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

In addition, researchers also use other data in the form of data obtained by researchers from GCM data through the website https://cds.climate.copernicus.eu/ with domain grid 3 × 3 to 12 × 12. 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 rainfall data.

The research steps in this study are attached as follows

![Flowchart of Research](image)

**Figure 2.** Flowchart of Research

4. Result and Discussion

The estimation accuracy of the method is assessed based on the performance indicators provided in the selection of the best imputation method. Table 1. presents the performance indicators collected from the missing value imputation of the two methods used

| Table 1. Performance Indicators |
|---------------------------------|
| NTT Daily Average Rainfall with Missing Value of Kupang | Performance Indicators |
| Models | RMSE | MAE |
| Kalman Filter | 0.979 | 0.236 |
| State-Space Model Arima (2,1,1) | **0.930** | **0.214** |
| State-Space Model Arima (2,2,2) | 0.938 | 0.215 |

Based on Table 1, the imputation of the best missing values is obtained using the Arima State-Space Model (2, 1, 1) with a Root Mean Square Error (RMSE) of 0.930. 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 3).

![Figure 3. Plot imputed value](image)

Daily GCM output data having a low resolution that is equal to $1.875^\circ \times 1.86442^\circ$ or 388 km. In this condition, SD technique can be used to connect global scale variables and local scale variables. The GCM output data was a predictor variable, while the data from the daily rainfall imputation of mussel was the response variable. The number of predictor variables based on the result of the multiplication grid $n \times n$. The size of the grid used in this study is $3 \times 3$ to $12 \times 12$ so that it is possible to have multicollinearity between the predictor variables. The existence of this multicollinearity can cause the modeling results to have a fairly large bias, one method to overcome this is the principal component analysis (PCA). The smallest number of PCs with a percentage of the cumulative variance of PCA was used as a predictor variable in the formation of the pre-estimate model using the principal component regression method. A total of 10 pre-estimated PCR models were obtained from a combination of grid domain sizes. Furthermore, the grid search stage is carried out to obtain the most representative grid as a predictor for local scale data. The choice of grid domain size is based on the lowest training MAPE value (see table 2).

| Domain Grid Size | Number of PCs | Cumulative Percentage (%) | MAPE (%) |
|------------------|---------------|----------------------------|----------|
| $3 \times 3$     | 3             | 93.37%                     | 1.51     |
| $4 \times 4$     | 6             | 94.34%                     | 1.55     |
| $5 \times 5$     | 8             | 93.62%                     | 2.67     |
| $6 \times 6^*$   | 11            | 94.01%                     | 1.30     |
| $7 \times 7$     | 14            | 93.81%                     | 1.33     |
| $8 \times 8$     | 17            | 93.44%                     | 1.66     |
| $9 \times 9$     | 20            | 93.95%                     | 1.71     |
| $10 \times 10$   | 25            | 93.74%                     | 1.38     |
Based on the Table 2, the best pre-estimated PCR model was obtained on grid $6 \times 6$ with a MAPE is 1.30% and the number of PCs 11 with a cumulative percentage of 94.01%. Thus, the results of the imputation of the missing value in the daily rainfall of Kupang city using the Kalman filter produce small error value for use in statistical downscaling. Model can also be regarded as the best model because they have less of the value of MAPE 10% [24].

5. Conclusion

Based on the results and discussion, it can be explained that the model Kalman Filter with the Arima State-Space Model (2,1,1) is the most efficient and suitable model to be used. It has the Root Mean Square Error (RMSE) of 0.930. The model also shown a good compatibility in entering missing value in NTT daily average rainfall data. The graphs generated from the model can follow trends in missing value patterns. The best model of imputation of missing value on the average daily rainfall in NTT is used to impute the lost value in the daily rainfall of Kupang city. The results of this imputation resulted in a MAPE value of 1.30% in the statistical downscaling model with 11 PCs and a cumulative percentage of 94.01%.

Acknowledgement

We would like to thank Ministry of Research and Technology/National Research and Innovation Agency of Indonesia, especially The Directorate of Research and Community Empowerment (DRPM). This paper was part of the research that funded by the DRPM with contract number 175/SP2H/AMD/LT/DRPM/2020. We also thanks to all member of Data Science Research Group at Department of Matematics, University of Jember.

References

[1] Anuchaivong P, Sukawat D, and Luadsong A 2017 Statistical Downscaling for Rainfall Forecasts Using Modified Constructed Analog Method in Thailand https://doi.org/10.1155/2017/1075868
[2] Sahriman S, Kalondeng A, and Koerniawan V 2019 Pemodelan Statistical Downscaling Dengan Peubah Dummy Berdasarkan Teknik Cluster Hierarki Dan Non-Hierarki Untuk Pendugaan Curah Hujan Indonesian Journal of Statistics and Its Applications (eISSN:2599-0802)Vol 3No 3(2019), 295 -309
[3] Cryer J 1986 Time series analysis Boston: PWS-KENT Publishing Company
[4] Anderson B D O dan Moore J B 1979 Optimal Filtering New Jersey: Prentice-Hall, Englewood Cliffs
[5] Muflihah 2015 Prediksi curah hujan melalui model State Space untuk data hilang Makassar: Universitas Hasanuddin
[6] Fung D S 2006 Methods for the estimation of missing values in time series https://ro.ecu.edu.au/theses/63
[7] Muflihah and Pahlawan R Y 2017 perbandingan teknik interpolasi berbasis r dalam estimasi data curah hujan bulanan yang hilang di sulawesi Jurnal Meteorologi dan Geofisika Vol. 18, 107-111
[8] Albinsson M and Gillsbro E 2017 Imputation Methods in Dialysis Data Sweden: Lund University
[9] Sutikno, Setiawan S, and Purnomoadi H 2010 Statistical Downscaling Output GCM Modeling with Continuum Regression and Pre-Processing PCA Approach The Journal for Technology and Science Vol. 21 No 3
[10] Joyce R J, Janowiak J E, Arkin P A, and Xie P 2004 CMORPH: A Method That Produces Global Precipitation Estimates From Data High Spatial and Temporal Resolution. J Hydrometeor 5 487-503
[11] de Jong R, van Buuren S, and Spiess M 2016 Multiple imputation of predictor variables using generalized additive models Communications in Statistics Simulation and Computation 45 968–985
[12] Harvey A C 1989 Forecasting structural time series models and the Kalman filter Cambridge UK: Cambridge University Press
[13] Hamilton J D 1994 Time Series Analysis Princeton NJ: Princeton University Press
[14] Yamoah E A, Mueller U A, Taylor S M, and Fisher A J 2020 Missing data imputation of high-resolution temporal climate time series data Meteorol Appl 27 e1873
[15] R Core Team 2013 R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing
[16] Moritz S and Bartz-Beielstein T 2017 imputeTS: Time series missing value imputation in R R Journal 9(1) 207–218
[17] Sailor D, Hu T, Li X, and Rosen J N 2000 A Neural Network Approach to Local Downscaling of GCM Output for Assessing Wind Power Implications of Climate Change. Renewable Energy 359–379
[18] Zorita E and Storch, H V 1999 The Analog Method as A Simple Statistical Downscaling Technique: Comparison With More Complicated Methods. Journal Climatology 12 2474-2489
[19] Ohyver M 2012 Pemodelan Principal Component Regression dengan Software R ComTech Vol 3 No 1 177-185
[20] Rubin D B 1976 Inference and missing data Biometrics 63 (3) 581-592
[21] O. Anava, E. Hazan, and A. Zeevi 2015 Online time series prediction with missing data in: ICML pp. 2191–2199.
[22] Weeks M 1999 Methods of Imputation for Missing Data (Fifth Draft) Cambridge : University of Cambridge
[23] Chaudhry A, Li W, Basri A, and Patenaude F 2019 A Method for Improving Imputation and Prediction Accuracy of Highly Seasonal Univariate Data with Large Periods of Missingness Wireless Communication and Mobile Computing https://doi.org/10.1155/2019/4039758
[24] Ramli J H B 2013 Analisis Peramalan Ekspor Indonesia Pasca Krisis Keuangan Eropa Dan Global Tahun 2008 Dengan Metode Dekomposisi. Jurnal Ekonomi dan Keuangan. vol. 1, No. 3, pp. 117–133.