IJDASEA
International Journal of Data Science
Engineering, and Analytics

Model Selection for Forecasting Rainfall Dataset

Amri Muhaimin [1]*, Hendri Prabowo [2] and Suhartono [2],

[1] Universitas Pembangunan Nasional "Veteran" Jawa Timur; amri.muhaimin.stat@upnjatim.ac.id
[2] Institut Teknologi Sepuluh Nopember;
* Correspondence: amri.muhaimin.stat@upnjatim.ac.id;

Citation: Muhaimin, A.; Prabowo, H.; Suhartono. Model Selection for Forecasting Rainfall Dataset. *C* **2021**, Vol *1*, Page 1-10. https://doi.org/10.3390/xxxxx

Academic Editor: Amri Muhaimin

Received: 28 June, 2021
Accepted: 15 July, 2021
Published: 19 July, 2021

Abstract: The objective of this research is to obtain the best method for forecasting rainfall in the Wonorejo reservoir in Surabaya. Time series and causal approaches using statistical methods and machine learning will be compared to forecast rainfall. Time series regression (TSR), autoregressive integrated moving average (ARIMA), linear regression (LR), and transfer function (TF) are used as a statistical method. Feedforward neural network (FFNN) and deep feed-forward neural network (DFFNN) is used as a machine learning method. Statistical methods are used to capture linear patterns, whereas the machine learning method is used to capture nonlinear patterns. Data about hourly rainfall in the Wonorejo reservoir is used as a case study. The data has a seasonal pattern, i.e. monthly seasonality. Based on the cross-validation and information criteria, the results showed that DFFNN using the time series approach has a more accurate forecast than other methods. In general, machine learning methods have better accuracy than statistical methods. Furthermore, additional information is obtained, through this research the parameter that best to make a neural network model is known. Moreover, these results are also not in line with the results of M3 and M4 competition, i.e. more complex methods do not necessarily produce better forecasts than simpler methods.

Keywords: causal; machine learning; model selection; neural network; statistical; time series

1. Introduction

Rain has an important role in the management and planning system of water resources, especially in tropical countries like Indonesia [1]. The Rainfall in Indonesia, especially Wonorejo, Surabaya city as a tropic area has a big variation. Therefore, accurate rainfall forecasting is very important in managing the water resources, such as drinking water demand, the availability of groundwater, hydroelectric power plant, irrigation water demand, and flood control [2]. The best accurate rainfall forecasting can be obtained by using the best forecasting method that suits the pattern of rainfall data in Wonorejo, Surabaya city.

In forecasting, a time-series approach and causal approach can be used. The forecasting method commonly is used for solving practical problems in statistical methods [3]. Time series regression (TSR) in general is the same as the linear regression model [4]. Another linear forecasting model is Autoregressive Integrated Moving Average or ARIMA method. ARIMA method is one of the most popular methods in time series forecasting [5, 6]. TSR and ARIMA is a time series approach. The statistical method that using a causal approach is the transfer function (TF) [7]. Furthermore, TSR, ARIMA, and TF can be used for forecasting data that follow linear patterns. However, many real data not only follow linear patterns. Thus, a nonlinear model is needed to handle this nonlinearity pattern. Recently, many nonlinear methods were proposed and applied for time series forecasting.

Neural Network (NN) is one of the nonlinear methods that frequently used for solving forecasting problems [8].

Some research in the prediction of rainfall has been carried out. Azumanga and Saranya [9] used Seasonal ARIMA (SARIMA) model in forecasting rainfall in India. Chattopadhyay and Chattopadhyay [10] compared ARIMA and Artificial Neural Network (ANN) in forecasting rainfall in India. Yu et al. [11] also compared Random Forest (RF) and Support Vector Machine (SVM) in forecasting rainfall in Taiwan. The result was showed that SVM is better than Random Forest.

This study focused on three statistical methods, i.e. TSR, ARIMA, and TF, and two machine learning methods, i.e. FFNN and DFFNN to forecast rainfall in Wonorejo, Surabaya city. The forecasting will use time series and causal approaches. The rest of the paper is organized as follows: Section 2 reviews the methodology, i.e. TSR, ARIMA, TF, and FFNN; Section 3 presents the dataset and methodology; Section 4 presents the results, analysis, and discussion; and Section 5 presents the conclusion from this study.

2. Related Works

There some methods and approaches to create a forecast model. Such as linear and non-linear methods. Several linear methods are often used to forecast, such as TSR, ARIMA, and TF. Otherwise, the non-linear method that is often used is ANN. TSR is a method using regression-based. Generally, the TSR model is almost the same as linear regression that is the predictor variables that influenced the response variable [4]. ARIMA is one of the popular methods in time-series forecasting. ARIMA workflow is based on the lag of the data, also model identification is needed to extract which lag significantly affects the data. The transfer function is a model which is based on the relationship between time-series data as response variable (output series) with one or more predictor variables (input series) [7]. Many researchers about rainfall forecasting are already done, either use linear or non-linear methods.

Novel hybrid already did by [2] with the linear and non-linear machine learning method to forecast a monthly rainfall. The non-linear method used is Extreme Learning Machine (ELM) and combined with Single Layer FFNN. The metrics score that used are R-Square, RMSE, MAE, root mean relative squared error (RSMRE), and mean absolute relative error (MARE). The lag that is included as a predictor is lag 1, lag 2, and lag 12 to the seasonal pattern. The result is the non-linear method outperforms the linear method in almost all metrics.

Ref [12] did rainfall and temperature forecasting using a monthly dataset. The method used is seasonal ARIMA, and the metric evaluation used is the root mean squared error (RMSE) and R^2. It stated that seasonal ARIMA produces a reliable forecast value with an RMSE score is 62.40 and an R-square score is 72%. Ref [13] also did a rainfall forecasting using seasonal ARIMA. The model used is ARIMA (0,0,0)(0,1,0), which means the model used is integrated into the seasonal component. The metrics evaluation that used also RMSE and R-square.

Ref [14] uses the ANN method to forecast rainfall in the lake basin area. The ANN used is a feed-forward neural network (FFNN). Ref [14] compared it with ARIMA models. The metrics evaluation used is RMSE and mean absolute error (MAE). The testing data that used is 72 months ahead. The result is the ANN and ARIMA models are not significantly different. Yet in RMSE and MAE testing data, FFNN is better than the ARIMA model.

This study focused on two approaches in forecasting, i.e. time series approach and the causal approach. Statistical methods and machine learning methods were used on each approach. The time-series approach used univariate time series, the methods used were TSR, ARIMA, FFNN, and DFFNN, while the causal approach involved several variables (response variable and predictor variable). The methods used in the causal approach were LR, TF, and FFNN. Furthermore, this study will compare the predictor variables used in the model, and also the parameter that created the neural networks model.

3. Experiment and Analysis

The experiments carried out were to make comparisons between linear, nonlinear, and hybrid methods. Experimental matters include pre-processing data, determining input variables, and determining parameters in building a neural network model. Each treatment was carried out using the same data and the same amount of training-testing as well. Furthermore, the metrics score used contains information criterion score and cross-validation. Additional information that will be obtained is the best parameter to build a neural network model for rainfall data.

3.1. Dataset

Monthly data of rainfall in Wonorejo, Surabaya city is used in this study. In addition, there is also humidity data as the predictor data. Data used is time-series data of monthly rainfall from January 1998 until December 2018. The time series plot and a scatterplot of the data are shown in Figure 1.

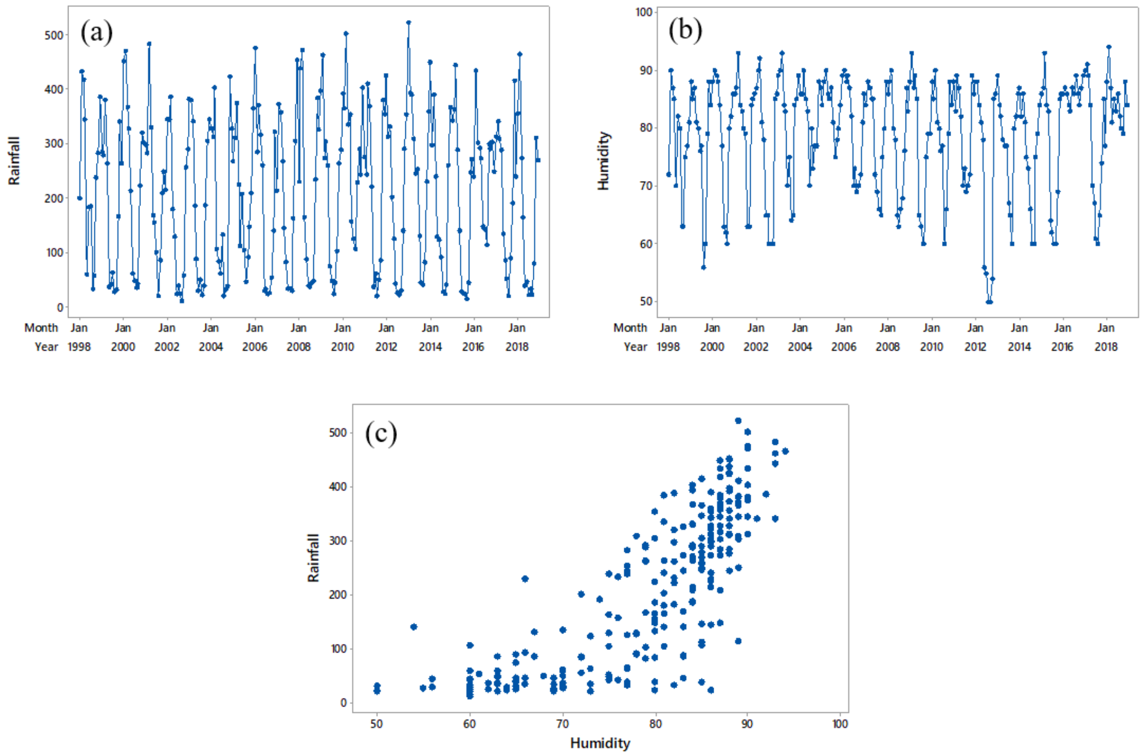

Figure 1. Time series plot rainfall (a), time series plot humidity (b), and scatter plot rainfall and humidity(c)

The rainfall and humidity data is monthly data, which is calculated by the summation of the rainfall every day within the same month. The rainfall pattern seems to be stationary in variance and mean, The humidity data used to create a transfer function model because the TF model needs another response to create the model. The scatter plot might be shown a correlation between rainfall and humidity.

The analysis is started with data visualization. Figure 2 shows the line plot of rainfall data in the Wonorejo reservoir in Surabaya City. It can be seen that rainfall has an annual seasonal pattern. From 1998 to 2018 rainfall tends to be high from November to March. While in May to September rainfall tends to be lower than others. In general, this condition occurs almost every year from 1998 to 2018, so rainfall data have an annual seasonal pattern. Although in 2010 and 2016 the rainfall in September was supposed to be low to high, this was caused by natural phenomena.

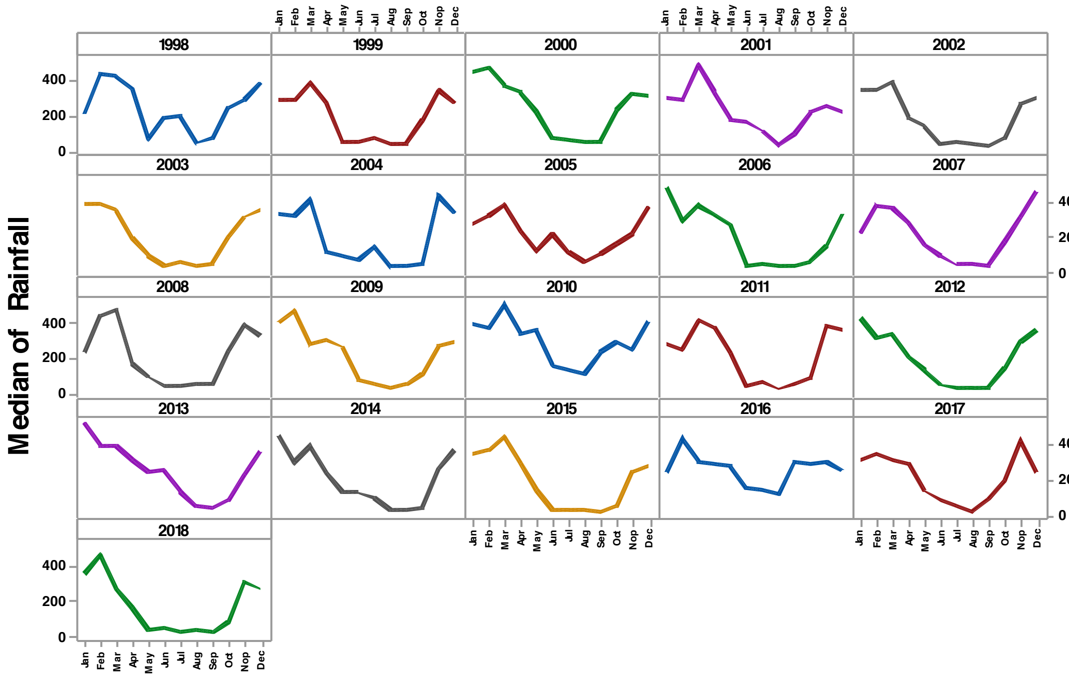

Figure 2. Line plot of rainfall data

3.2. Methodology

As mentioned before, the data used in this study is from January 1998 until December 2018. The data was divided into in-sample data and out-of-sample data. Data from January 1998 until December 2017 as in-sample data, while the data from January 2018 until December 2018 as out-of-sample data. To select the best model, the researcher used two scenarios that are information criteria and cross-validation. Furthermore, the results were compared. The best model was chosen by the information criteria has been done by calculating the Akaike information criteria (AIC) and Bayesian information criteria (BIC) of in-sample data [15]. The formula of AC and BIC are as follows:

$$AIC = n \ln\left(\frac{S}{n}\right) + 2p, \tag{1}$$

$$BIC = n \ln\left(\frac{S}{n}\right) + p + p \ln(n), \tag{2}$$

where n is the number of observations of in-sample data, S is sum square error (SSE) and p is the number of parameters in the model. The best model was also chosen by cross-validation has been done by calculating the root mean square error prediction (RMSEP) and mean absolute percentage error prediction (MAPEP) [7] of out-of-sample data. The formula of RMSEP and MAPEP is defined as:

$$RMSEP = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(Y_{n+l} - \hat{Y}_n(l)\right)^2}, \tag{3}$$

$$MAPEP = \left(\frac{1}{L}\sum_{l=1}^{L}\frac{|Y_{n+l} - \hat{Y}_n(l)|}{|Y_{n+l}|}\right)100\%, \tag{4}$$

where, Y_{n+l} is the actual value of out-of-sample data, $\hat{Y}_n(l)$ is the prediction value of out-of-sample data and L is the size of out-of-sample data. Figure 3 shows the methodology of choosing the best model in the general study.

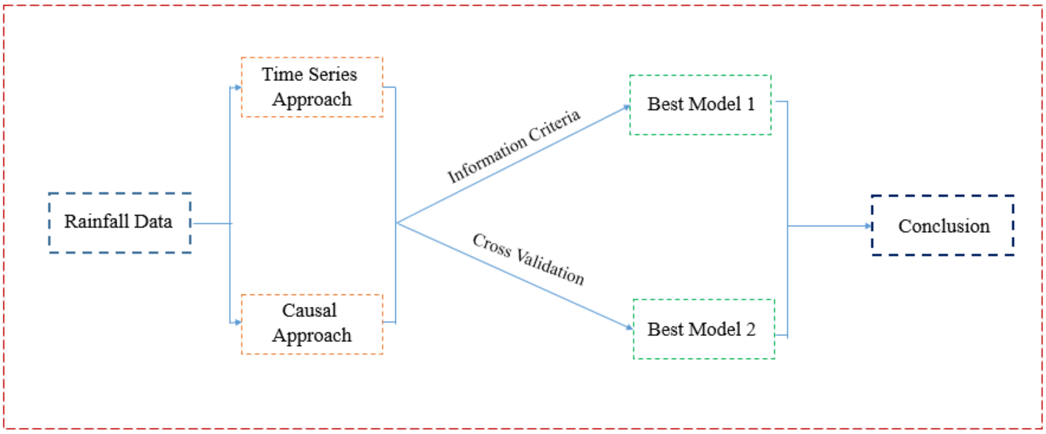

Figure 3. The best model selection scheme

3.3 Forecasting Rainfall with Time Series Model

The time series model used in predicting rainfall in the Wonorejo reservoir in Surabaya uses statistical methods (TSR and ARIMA) and machine learning (FFNN and DFFNN). From the identification of data, patterns have been known that the data has a seasonal pattern. So the TSR model used in this study uses one type of predictor in the form of a seasonal dummy month in one year. So there are 12 predictors in the TSR model. The use of this dummy variable aims to capture seasonal patterns from rainfall data. The following is the TSR model:

$$Y_t = 342B_1 + 356.6B_2 + 385.5B_3 + 267.3B_4 + 163.9B_5 + 93.8B_6 + 77.4B_7 + 37{,}7B_8 + 66.6B_9 + 150.2B_{10} + 292.8B_{11} + 324.5B_{12} + \varepsilon_t \quad (5)$$

Furthermore, the ARIMA model was obtained based on the Box Jenkins procedure. The Box Jenkins procedure begins with the identification of the stationarity of the data. From the identification step, it is found that the rainfall data is not stationary. Then differencing in seasonal lag 12 is applied on the data so the data becomes stationary. From the ARIMA order identification results, it is found that the best ARIMA model is ARIMA (0,0,[1,4])(0,1,1)12. This ARIMA model has residuals that meet the assumptions of white noise and are normally distributed. The best ARIMA rainfall model can be written as follows:

$$Y_t = Y_{t-12} + a_t + 0.17a_{t-1} + 0.16a_{t-4} - 0.75a_{t-12} - 0.13a_{t-13} - 0.12a_{t-16} \quad (6)$$

The formation of the NN model is based on the ARIMA model, especially in determining the NN input. There are 2 types of input in the NN model, which are based on the AR lag of the ARIMA model and based on the PACF lag. Based on equation (6) the input in rainfall forecasting based on the ARIMA model with NN is lag 12. From Figure 4 the input based on PACF is lag 1, 3, 5, 12, 13, 15, 17, 21, 24, 25, 26, 33, 36, 38 and 48. Before modeling the data, pre-processing was done using 3 scenarios i.e. normalized, adjusted normalized, and standardized. The NN model in this study uses tanh and logistical activation functions tried 1 and 2 hidden layers and tried 1 to 5 neurons. For 1 hidden layer, it is called a feed-forward neural network (FFNN) and for 2 hidden layers, it is called a deep feed-forward neural network (DFFNN). In addition, 5 optimization algorithms were also tried, namely backprop, rprop +, rprop-, sag and slr. 10 replications were made in forming the FFNN and DFFNN models. After that, it will be compared to get the best FFNN and DFFNN architecture in predicting rainfall. Table 1 is the best model of FFNN and DFFNN based on cross-validation and information criteria.

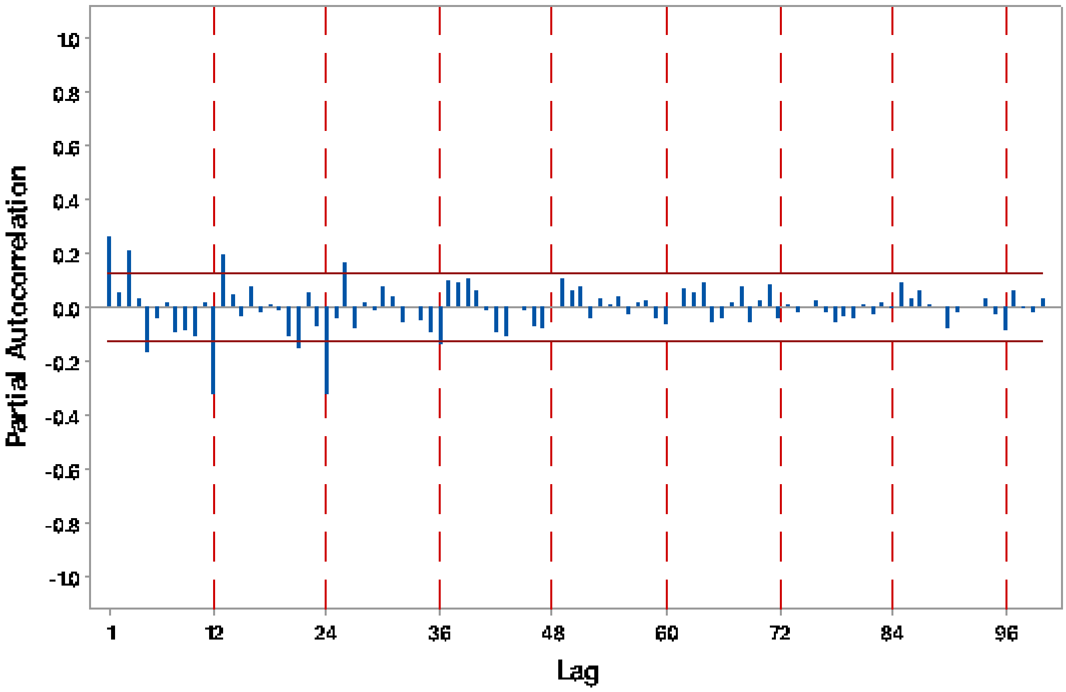

Figure 4. PACF plot rainfall data with 12 difference

From Table 1 it can be seen that based on cross-validation (RMSEP and MAPEP) and information criteria (AIC and BIC) obtained the best architecture to forecast rainfall in the Wonorejo reservoir in Surabaya using FFNN and DFFNN is different. In general, the PACF input lag is better than using the ARIMA input lag. The best FFNN model based on cross-validation is called FFNN1 and based on information criteria is called FFNN2. Similarly, in DFFNN there are DFFNN1 and DFFNN2.

Tabel 1. The best FFNN and DFFNN model

Properties	FFNN		DFFNN	
	Cross Validation	**Information Criteria**	**Cross Validation**	**Information Criteria**
Number of Neuron	3	5	(5,2)	(5,5)
Input	PACF	PACF	PACF	PACF
Preprocessing	Normalized	Standardized	Normalized	Normalized
Algorithm	backprop	Backprop	backprop	backprop
Activation Function	Tanh	Logistic	Tanh	Tanh

3.4 Forecasting Rainfall with Causal Model

The difference between the time series model and the causal model is that the causal model involves other variables (predictors). Causal models that used in forecasting rainfall in the Wonorejo reservoir in Surabaya are statistical methods i.e. linear regression (LR) and transfer function (TF), and machine learning method (FFNN). The predictor used in this study is humidity. The problem that arises when forecasting a causal approach is to have to predict the predictor variables first. This applies to LR and FFNN models, whereas in TF it is not necessary to predict the predictors first.

Humidity is predicted using a model ARIMA (0,0,4)(0,1,1)[12]. This model has a residual that meets the white noise assumption and is normally distributed. The linear regression model (LR) used in this study uses a predictor in the form of humidity, where humidity was previously predicted by the ARIMA model (0,0,4)(0,1,1)[12]. The following is the LR model:

$$Y_t = -682.4B_1 + 11.386X_t + \varepsilon_t \tag{7}$$

The transfer function (FT) model is obtained through several stages. The first is the pre-whitening of the input series by making the input series have a residual with white noise series. Then filtering is done in the output sequence. So we can get the cross-correlation function (CCF). This CCF is used to guess orders from models b, s, and r. The best model is obtained to predict rainfall with the transfer function, i.e. b = 0, s = 0, r = 0, p = 0, q = [1,4], P = 0, Q = 1 and S = 12. This means that humidity affects rainfall at the same time. This transfer function model has a residual that meets the white noise assumption and is normally distributed. The following is the model obtained:

$$Y_t = Y_{t-12} + 6.43X_t - 6.43X_{t-12} + a_t + 0.22a_{t-1} + 0.16a_{t-4} - 0.79a_{t-12} - 0.17a_{t-13} - 0.13a_{t-16} \tag{8}$$

The formation of the FFNN model with a causal approach is based on the transfer function model, especially in determining FFNN inputs. Y and X lag of the transfer function model is used as input in FFNN. Based on equation (8) the input in forecasting rainfall with FFNN is lag 12 of rainfall, humidity and lag 12 of humidity. The FFNN model in this study uses the tanh activation function, 1 hidden layer and tried 1 to 5 neurons. In addition, 5 optimization algorithms were also tried, namely backprop, rprop +, rprop-, sag, and slr. After that, it will be compared to get the best FFNN architecture in predicting rainfall.

Based on cross-validation and information criteria, the same best model is obtained, with the number of neurons 1 and backprop algorithm. Figure 5 shows the optimum FFNN architecture with a causal approach. The best model of FFNN with a causal approach is called FFNN3.

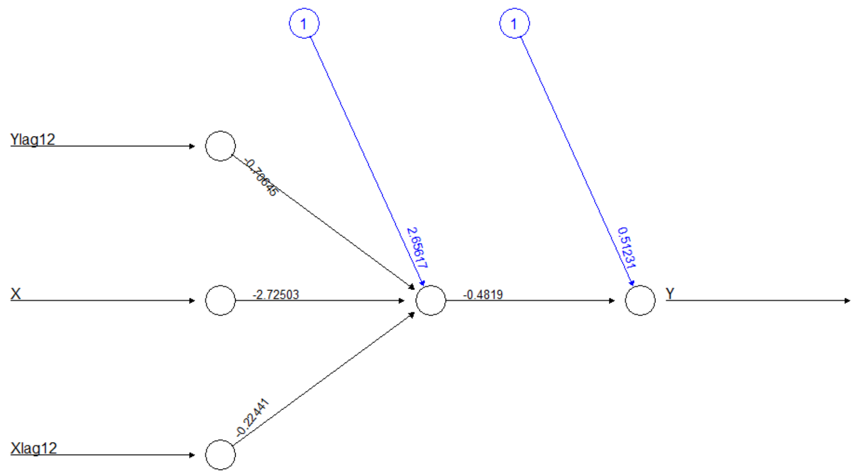

Figure 5. Optimum FFNN architecture with the causal approach

3.5 The Best Method for Forecasting Rainfall

After obtaining several best models with time series and causal approach. The results will be compared with cross-validation strategies (RMSEP and MAPEP) and information criteria (AIC and BIC).

Tabel 2. Comparison of the best method

Method	RMSEP	MAPEP	AIC	BIC
Time Series Approach				
TSR	74.012	91.551	2042.013	2095.781
ARIMA	77.619	112.551	1977.095	1990.383
FFNN 1	46.160	**30.958**	1608.301	1757.314
FFNN 2	166.403	199.422	**1380.925**	**1529.937**
DFFNN1	**33.332**	46.239	1498.659	1647.672
DFFNN2	93.419	97.047	1413.782	1562.794
Causal Approach				
RL	91.68749	128.915	2106.2	2115.161
FT	76.35959	109.2215	1913.958	1931.676
FFNN3	63.53907	65.98651	1891.749	1918.325

From Table 2 it can be seen that the best method for predicting rainfall is different based on cross-validation and information criteria. Based on the cross-validation criteria the best method is FFNN1 and DFFNN1. Based on the information criteria the best method is FFNN2. FFNN2 has a small AIC and BIC but the RMSEP and MAPEP values are quite large. So if based on a combination of cross-validation and information criteria, the best method obtained is DFFNN1. This shows that the data has a nonlinear pattern. In general, the time-series approach with the machine learning method produces better accuracy than the causal approach using statistical methods or machine learning. It happens because the data tends to be more affected by rainfall lag than humidity. Although visually in Figure 1 shows that there is a nonlinear relationship between rainfall and humidity.

The best method used to predict rainfall is the machine learning method i.e. DFFNN1. In general, in this study machine learning methods have better accuracy than statistical methods. These results are in line with research that conducted by Chattopadhyay and Chattopadhyay [10] and Xiang et al. [16]. But this result is not in line with the result of M3 and M4 forecasting competition, i.e. more complex methods on average tend to produce more accurate forecast than simpler methods [17,18]. Moreover, M3 and M4 forecasting competition are big events for forecasting researchers in the world to do a competition to find the best forecasting method in many practical problems. Finally, the forecast values at the out-of-sample dataset by using the best method and the actual data are shown in Figure 6.

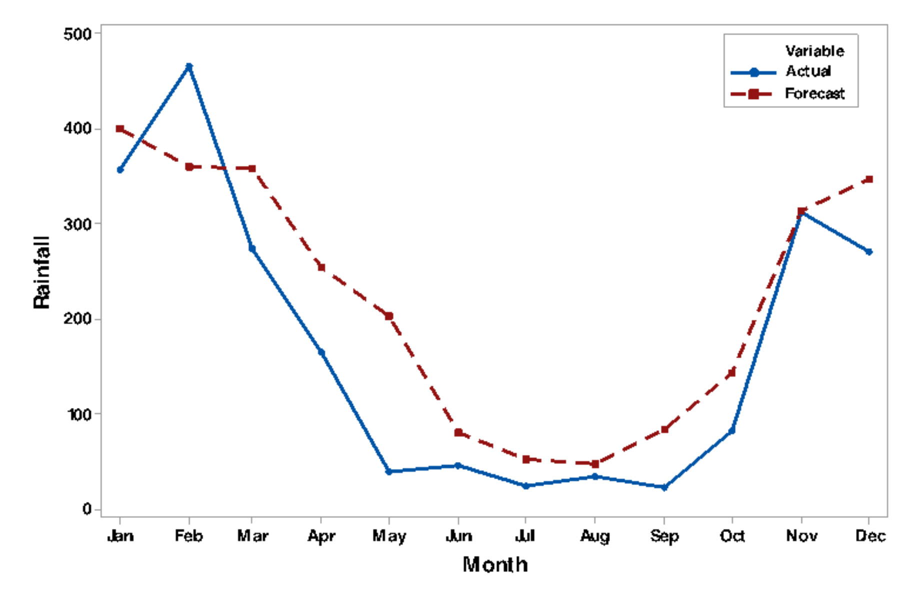

Figure 6. Comparison between forecast and actual data

4. Conclusion

In this research, time series and causal approaches using statistical methods and machine learning will be compared to predict rainfall in the Wonorejo reservoir in Surabaya. The results show that DFFNN using the time series approach is the best method in this research. In general, the machine learning method used has better accuracy than the statistical method. This result is not in line with the results of M3 and M4 forecasting competition, i.e. more complex methods on average tend to produce more accurate estimates than simpler methods [17,18]. These results indicate that the data has a nonlinear pattern and is greatly influenced by rainfall lag compared to humidity. Because it has a nonlinear pattern, then in future studies other nonlinear methods such as long short-term memory (LSTM) and support vector regression (SVR) can be used. Hybrid methods also can be used by combining linear and nonlinear models to get better forecast results.

Author Contributions: Muhaimin is the one who writes this paper and did the analysis of the linear and non-linear method. The concept of this paper comes from Suhartono. Moreover, Prabowo and Maulida also help to analyst and running the experiment.

Funding: This research received no external funding.

Data Availability Statement: The data is gathered by a weather agency located in Surabaya Indonesia. The access to this data can be checked on dataonline.bmkg.go.id/akses_data

Acknowledgment: Thanks to everyone who always be there for me, especially my wife.

Conflict of Interest: The authors declare no conflict of interest.

References

1. Lux, K. C., Ball, J. E., & Sharma, A. An Application of Artificial Neural Networks For Rainfall Forecasting. *Mathematics Computation Model 2001, Vol. 33*, 683-693.
2. Zeynoddin, M., Bonakdari, H., Azari, A., Ebtehaj, I., Gharabagi, B., & Madavar, H. R. Novel Hybrid Linear Stochastic with Non-Linear Extreme Learning Machine Methods for Forecasting Monthly Rainfall a Tropical Climate. *Journal of Environmental Management 2018 Vol. 222*, 190-206.
3. Hanke, J. E., & Wichern, D. W. *Bussines Forecasting Eight Edition.* Pearson Practice Hall: New Jersey, 2005.
4. Shummway, R. H., & Stoffer, D. S. *Time Series Analysis and Its Application with R Examples.* Springer: Pittsburg, 2006.

5. Robles , R. A., Ortega, J. C., Fu, J. S., Reed, G. D., Chow, J. C., Watson, J. G., & Herrera, J. A. A Hybrid ARIMA and Artificial Neural Network Model to Forecast Particulate Matter in Urban Areas: The Case of Temuco, Chile. *Journal Atmosphere Environment 2008 Vol. 42*, 8331-8440.
6. Cheng, Y., Zhang, H., Liu, Z., Chen, L., & Wang, P. Hybrid Algorithm for Short-Term Forecasting of PM2.5 in China. *Atmospheric Environment 2019 Vol. 200*, 264-279.
7. Wei, W. W. *Time Series Analysis Univariate and Multivariate Methods (2nd ed).* Pearson Education, Inc: the United States of America, 2006.
8. Tealab, A. Time Series Forecasting using Artificial Neural Networks Methodologies: A Systematic Review. *Future Computing and Informatics Journal 2018 Vol. 3 (2)*, 334-340.
9. Arumugam, P., & Saranya, R. Outlier Detection and Missing Value in Seasonal ARIMA Model Using Rainfall Data. *Materials Today: Proceedings 2018 Vol. 5*, 1791-1799.
10. Chattopadhyay, S., & Chattopadhyay, G. Univariate modeling of summer-monsoon rainfall time series: Comparison between ARIMA and ARNN. *Comptes Rendus Geoscience 2010 Vol. 342*, 100-107.
11. Yu, P. S., Tao, C. Y., Szu, Y. C., Chen, M. K., & Hung, W. T. Comparison of Random Forests and Support Vector Machine for Real-Time Radar-Derived Rainfall Forecasting. *Journal of Hydrology 2017 Vol. 552*, 92-104.
12. I. Kaushik, S. Madhvi Singh, Seasonal ARIMA model for forecasting of monthly rainfall and temperature, J.Environ. Res. Dev. 3 (2008) 506–514.
13. Graham, A., Mishra E.P. 2017. Time series analysis model to forecast rainfall for Allahabad region. Journal of Pharmacognosy and Phytochemistry 2017; 6(5): 1418-1421
14. Farajzadeh, J., Fard, A.M., Lotfi, S. Modeling of monthly rainfall and runoff of Urmia lake basin using “feed-forward neural network” and “time series analysis” model. Water Resources and Industry 7-8 (2014) 38–48
15. Suhartono. New Procedures for Model Selection in Feedforward Neural Networks. *Jurnal Ilmu Dasar 2008 Vol. 9*, 104-113.
16. Farraway, J., & Chatfield, C. Time Series Forecasting with Neural Network: A Comparative Study Using The Airline Data. *Applied Statistics 1998 Vol. 47*, 231-250.
17. Xiang, Y., Gou, L., He, L., Xia, S., & Wang, W. A SVR-ANN Combined Model-Based on Ensemble EMD for Rainfall Prediction. *Applied Soft Computing 2018 Vol. 73*, 874-883.
18. Makridakis, S., & Hibbon, M. The M3-Competition Result, Conclusions and Implications. *International Journal of Forecasting 2000 Vol. 16*, 451-676.
19. Makridakis, S., Spiliotis, E., & Assimakopoulus, V. The M4 Competition: Results, Findings, Conclusion and Way Forward. *International Journal of Forecasting 2018 Vol. 34*, 802-808.