E3S Web of Conferences **680**, 00067 (2025) https://doi.org/10.1051/e3sconf/202568000067
ICEGC'2025

PM2.5 INDEX PREDICTION IN CENTRAL JAKARTA USING POLLUTION AND METEOROLOGICAL DATA WITH CNN-LSTM

Stevanus Febrian[1*], and *Syarifah Diana* Permai[2*]

[1] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480

[2] Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480

Abstract. Jakarta is among the most polluted cities globally, with $PM_{2.5}$ posing the greatest health risks. Accurate prediction of $PM_{2.5}$ levels is therefore essential to support early warning systems and public health decisions. This study develops a hybrid deep learning model, CNN-LSTM, to forecast daily $PM_{2.5}$ concentrations in Central Jakarta for the next seven days. The dataset combines air pollutant records from Satu Data Jakarta and meteorological data from BMKG, covering January 2023 to December 2024. Data preparation included cleaning, handling missing values, outlier analysis, feature selection using Pearson correlation and linear regression, and normalization. Model performance was evaluated using MSE, MAE, RMSE, and MAPE, and compared against standalone CNN and LSTM models under two data configurations: pollutant-only and pollutant & meteorology. Results show that CNN-LSTM achieved the lowest MSE and RMSE when using combined data, indicating high predictive accuracy. However, the LSTM model with pollutant-only input provided the most consistent and efficient results, with lower error values across all metrics. These findings suggest that historical pollutant patterns alone are sufficient for short-term $PM_{2.5}$ forecasting, while meteorological factors offer limited additional benefit. The study demonstrates the potential of deep learning in supporting air quality monitoring and public health protection.

1 Introduction

Jakarta is ranked as one of the most polluted city in the world [1] and holds the highest air pollution levels compared to other cities in Indonesia by Jakarta Environmental Service [2]. This condition is driven by human activities, particularly the increase of personal transportation usage and industrial operations, which contribute to the increasing the pollution levels in Jakarta [3]. These pollutions include carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM_{10} and $PM_{2.5}$). $PM_{2.5}$ poses the greatest threat to human health [4]. Due to their small size, $PM_{2.5}$ particles can penetrate deep into the respiratory system, causing serious health problems including cancer, cardiovascular diseases, stunting, and acute respiratory infections [5]. Prolonged exposure above safe thresholds not only endangers public health but also imposes substantial economic losses. In 2019, the health-related economic burden of air pollution in Jakarta was estimated at USD 2,943.42 millions, equivalent to 2.2% of the province's GDP.

Meteorological factors have been shown to influence $PM_{2.5}$ concentrations in various ways [6]. Low rainfall is generally associated with higher $PM_{2.5}$ levels, while increased precipitation tends to reduce them. Higher temperatures and longer sunshine durations

* Corresponding author: stevanus.febrian@binus.ac.id, syarifah.permai@binus.ac.id

 https://doi.org/10.1051/e3sconf/202568000067

correlate with elevated $PM_{2.5}$ concentrations, whereas low wind speeds can exacerbate particle accumulation. Conversely, higher humidity, that often associated with heavy rainfall, can help lower $PM_{2.5}$ levels. These relationships suggest that incorporating meteorological data into predictive models could improve the accuracy of $PM_{2.5}$ forecasting [7].

However, there aren't many studies in Jakarta that consider meteorological factors to improve $PM_{2.5}$ prediction. Most existing studies in Jakarta solely use air pollution data without integrating weather-related parameters such as rainfall, temperature, humidity, wind speed and direction, and sunshine duration, which have been shown to influence $PM_{2.5}$ concentrations. To address this gap, this study develop a deep learning forecasting model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to predict $PM_{2.5}$ concentrations in Central Jakarta for seven days ahead. The model incorporates both air pollution data from Satu Data Jakarta and meteorological data from Data Online Meteorology, Climatology, and Geophysical Agency (BMKG).

The development of a $PM_{2.5}$ prediction model for Central Jakarta offers significant benefits for both the public and policymakers. For the community, it can serve as an early warning system, providing seven-day forecasts of $PM_{2.5}$ concentrations to help residents plan activities and minimize exposure to harmful air pollutants. For the Jakarta government, the model can support evidence-based decision-making, enabling the formulation of targeted mitigation strategies to improve air quality. Effective policies informed by accurate predictions may also help reduce the economic losses associated with pollution-related health impacts.

2 Related Works

Several studies have been conducted to employ CNN-LSTM model using pollution and meteorological. A study conducted in Qingdao, China [8], employed a CNN-LSTM architecture to predict $PM_{2.5}$ concentrations by leveraging both air quality and meteorological data. The dataset consists of PM2.5 measurements from 20 monitoring stations throughout 2020, along with temperature, wind speed, and air pressure variables. Model performance was benchmarked against standalone CNN and LSTM models, with the CNN-LSTM achieving the highest accuracy, obtaining an RMSE of 8.216 µg/m³, and MAPE of 38.75%. In comparison, the CNN and LSTM models recorded higher RMSE values of 11.356 µg/m³ and 14.367 µg/m³, and significantly larger MAPE values of 85.36% and 54.21%, respectively.

In another research, D. Li, Liu, & Zhao [9] came up with proposed model of CBAM-CNN-LSTM and compares it with CNN, LSTM, BiLSTM, and CNN-LSTM for forecasting hourly PM2.5 concentration at multiple sites in Beijing, China using hourly pollutant data ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3) and hourly meteorological data (temperature, pressure, dew point, precipitation, wind direction, and wind speed). While its CBAM-CNN-LSTM model has the lowest prediction error in long-term prediction (h13 – h18), it also shows that CNN-LSTM has the second lowest prediction error, then followed by BiLSTM, LSTM, and CNN, with RMSE of 37.33, 42.84, 48.84, 49.71, and 58.79, respectively. And then has MAE of 26.54, 30.41, 34.89, 34.87, and 41.53, respectively.

3 Methodology

The proposed methodology in this research consists of data gathering, data preparation, and modelling and evaluation as shown in the figure 1.

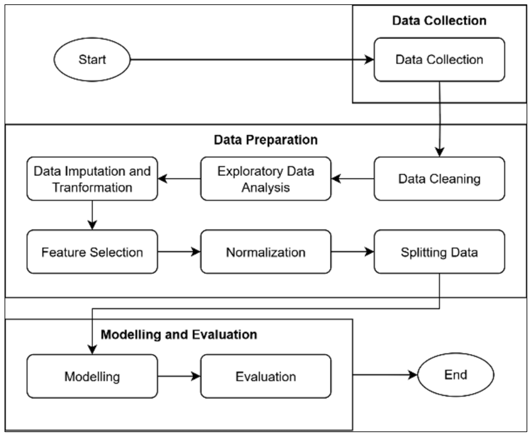

Fig. 1. Proposed methodology

3.1 Data Collection

There are two data used for this research, pollutant and meteorological data in Central jakarta. For pollutant data was collected from Satu Data Jakarta [10] website where the sensor is located in latitude and longitude of -6.194587 and 106.823600, respectively. The data include daily index of $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3. And for the meteorological data was collected from Data Online Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) [11] where the sensor is located in latitude and longitude of -6.156000 and 106.842000, respectively. And the data include average temperature (°C), relative humidity (%), rainfall intensity (mm), sunshine duration (hour), average wind speed (m/s), and most wind direction (°). Both datasets were collected from 1 January 2023 to 31 December 2024, consisting of 731 data samples in total.

3.2 Data Preparation

Steps in data preparation includes data cleaning, exploratory data analysis, imputation and transformation, feature selection, normalization, and splitting data. In data cleaning consists of renaming variables, standardizing null values, and adjusting numerical data that was identified as string. And exploratory data analysis steps include univariate analysis to visualize distribution for numerical data and bar plot for categorical data. And bivariate analysis is done with correlation plot to identify $PM_{2.5}$ correlation to the independent variables. Also identifying and analysis of outliers was done to understand whether the outliers were part of the natural occurrences or not. Then identifying missing value counts helps to choose what imputation method can be used [12].

The imputation and transformation steps include handling missing value, transform wind direction data, combining pollutant and meteorological data. If the missing value happened for all variables in a row, linear interpolation is used, otherwise use kNN [13]. For categorical data in wind direction data, it is transformed from degree to cos and sine representation using trigonometry [14], [15]. And after cleaning the data, both pollutant and meteorological data then combined as one dataset so model can learn both data simultaneously.

Then feature selection is done with two approached, pearson correlation and linear regression test. Pearson correlation is used to measure the strength and direction of linear relationship of each independent variables to the $PM_{2.5}$ and linear regression test done to evaluate the individual influence of each independent variable on $PM_{2.5}$ through estimated regression coefficients, R^2 values, and p-values. And then normalization was performed using Min-Max Normalization to convert each value from all variable to range between 0 and 1 [16]. Then the data was splitted using sliding window technique with window and horizon equals to 7 and it got total of 718 pairs sample of 7 days of historical data and 7 days ahead for $PM_{2.5}$ value. Then the data was splitted into training, validation, and test set with percentage of 92%, 4%, and 4%, respectively.

3.3 Model Architecture and Working Principle

The CNN-LSTM predictive model in this study was designed using a sequential architecture to leverage the strengths of both convolutional and recurrent neural networks. The one-dimensional CNN that designed to extract spatial and temporal features from the multivariate input data [17]. It applies one-dimensional convolutional filters that capture local dependencies between pollutant and meteorological variables across consecutive days. And LSTM network is then used to capture long-term temporal dependencies by maintaining information through gated memory cells (input, output, and forget gates). This allows the model to learn patterns and trends in air quality over time, even when the relationships are non-linear or delayed [18].

In the hybrid CNN-LSTM model, the convolutional layer first generates feature maps by applying filters that capture local temporal and patterns between variables in the input sequence. These feature maps preserve the temporal structure of the data and are passed to the two LSTM layers, which model long-term dependencies through memory cells and gated mechanisms that control how information is retained or discarded over time . Finally, the sequential representation learned by the LSTM layers is flattened and passed to a dense layer to produce PM2.5 predictions for the next seven days.

3.4 Evaluation

To validate the performance gain of the hybrid approach, two baseline models were also implemented: a CNN only model and an LSTM only model. The CNN only model relies solely on convolution and pooling operations, while the LSTM-only model focuses on learning temporal dependencies without convolutional feature extraction. This comparison allows the evaluation of whether integrating CNN and LSTM provides a significant improvement over single architectures. Those CNN-LSTM, CNN, and LSTM were iterated several times for hyperparameter tuning, done by exploring with difference units or filters, number of layers, epoch, and batch size.

The evaluation was conducted using four metrics: mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Models were trained and tested on two different input configurations: (1) pollutant data only and (2) pollutant data combined with meteorological variables. The test set was used to evaluate generalization performance on unseen data. For all metrics, lower values indicate better predictive performance. Then another evaluation is done statistically to determine whether the proposed model and configuration's performance has significant a difference to comparator model and configuration by running the chosen models for 10 times then fetch each evaluation metric result. The result will be processed for normality test, t-test, and Mann-Whitney test.

4 Result and Discussion

Both pollutant and meteorological dataset was cleaned and any kind of missing values is replaced with NaN in python to be analyzed. After those processes, exploratory data analysis (EDA) is done with visualization with boxplot and checking its outlier.

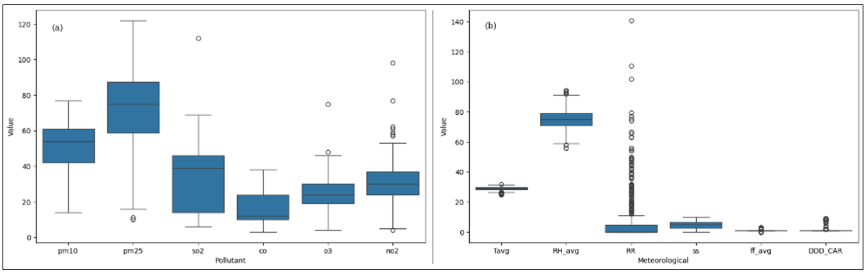

Fig. 2. Boxplot result. (a) Pollutant data. (b) Meteorological data

As shown in the Figure 2a, $PM_{2.5}$ and PM_{10} shows relatively variations also have higher medians compared toother pollutants. Meanwhile, SO_2, CO, O_3, and NO_2 display lower and more symmetric spreads, but with several outliers occurred. For meteorological data shown in the Figure 2b, RH_avg (relative humidity) shows the widest range, while other variables such as Tavg (average temperature), ff_avg (average wind speed), and DDD_CAR (dominant wind direction) tend to have narrower distributions. The RR (rainfall) variable shows many outliers. Reflecting irregular distribution due to varying rainfall intensities.

Based on the outlier analysis of air pollution data, most of the outliers that occurred can be explained by meteorological factors influencing air conditions, such as high relative humidity, low sunshine duration, calm wind direction, and the absence of rainfall. Some outliers are also linked to photochemical reactions (high NO2 and sunshine duration) increase O_3 concentrations or pollutant accumulation under specific atmospheric conditions. Or outliers occurred due to low human activity.

Outliers in Tavg (average temperature) generally occurred during the rainy or transitional seasons with high rainfall and low sunshine duration, while high temperatures outliers were caused by strong solar radiation. High outliers in RH_avg (average humidity) aligned with the rainy season, whereas low humidity outliers below 60% typically occurred in the dry season, although some cases in the rainy season were validated by meteorological data from the Tanjung Priok station. RR (rainfall) showed many outliers in the form of heavy rain, which is common in Jakarta's tropical climate. The ss (sunshine duration) showed a normal distribution without significant outliers. Meanwhile, ff_avg (average wind speed) and DDD_CAR (dominant wind direction) contained outliers, but these were more related to the limited variability in the dataset and still reflected actual conditions. Therefore, most meteorological outliers were not removed as they represented natural weather phenomena.

Then missing values for both dataset were handled individually with linear interpolation if all variables happens to be null, otherwise handled with kNN using kNN Imputer in python. And DDD_CAR was transformed from categorical direction (North, Northwest, West, etc) to circular degree, then that circular degree transformed into cos and sine representation with trigonometry. Then both dataset with combined.

Table 1. Pearson correlation and linear regression result

Features	Correlation	Pearson p-value	Coefficient	R2	Linear Regression p-value
pm10	0.9373	0.0000	89.8701	0.8786	0.0000
no2	0.5439	0.0000	92.8910	0.2959	0.0000
o3	0.4698	0.0000	80.3422	0.2207	0.0090
Tavg	0.3798	0.0000	48.7820	0.1443	0.0130

ss	0.3491	0.0000	26.6631	0.1219	0.0240
DDD_CAR_sin	0.2953	0.0000	52.6836	0.0872	0.0250
co	0.2619	0.0000	23.1870	0.0686	0.0120
DDD_CAR_cos	0.1256	0.0007	16.7197	0.0158	0.1020
so2	-0.1264	0.0006	-15.8845	0.0160	0.0000
ff_avg	-0.1799	0.0000	-18.8593	0.0324	0.7040
RR	-0.3477	0.0000	-68.1670	0.1210	0.2380
RH_avg	-0.3829	0.0000	-44.7082	0.1466	0.0210

After the data was cleaned, pearson correlation and linear regression test were done for feature selection. From Table 1, the Pearson correlation results show that PM_{10} has the strongest positive correlation with $PM_{2.5}$ ($r = 0.9373$), followed by NO_2 and O_3 with moderate correlations, while RH_avg has a moderate negative correlation; other variables such as Tavg, ss, CO, and DDD_CAR_sin component display weak but significant associations ($p < 0.05$). Linear regression further confirms PM_{10} as the dominant predictor with the highest explanatory power ($R^2 = 0.8786$, coef. = 89.87, $p < 0.001$), while NO_2 and O_3 also contribute positively but with lower R^2 values. Meteorological factors such as RH_avg (coef. = -44.71, $p = 0.021$) and Tavg (coef. = 48.78, $p = 0.013$) show significant effects, although weaker, and sunshine duration along with wind direction sine are also significant but minor contributors. In contrast, rainfall (RR), wind direction cosine, and wind speed are not significant predictors. Overall, the findings highlight that pollutant variables, especially PM_{10}, are the main determinants of $PM_{2.5}$, while meteorological factors play a secondary but supportive role in prediction and all features gonna be used for modelling.

Table 2. CNN & LSTM modelling result

Model	No	Units/Filters	Epoch	Batch Size	MSE	MAE	MAPE	RMSE
CNN (pol. Data and met. Data)	1	64	20	42	532.0422	18.0584	40.3891	23.0660
	2	128, 32	20	42	522.8632	18.3753	41.2809	22.8662
	11	**128, 256, 128**	**20**	**42**	**517.7249**	**17.3187**	**38.1526**	**22.7536**
	19	256, 128, 512	20	42	681.5029	21.8167	50.0377	26.1056
	20	512, 128	50	42	620.0514	19.4827	41.8904	24.9008
CNN (pol. Data only)	1	64	20	42	519.0563	18.6707	46.1617	22.7828
	2	128, 32	20	42	544.1181	18.2797	40.9557	23.3263
	18	**128, 512**	**20**	**42**	**508.4966**	**17.6164**	**40.7820**	**22.5499**
	19	256, 128, 512	20	42	946.5297	23.2624	46.4645	30.7657
	20	512, 128	50	42	779.9055	21.9331	48.2355	27.9268
LSTM (pol. Data	1	32	20	42	530.0662	17.9308	40.2165	23.0232
	2	32, 32	20	42	595.5813	19.3811	40.8632	24.4045

and met. Data)	**11**	**128, 128**	**20**	**42**	**508.2342**	**17.8310**	**39.8435**	**22.5441**
	24	512, 256	20	42	532.7318	18.2944	39.8220	23.0810
	25	1024, 512	20	42	524.9836	18.3667	40.1733	22.9125
LSTM (pol. Data only)	1	32	20	42	517.5111	18.4543	45.0459	22.7489
	2	32, 32	20	42	494.2478	18.2122	42.8390	22.2317
	17	**512, 128**	**20**	**42**	**463.2636**	**17.1121**	**38.7530**	**21.5236**
	24	512, 256	20	42	464.5804	17.3516	40.3745	21.5541
	25	1024, 512	20	42	490.4218	18.5823	47.5220	22.1454

Table 3. CNN-LSTM with pollutant and meteorological data modelling result

No	CNN Filters	LSTM Units	Epoch	Batch Size	MSE	MAE	MAPE	RMSE
1	64	64	20	42	537.9931	18.1834	39.2839	23.1947
2	64	32	20	42	608.2287	19.5500	42.4178	24.6623
37	300	64, 256	20	42	508.1072	18.2648	42.0780	22.5412
38	32	64, 16	20	42	545.9707	18.2972	39.3336	23.3660
39	**32**	**64, 64**	**20**	**42**	**474.3771**	**17.5033**	**42.8834**	**21.7802**

Table 4. CNN-LSTM with pollutant data modelling result

No	CNN Filter s	LSTM Units	Epoch	Batch Size	MSE	MAE	MAPE	RMSE
1	64	64	20	42	487.1058	17.8497	41.8720	22.0705
2	64	32	20	42	508.2961	18.4842	43.6100	22.5454
27	**300**	**64, 32**	**40**	**42**	**466.5632**	**17.3758**	**41.4981**	**21.6000**
38	32	64, 16	20	42	463.5708	17.9327	44.6085	21.5307
39	32	64, 64	20	42	576.7006	19.3063	42.1496	24.0146

Table 2, Table 3, and Table 4 shows modelling results that done through iteration for the purpose of hyperparameter tuning process and two different data configurations. All the activation functions used in the modelling were ReLU. Two LSTM models trained using LeakyReLU activation function but didn't show any improvement, all CNN kernel size that use in the configuration is 1. All the bolded row in the Table 2, Table 3, and Table 4 are the chosen models that has best performance from all the evaluation metrics.

The overall comparison shows that CNN-LSTM achieved the best performance in terms of MSE and RMSE when using combined pollutant and meteorological data, indicating predictions closest to the actual values, although its MAPE remained relatively high. MAE values across all models were similar. When using only pollutant data, LSTM showed the most improvement, with significant reductions across all error metrics, suggesting it is more effective in capturing temporal patterns from pollutant data alone. Meanwhile, CNN and CNN-LSTM showed only slight performance changes, with CNN-LSTM still maintaining higher proportional errors (MAPE) despite close predictions to actual values. This suggests that, pollutant data alone contained sufficient information for $PM_{2.5}$ index prediction, while the addition of meteorological features did not always improve performance. Nonetheless, the proposed CNN-LSTM with pollutant and meteorological data still provided acceptable predictive performance, highlighting its potential for integrating multiple data sources in air quality prediction tasks.

5 Conclusion

This study shows that adding meteorological data does not always improve model performance in predicting $PM_{2.5}$. The LSTM model worked better when it used only pollutant data, suggesting that historical pollutant patterns alone are strong enough for short-term $PM_{2.5}$ prediction in Central Jakarta. The correlation and regression tests also supported this, showing that pollutants such as PM_{10}, NO_2, and O_3 have the biggest influence on $PM_{2.5}$ levels. Meteorological factors like humidity, temperature, sunshine, and wind direction still had some effect, but not as strong, while rainfall and wind speed were found to have little impact.

In terms of model performance, the CNN-LSTM gave the best results on MSE and RMSE when combining pollutant and meteorological data, meaning its predictions were closest to the actual values. However, its MAPE was still quite high, showing that errors were larger when $PM_{2.5}$ values were small. Overall, the LSTM model with pollutant-only input gave the most reliable and efficient results, with the lowest error values across all metrics. This makes it the best model and data combination for predicting $PM_{2.5}$ in Central Jakarta.

For future work, the study can be expanded to cover a wider area, such as the whole of Jakarta or even at the national level, by combining data from multiple air monitoring stations to improve spatial accuracy. The use of higher-resolution data, such as hourly measurements, is also recommended to better capture rapid fluctuations in air pollution, especially during peak hours. In addition, further exploration of the regressor layers in CNN-LSTM models, such as experimenting with the number of dense layers, units, and activation functions could help improve prediction accuracy.

Data Availability Statement

The pollutant dataset used in this study was obtained from Satu Data Jakarta. The dataset can be accessed for free on the following Satu Data Jakarta website: Satu Data Jakarta. And meteorological data in this study was obtained from Meteorology, Climatology, and Geophysics Agency (BMKG) Data Online. The dataset can be downloaded for free after signing up an account on the following BMKG Data Online website: BMKG Data Online.

Credit Authorship Contribution Statement

Stevanus Febrian: Data curation, Formal analysis, Methodology. **Syarifah Diana Permai**: Supervision, Writing - Review & Editing.

References

1. F.S. Pramudya, F.I. Kurniadi, A.N. Rakhiemah, *2023 7th International Conference on New Media Studies (CONMEDIA)*, Proc. Int. Conf. New Media Stud. (CONMEDIA), 162 (2023)
2. G. Syuhada, et al., Int. J. Environ. Res. Public Health **20**, 2916 (2023)
3. T. Istiana, et al., Aerosol Air Qual. Res. **23**, 230014 (2023)
4. S. Akmal, A. Carolina, Relawan J. Indones. **3**, 22 (2023)
5. D.D. Akbar, Interaction: J. Pendidik. Bahasa **10**, 764 (2023)
6. T. Handhayani, Sci. Rep. **13**, 5798 (2023)
7. N. Kadir, M. Faisal, F. Kurniawan, Appl. Inf. Syst. Manag. **7**, (2024)
8. X. Bai, N. Zhang, X. Cao, W. Chen, PeerJ **12**, e17811 (2024)
9. D. Li, J. Liu, Y. Zhao, Atmosphere (Basel) **13**, 1719 (2022)

10. Dinas Komunikasi Informatika Provinsi DKI Jakarta, Satu Data Jakarta (online resource)
11. Direktorat Data dan Komputasi BMKG, Data Online BMKG (online resource)
12. J. Peng, et al., *Proceedings of the 2021 International Conference on Management of Data*, Proc. Int. Conf. Manage. Data, 2271 (2021)
13. I. Pratama, A.E. Permanasari, I. Ardiyanto, R. Indrayani. *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, Proc. Int. Conf. Inf. Technol. Syst. Innov., 1 (2016)
14. R.O. Weber, J. Appl. Meteorol. **36**, 1403 (1997)
15. Z. Zeng, W.W. Hsieh, W.R. Burrows, A. Giles, A. Shabbar, Atmos.-Ocean **49**, 22 (2011)
16. A. Ambarwari, Q.J. Adrian, Y. Herdiyeni, J. RESTI **4**, 117 (2020)
17. Z. Zhang, S. Zhang, C. Chen, J. Yuan, Alex. Eng. J. **93**, 128 (2024)
18. I.H. Sarker, SN Comput. Sci. **2**, 420 (2021)