International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 6, pp 142-148, June-2025

https://doi.org/10.47001/IRJIET/2025.906018

Missing Data Imputation for LSTM-Based Flood Early Warning System in Jakarta

[1]*Akmarina Khairunnisa, [2]Bagus Sartono, [3]Muhammad Nur Aidi

[1,2,3]Department of Statistics and Data Science, IPB University, Bogor, Indonesia

*Corresponding Author's E-mail: akmarinakkhairunnisa@apps.ipb.ac.id

***Abstract* - The technological advancements of data storage capacity and computational capabilities have implications for the recording of time series data with increasingly narrow intervals, called high-frequency time series data. Sensor data, as a prominent example of high-frequency time series generated through the utilization of the Internet of Things (IoT), is susceptible to issues related to missing data due to the likelihood of device failures. Furthermore, both the quantity and quality of data significantly impact the performance of forecasting models. This study examines the effects of imputing missing data within a forecasting workflow for sensor data that records water levels at four observation sites. The analysis will be conducted by evaluating 6 imputation methods in a simulation study using 10 datasets with 18 missing scenarios each. The forecasting outcomes of the IMV-LSTM (Interpretable Multi Variable Long Short-Term Memory) model, trained using empirical data reconstructed through the best imputation methods from the simulation study, will also be evaluated. The results indicate that the imputed data using the Kalman-Structural method enhances forecast accuracy, evidenced by a 32% reduction in RMSE compared to the model trained on data without imputation treatment as the benchmark. Additionally, imputed data employing Kalman-ARIMA improves the performance of the IMV-LSTM model, yielding a 29% lower RMSE compared to the benchmark. The best-performing model demonstrates that the forecasts of water levels deviate by only approximately 0.1% from the actual data.**

Keywords: Missing Data, High-Frequency Time Series, River Flood, Early Warning, IMV-LSTM.

I. INTRODUCTION

Hydrological factors have become a key concern in urban planning, particularly in densely populated areas [1] where there is limited groundwater recharge. One of the key factors is precipitation, which has a positive impact on flood vulnerability [2]. Typically, rainwater in urban areas is channeled through drainage systems (gutters) and eventually flows into rivers [3]. According to the population density projections by the Central Statistics Agency, Jakarta is the most densely populated city in Indonesia. This trend has consistently increased since 2000 and is expected to continue growing in the future. The largest contribution to the potential flood events in Jakarta is the Ciliwung River, which has the largest and longest watershed in Jakarta [4].

The Jakarta Flood Early Warning System (J-FEWS) is a tool utilized by the Jakarta government to predict flood events and their inundation areas within the Jakarta region [5]. J-FEWS integrates observational data on rainfall, water levels, and weather forecasts to generate water level predictions using the rainfall-runoff method. According to the J-FEWS bulletin, the accuracy of water level forecasts in J-FEWS is highly dependent on the precision of hydrometeorological observations, the quality of simulation models, and the accuracy of weather forecasts. Efforts to enhance the accuracy of J-FEWS have included improving radar-based rainfall measurements, updating simulation models, and incorporating high-resolution weather forecast datasets. J-FEWS employs the Delft-FEWS platform to integrate hydrometeorological observations, hydrological models, and weather forecasts, thereby facilitating the generation of flood forecasts.

River water level data is crucial for Jakarta's flood early warning system, as overflowing rivers are a major cause of flooding in the city. The water level sensor data on the Ciliwung River is collected every 10 minutes from sensors at various monitoring posts, but this data is often missing due to inconsistent data transmission. Missing data is a common issue in sensor data recording, often caused by device malfunction, communication failure, power outages, or other disruptions [6], [7]. Studies have shown that missing data in high-frequency time series can significantly impair forecasting performance if not properly addressed [8], [9]. Therefore, reconstruction of missing data is required to enhance the performance and accuracy of water level forecasting.

Imputation techniques offer a robust solution for reconstructing missing data, as demonstrated by studies such as [10], [11], and [12]. Key factors in selecting an imputation technique include convergence time, sensitivity to outliers, and the method's ability to account for autocorrelation [13].

Including an imputation stage within the forecasting workflow has been shown to improve forecast accuracy compared to workflows that do not address missing data [14], [15]. However, studies examining the impact of imputation on LSTM forecasting performance have predominantly focused on low-frequency time series.

Studies comparing forecasting methods have concluded that models based on Long Short-Term Memory (LSTM) networks outperform others in terms of both performance and forecast accuracy [16], [17]. LSTM networks excel in time series forecasting because they mitigate the exploding and vanishing gradient problems, enabling the effective learning of long-term dependencies in the data [8], [18]. Research [17] further supports the conclusion that LSTM-based models are particularly effective for forecasting high-frequency time series.

This study aims to evaluate imputation methods within the forecasting workflow for sensor data (high-frequency time series) by assessing the quality of reconstructed missing data and the forecasting performance. Three groups of imputation methods are compared: univariate methods (Kalman-Structural and Kalman-ARIMA), multivariate-global methods (SVDImpute and PPCA), and multivariate-local methods (kNN and MICE). Interpretable Multi-Variable LSTM (IMV-LSTM) as the forecast model [19] were developed to enable the simultaneous processing of multiple time series. IMV-LSTM is claimed to produce more accurate forecasts and offer greater interpretability compared to standard LSTM models. The optimal IMV-LSTM model is proposed as a potential alternative to J-FEWS. This model is designed to function as a flood early warning system for Jakarta, relying solely on historical water level data. In contrast to J-FEWS—a government-implemented system that integrates multiple types of projection data—this study focuses exclusively on past water level records to develop a forecasting approach.

II. METHODS & DATA

2.1 Time Series Missing Data Imputation

High-frequency time series data can exhibit various patterns of missingness, including large gaps, small gaps, or individual missing observations occurring randomly [20], [21]. According to [22], as cited by [23], missing data mechanisms are categorized into three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). In this study, three groups of imputation methods are compared: univariate, multivariate-global, and multivariate-local methods.

Univariate imputation methods assume that a model can be derived from the data to estimate the missing values. The univariate methods for this study are the Kalman Smoothing method [24], [25] applied to a structural time series model [26] and the state-space representation of the ARIMA model [27], [28]. Multivariate-global methods utilize the global structure of the data in matrix form, with estimation performed through matrix completion algorithms. This multivariate-global methods are SVD Impute [11], which uses Singular Value Decomposition (SVD), and PPCA (Probability PCA), based on Principal Component Analysis (PCA). Multivariate-local methods operate on the assumption of high similarity among series. Two methods for this category are employed: Multiple Imputation by Chained Equations (MICE) [29], [30] and k Nearest Neighbors (kNN) [11]. MICE iteratively impute missing values by modeling each variable based on the others, while kNN estimates missing values using the nearest observations to the missing data points.

2.2 Interpretable Multi-Variable Long Short-Term Memory (IMV-LSTM)

The Interpretable Multi-Variable Long Short-Term Memory (IMV-LSTM) model [19] is capable of simultaneously processing multiple time series. IMV-LSTM is reported to deliver more accurate forecasts and greater interpretability compared to standard LSTM models. Its enhancements include optimizing the LSTM structure by assigning separate LSTM processes for each input variable, ensuring that hidden states and model parameters are uniquely associated with specific variables. IMV-LSTM captures the dynamics of input time series variables and quantifies the contribution of each variable to the forecast through a mixture attention mechanism.

2.3 Data Source

Ten complete datasets for the simulation study are categorized based on their data sources or methods of acquisition, those are: aggregated empirical data, secondary data, and data generated using the Vector Autoregression (VAR) model. All datasets meet the characteristics required for the simulation study, specifically multivariate time-series data without missing values. The data removal scenarios shown in Figure 1 were applied to these ten datasets, resulting in 180 incomplete datasets (containing missing values). The random removal of data was repeated five times, generating a total of 900 datasets to be used in the simulation study for each method.

The empirical data used in this study is water level sensor data collected from the Ciliwung River with a recording frequency of every 10 minutes. The data was recorded and transmitted by IoT sensors operated by the Jakarta Water Resources Agency. Figure 2 provides a map showing the four observation locations used in this study, namely: the Cibalok-

Gadog weir, the Katulampa weir, the Depok monitoring station, and the Manggarai floodgate. The empirical data contains 5% missing values spread across the four location variables. The timeframe for the empirical data used in this study is from November 17, 2022, to December 31, 2023.

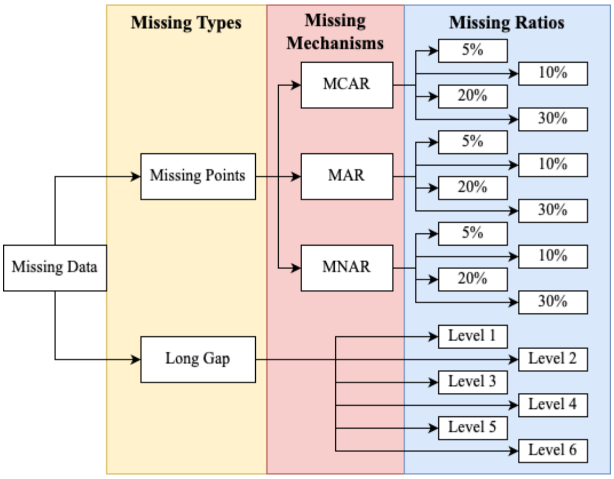

Figure 1: Data Removal Scenarios for Simulation Studies

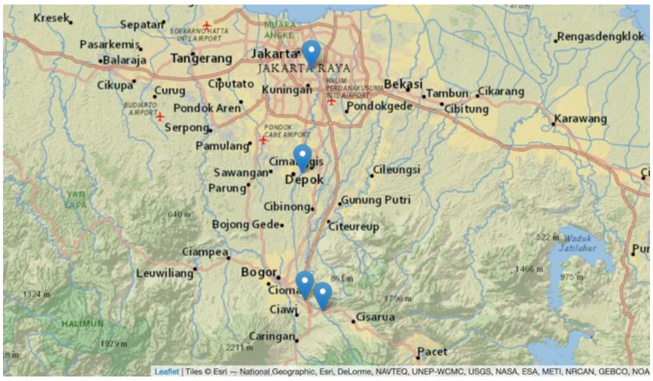

Figure 2: Observation Points for Water Level Monitoring in the Ciliwung Watershed

2.4 Research Methodology

This study consists of a simulation study and an empirical study. The simulation study evaluates methods for handling missing data in simulated datasets, while the empirical study compares the performance of these methods based on forecasting accuracy using empirical data. The primary objective is to identify the best imputation method that produces the highest forecasting accuracy when applied to the IMV-LSTM model. The best imputation method, as determined by the empirical study, will then be used to process the entire dataset, followed by model interpretation based on the optimal model at the final stage. The overall research workflow is illustrated in Figure 3.

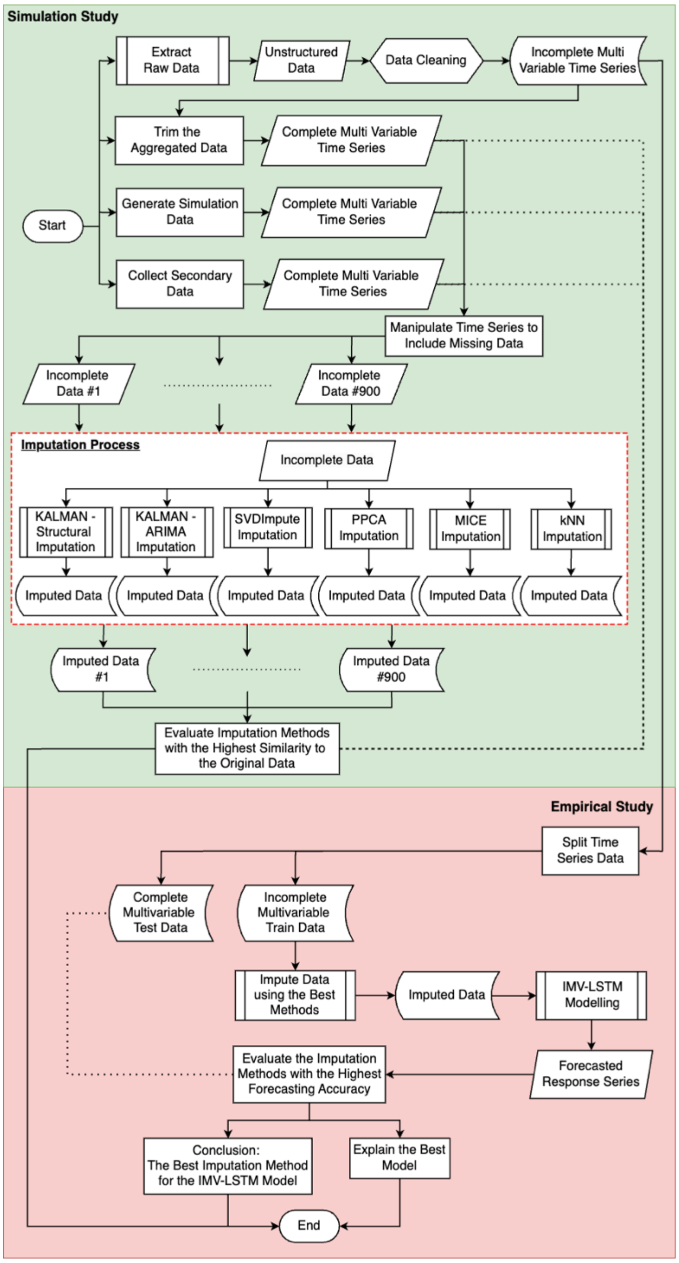
Simulation Study
Start
Extract Raw Data
Unstructured Data
Data Cleaning
Incomplete Multi Variable Time Series
Trim the Aggregated Data
Complete Multi Variable Time Series
Generate Simulation Data
Complete Multi Variable Time Series
Collect Secondary Data
Complete Multi Variable Time Series
Manipulate Time Series to Include Missing Data
Incomplete Data #1
Incomplete Data #900
Imputation Process
Incomplete Data
KALMAN - Structural Imputation
KALMAN - ARIMA Imputation
SVDImpute Imputation
PPCA Imputation
MICE Imputation
kNN Imputation
Imputed Data
Imputed Data
Imputed Data
Imputed Data
Imputed Data
Imputed Data
Imputed Data #1
Imputed Data #900
Evaluate Imputation Methods with the Highest Similarity to the Original Data
Empirical Study
Split Time Series Data
Complete Multivariable Test Data
Incomplete Multivariable Train Data
Impute Data using the Best Methods
Imputed Data
IMV-LSTM Modelling
Forecasted Response Series
Evaluate the Imputation Methods with the Highest Forecasting Accuracy
Conclusion: The Best Imputation Method for the IMV-LSTM Model
Explain the Best Model
End

Figure 3: Research Workflow

III. RESULTS AND DISCUSSION

3.1 Imputation as the Pre-Forecasting Stage

Two simple imputation techniques, Moving Average (MA) and Last Observation Carried Forward (LOCF), were added to the simulation study as benchmarks. The evaluation of imputation methods is conducted based on the type, mechanism, and ratio of missing data. A visualization of the Root Mean Square Error (RMSE) distribution across all imputation trials, grouped by imputation method, is presented. Figures 4 and 5 shows the RMSE distribution for missing points, evaluated by the missing data mechanism and ratio. Figure 6 shows the RMSE distribution for gap missing data, evaluated based on the size of the gap.

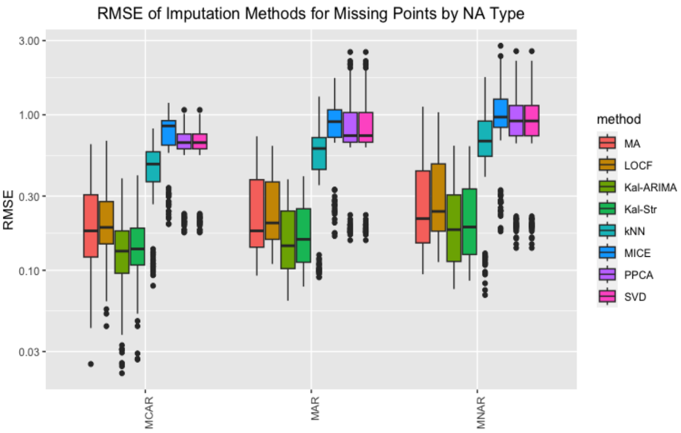

Figure 4: RMSEs Distribution of Imputation Methods for Missing Points by Missing Mechanisms

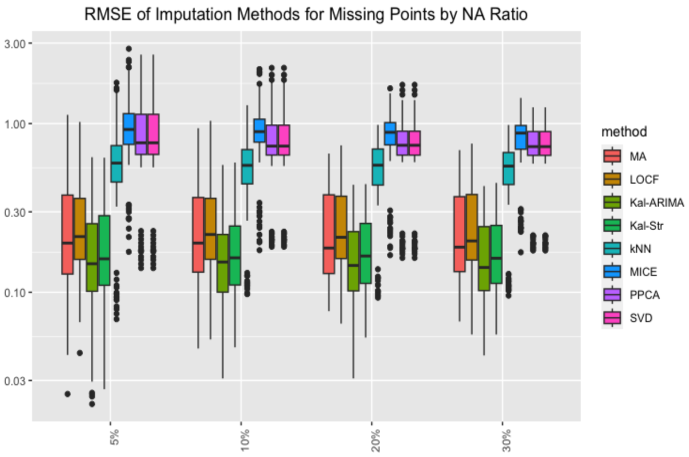

Figure 5: RMSEs Distribution of Imputation Methods for Missing Points by Missing Ratios

The boxplots in Figures 4 and 5 indicate that univariate methods, specifically Kalman-ARIMA and Kalman-Structural, produce better imputation results compared to other methods. Two-way analysis of variance (ANOVA) was performed to examine the effects of the missing data mechanism and ratio on the RMSE values for each method. The ANOVA results in Table 1 demonstrate significant differences in the mean RMSE among the method groups at each level of the missing data ratio and mechanism factors for point missing data.

The Tukey post-hoc test supports the finding that univariate methods yield imputations that most closely match the actual values compared to other methods. This finding is particularly applicable to high-frequency time-series data with point missing data. The superior performance of univariate methods suggests that the best imputation methods are those that account for time-series patterns when predicting missing values. The absence of significant differences between the Kalman-Structural and Kalman-ARIMA methods across all levels of the missing data ratio and mechanism factors indicates that both methods perform equally well.

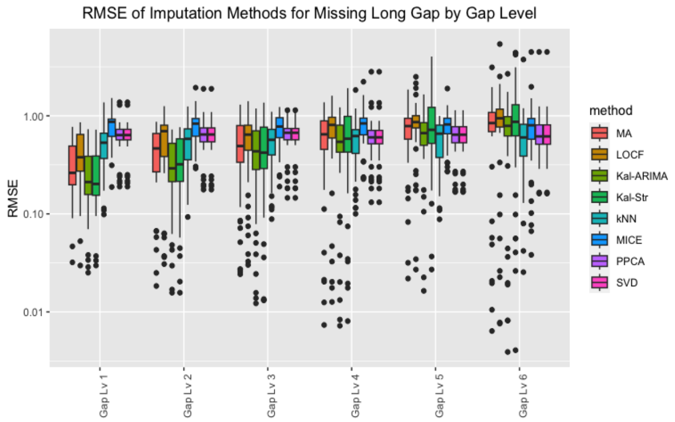

Figure 6: RMSEs Distribution of Imputation Methods for Missing Gaps

The boxplot in Figure 6 reveals that all imputation methods struggled to correctly fill missing gap, as the number of experiments yielding RMSE > 1 increased with larger gap sizes. Visual differentiation among method groups was challenging; therefore, ANOVA results in Table 1 were used to identify superior methods based on Figure 6. The analysis indicated significant differences in average RMSE among imputation method groups at each gap size level.

Table 1: Two-Way Analysis of Variance (ANOVA) of RMSE in the Simulation Study

Source of Variation	df*	SS*	MS*	F*	*p value*
Missing Type: **Missing Point**					
Methods	7	410.5	58.64	940.492	< 0.0001 **
Missing Mechanisms	2	30.1	15.03	240.985	< 0.0001 **
Missing Ratios	3	6.0	1.99	31.902	< 0.0001 **
Methods × Missing Mechanisms	14	7.9	0.56	9.012	< 0.0001 **
Methods × Missing Ratios	21	3.4	0.16	2.633	< 0.0001 **
Residual	4744	295.8	0.06	-	-
Missing Type: **Missing Gap**					
Methods	7	15.9	2.275	14.239	< 0.0001 **
Missing Ratios	5	34.0	6.799	42.554	< 0.0001 **
Methods × Missing Ratios	35	29.1	0.832	5.208	< 0.0001 **
Residual	2349	375.3	0.160	-	-

*) df: degree of freedom; SS: Sum of Square; MS: Mean of Square; F: Test Statistics F

**) significant

3.2 Water Level Forecast after Missing Data Imputation

Table 2 presents the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for evaluating model performance by comparing forecast results with test (actual) data. The models were differentiated by the preprocessing applied, including training data without imputation and training data with imputation using Kalman-Structural and Kalman-ARIMA methods. Relative improvement percentages were calculated using:

$$\Delta\text{Eval}(\%) = \frac{E_0 - E_1}{E_0} \times 100$$

Where E_0 is the metrics for the "No Imputation" treatment. The results indicate that training data imputed with Kalman-Structural and Kalman-ARIMA methods improved forecast accuracy, reducing RMSE by 32% and 30%, respectively, and reducing MAPE by 50%. Imputed data using the Kalman-Structural method resulted in an average forecast error of 1.2 cm, with the forecast plot shown in Figure 7.

Table 2: Forecast Evaluation

Evaluation Metrics	Treatment for Train Data		
	No Imputation	Kalman-Structural Imputation	Kalman-ARIMA Imputation
RMSE	1.775 cm	1.206 cm	1.243 cm
MAPE	0.210%	0.105%	0.115%

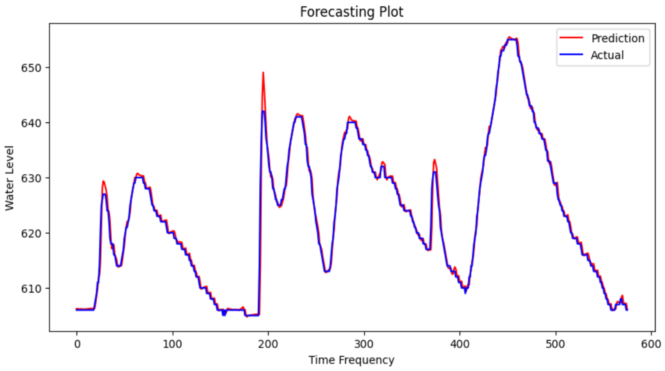

Figure 7: Plot of Manggarai Water Level Forecast

Figure 8 illustrates how the model assigns attention weights to different time steps when generating predictions. Observing the heat map patterns provides insights into which time steps contribute more significantly to the model's predictions. Brighter bands or regions in the heat map highlight segments of the input series that strongly influence prediction accuracy. For the best-performing model, it was observed that the temporal importance values are mostly uniform as the water level itself has relatively long-term autocorrelation. "Cibalok", "Katulampa", and "Manggarai" are relatively long-term correlated to the target, and short history of "Depok" variables contributes more to the forecasting.

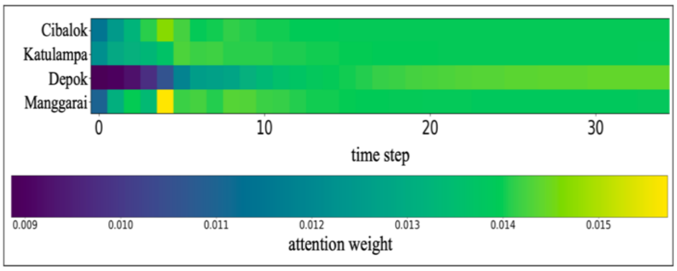

Figure 8: Attention Weights Visualization from the Best Model

IV. CONCLUSION

Based on the simulation results, the univariate method category, including the Kalman-Structural and Kalman-ARIMA methods, produced the best imputation results for high-frequency multivariate time series data. This conclusion was drawn for point missing data based on the similarity between the imputed values and the actual values, as well as the methods' performance under various missing data conditions. However, the ability of these two methods to impute gap missing data decreased as the size of the gaps increased. According to post-hoc analysis, at gap sizes corresponding to levels 5 and 6, Kalman-ARIMA outperformed Kalman-Structural.

These two best imputation methods were applied to forecasting in the empirical study, where the training data contained approximately 5% missing values. The forecasting results demonstrated that using the Kalman-Structural method for imputation during the pre-forecasting stage of water level sensor data provided the best forecasting outcomes, improving accuracy by reducing RMSE by 32% and MAPE by 50%. The best model achieved a forecasting accuracy of 0.1%, with an average error of 1.2 cm.

This study demonstrates that utilizing historical data, supported by a forecasting workflow that includes imputation during the pre-forecasting stage, is sufficient to develop an accurate flood early warning model. This model does not rely on precipitation data or other hydrometeorological information, yet it still delivers highly accurate forecasts. This solution can be considered as an alternative model for government use.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Jakarta Water Resources Service for providing the Ciliwung River water level data used in this study. The first author would like to thank the Indonesia Endowment Fund for Education (LPDP), Ministry of Finance of the Republic of Indonesia, for granting the scholarship that supported her graduate studies and made this research possible.

ISSN (online): 2581-3048
Volume 9, Issue 6, pp 142-148, June-2025
https://doi.org/10.47001/IRJIET/2025.906018

REFERENCES

[1] H. Kurdi and Novitasari, 'Evaluasi Terhadap Aspek Hidrologi pada Kawasan Rencana Pengembangan Kota di Kota Balangan', *Jurnal Teknologi Berkelanjutan (Sustainable Technology Journal)*, vol. 9, no. 2, pp. 96–109, 2020, [Online]. Available: http://jtb.ulm.ac.id/index.php/JTB

[2] M. N. Aidi, 'The Influence of Precipitation, Stream Discharge, and Physiographic Factors on Flood Vulnerability at Cimanuk River West Java, Indonesia', *J Sustain Sci Manag*, vol. 14, pp. 125–136, Feb. 2019.

[3] N. Koyama, M. Sakai, and T. Yamada, 'Study on a Water-Level-Forecast Method Based on a Time Series Analysis of Urban River Basins—A Case Study of Shibuya River Basin in Tokyo', *Water (Switzerland)*, vol. 15, no. 1, Jan. 2023, doi: 10.3390/w15010161.

[4] B. Harsoyo, 'Mengulas PenyebabBanjir di Wilayah DKI Jakarta dariSudut Pandang Geologi, Geomorfologi dan Morfometri Sungai', *Jurnal Sains & Teknolohi Modifikasi Cuaca*, vol. 14, no. 1, pp. 37–43, 2013.

[5] S. Ginting and W. M. Putuhena, 'Sistem Peringatan Dini Banjir Jakarta: Jakarta-Flood Early Warning Sytem (J-FEWS)', *Jurnal Sumber Data Air*, vol. 10, no. 1, pp. 71–84, May 2014.

[6] M. Halatchev and L. Gruenwald, 'Estimating Missing Values in Related Sensor Data Streams', 2005.

[7] R. N. Faizin, M. Riasetiawan, and A. Ashari, 'A Review of Missing Sensor Data Imputation Methods', in *5th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia, Jul. 2019.

[8] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, 'LSTM-based traffic flow prediction with missing data', *Neurocomputing*, vol. 318, pp. 297–305, Nov. 2018, doi: 10.1016/j.neucom.2018.08.067.

[9] G. Chang and T. Ge, 'Comparison of Missing Data Imputation Methods for Traffic Flow', in *International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, IEEE, 2011, pp. 639–642.

[10] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page, 'A Survey on Data Imputation Techniques: Water Distribution System as a Use Case', *IEEE Access*, vol. 6, pp. 63279–63291, 2018, doi: 10.1109/ACCESS.2018.2877269.

[11] O. Troyanskaya *et al.*, 'Missing value estimation methods for DNA microarrays', 2001. [Online]. Available: http://smi-web.

[12] T. Schneider, '853 Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values', 2001.

[13] M. K. Hasan, M. A. Alam, S. Roy, A. Dutta, M. T. Jawad, and S. Das, 'Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)', Jan. 01, 2021, *Elsevier Ltd*. doi: 10.1016/j.imu.2021.100799.

[14] M. K. Gill, T. Asefa, Y. Kaheil, and M. McKee, 'Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique', *Water Resour Res*, vol. 43, no. 7, Jul. 2007, doi: 10.1029/2006WR005298.

[15] J. H. Yang, C. H. Cheng, and C. P. Chan, 'A time-series water level forecasting model based on imputation and variable selection method', *Comput Intell Neuro sci*, vol. 2017, 2017, doi: 10.1155/2017/8734214.

[16] D. Kumar, A. Singh, P. Samui, and R. K. Jha, 'Forecasting monthly precipitation using sequential modelling', *Hydrological Sciences Journal*, vol. 64, no. 6, pp. 690–700, Apr. 2019, doi: 10.1080/02626667.2019.1595624.

[17] Z. Li, J. Han, and Y. Song, 'On the forecasting of high-frequency financial time series based on ARIMA model improved by deep learning', *J Forecast*, vol. 39, no. 7, pp. 1081–1097, Nov. 2020, doi: 10.1002/for.2677.

[18] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] T. Guo, T. Lin, and N. Antulov-Fantulin, 'Exploring Interpretable LSTM Neural Networks over Multi-Variable Data', in *36th International Conference on Machine Learning*, May 2019. [Online]. Available: http://arxiv.org/abs/1905.12034

[20] H. M. Ahmed, B. Abdulrazak, F. Guillaume Blanchet, H. Aloulou, and M. Mokhtari, 'Long Gaps Missing IoT Sensors Time Series Data Imputation: A Bayesian Gaussian Approach', *IEEE Access*, vol. 10, pp. 116107–116119, 2022, doi: 10.1109/ACCESS.2022.3218785.

[21] J. Park *et al.*, 'Long-term missing value imputation for time series data using deep neural networks', *Neural Comput Appl*, Apr. 2022, doi: 10.1007/s00521-022-08165-6.

[22] D. B. Rubin, 'Inference and Missing Data', *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[23] Y. Zhang and P. J. Thorburn, 'Handling missing data in near real-time environmental monitoring: A system and a review of selected methods', *Future Generation Computer Systems*, vol. 128, pp. 63–72, Mar. 2022, doi: 10.1016/j.future.2021.09.033.

[24] R. E. Kalman, 'A New Approach to Linear Filtering and Prediction Problems', *Journal of Basic Engineering*, vol. 82, no. 1, 1960.

[25] M. S. Grewal, ‘Kalman Filtering’, in *International Encyclopedia of Statistical Science*, Lovric, M., Berlin, Heidelberg: Springer, 2011, pp. 705–708.

[26] J. T. Jalles, ‘Structural Time Series Models and the Kalman Filter: A Concise Review’, 2009, [Online]. Available: http://ssrn.com/abstract=1496864at:https://ssrn.com/abstract=1496864Electroniccopyavailableat:http://ssrn.com/abstract=1496864

[27] P. De Jong and J. Penzer, ‘The ARIMA model in state space form’, 2000.

[28] J. C. Abril, ‘Structural Time Series Models’, in *International Encyclopedia of Statistical Science*, Lovric, M., Berlin, Heidelberg: Springer, 2011, pp. 1555–1558.

[29] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, 1997.

[30] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. 2002.

Citation of this Article:

Akmarina Khairunnisa, Bagus Sartono, & Muhammad Nur Aidi. (2025). Missing Data Imputation for LSTM-Based Flood Early Warning System in Jakarta. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(6), 142-148. Article DOI https://doi.org/10.47001/IRJIET/2025.906018

****