

COMPARATIVE PERFORMANCE OF MLP AND LSTM SOFT-SENSOR MODELS FOR ESTIMATING CLASS A PAN WATER LEVEL FROM HIGH-RESOLUTION AUTOMATIC WEATHER STATION DATA

Asiyah¹⁾, Sajarwo Anggai²⁾, Abu Khalid Rivai³⁾, Sugiarto⁴⁾

^{1,2,3}Graduate Program of Informatics Engineering, Universitas Pamulang, Indonesia

⁴Badan Meteorologi Klimatologi dan Geofisika, Jakarta, Indonesia

email: 1asiyah.pwr@gmail.com, 2sajarwo@gmail.com, 3abukhalidr@gmail.com, 4sugiarto@bmkg.go.id

Abstract

Evaporation is a critical meteorological parameter for hydrology, agriculture, and climatology. However, its direct observation at the Indonesian Bureau of Meteorology, Climatology, and Geophysics (BMKG) stations heavily depends on manual hook-gauge readings and automatic water-level sensors, which are highly susceptible to data gaps during equipment malfunctions. This study developed and compared two soft-sensor models, a Multi-Layer Perceptron (MLP) and a Long Short-Term Memory (LSTM) network, to estimate the 5-minute water level in a Class A evaporation pan using Automatic Weather Station (AWS) variables. One year of 5-minute-resolution data (2025) from the Fatmawati Bengkulu Meteorological Station was utilized. The model inputs included air temperature, relative humidity, wind speed, solar radiation, barometric pressure, pan water temperature, and rainfall. Based on an Autocorrelation Function analysis, the temporal lookback window was fixed at 12 lags (60 minutes). The datasets were partitioned chronologically using a 70/15/15 split for training, validation, and testing, respectively. Evaluation metrics included RMSE, MAE, and R^2 . The MLP model achieved a lower RMSE (2.9112 mm) and a higher R^2 (0.9742), which may suggest a greater robustness against large prediction errors and a stronger global fit. In contrast, the LSTM model yielded the lowest MAE (1.1048 mm), indicating a lower average absolute prediction error under typical conditions, though it exhibited higher sensitivity to abrupt environmental adjustments (RMSE: 3.6215 mm). These findings indicate that both architectures represent viable soft-sensor options. Ultimately, model selection should be determined by whether the operational system prioritizes overall error stability or lower average absolute error, thereby providing practical guidance for BMKG data-continuity protocols.

Keywords: Soft Sensor; Pan Evaporation; MLP; LSTM; High-Resolution Meteorological Data; Time-Series Regression

1. INTRODUCTION

Accurate information on meteorological parameters, particularly evaporation, plays an important role in agriculture, hydrology, and climatology. Reliable estimates of evaporation and evapotranspiration directly influence water resource management, agricultural productivity, and infrastructure planning. Evaporation is commonly observed using a Class A open pan, and continuous, accurate data are essential for water-balance analysis, irrigation estimation, hydrological modelling, and drought early-warning systems. However, the recorded pan water level is also influenced by rainfall, manual refilling, drainage, leakage, maintenance, and sensor disturbances; therefore, pan water level should not be treated as equivalent to evaporation.

Ideally, the pan water level is recorded automatically using a high-precision water-level sensor; however, such sensors are costly, and many stations of the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG) still rely on conventional observation using a stilling well and manual hook-gauge readings in accordance with WMO-8 (World Meteorological Organization, 2010). This manual approach depends heavily on the presence and accuracy of an observer, leaving two vulnerabilities: instrument malfunction and human error. Both can create gaps that degrade the quality of any analysis or prediction that relies on continuous data. The quality control of Automatic Weather Station (AWS) observations, therefore, remains a persistent challenge (Wang & Wang, 2017), and robust, affordable AWS deployment is still an active concern for developing regions

(Nsabagwa et al., 2019), including efforts to expand AWS networks for climate monitoring in Indonesia (Pusparini et al., 2019). Related environmental studies have also used machine-learning methods to reconstruct missing sub-hourly sensor observations (Chivers et al., 2020).

To address this problem, a soft sensor offers a relevant solution. A soft sensor is a model-based estimation system that uses readily available, inexpensive, and already automated parameters to predict a quantity that is difficult to measure directly. In this context, AWS variables such as air temperature, relative humidity, wind speed, solar radiation, pressure, pan water temperature, and rainfall provide physically relevant predictors of the recorded pan water level. They can therefore be used to develop a pan-water-level soft sensor (Ghorbani et al., 2018).

Advances in machine learning and deep learning have shown remarkable success in modelling the complex, non-linear relationships found in hydrological time-series data (LeCun et al., 2015; Xu & Liang, 2021). The Multi-Layer Perceptron (MLP) effectively captures non-linear patterns, whereas the Long Short-Term Memory (LSTM) network is specifically designed to handle long-term dependencies in sequential data and has proven reliable for daily meteorological prediction (Albeladi et al., 2023; Patro & Bartakke, 2024). Deep neural networks have likewise been applied successfully to the estimation of reference evapotranspiration (Saggi & Jain, 2019).

Several studies have applied these methods to pan evaporation. Samii et al. (2023) compared deep LSTM and MLP for arid regions and reported LSTM as consistently superior, while Malik et al. (2022) showed that feed-forward deep learning outperformed gradient boosting machines. Abed et al. (2022) found convolutional and deep neural networks effective for estimating monthly pan evaporation, and Al-Mukhtar (2021) evaluated several machine learning methods for the same task. Other studies have refined LSTM-based pan-evaporation models through meta-heuristic optimisation under limited climatic data (Fu & Li, 2023; Kisi et al., 2022) and through dynamic sliding-window strategies for forecasting (Bhattarai et al., 2023). Most previous studies used monthly or daily targets, non-tropical or arid datasets, and different training procedures. Direct evidence remains limited for a controlled MLP–LSTM comparison using five-minute tropical AWS data to estimate Class A pan water level (Wang et al., 2023).

Accordingly, this paper presents a comparative analysis of two deep learning architectures, MLP and LSTM, for building an AWS-based soft sensor to estimate the absolute water level in a Class A evaporation pan at five-minute intervals. The objectives are to build and apply both models, compare their performance using RMSE, MAE, and R^2 , and assess their preliminary suitability for maintaining pan-water-level data continuity when direct measurement fails. The novelty lies in comparing both architectures under the same high-resolution tropical dataset and temporal context.

2. RESEARCH METHODS

This study used a quantitative approach based on a time-series dataset with a five-minute observation resolution, obtained directly from the integrated sensors of a digitalised AWS at the Fatmawati Bengkulu Meteorological Station. The input variables were selected from the weather elements most influential to evaporation dynamics: air temperature ($^{\circ}\text{C}$), relative humidity (%), wind speed (m/s), solar radiation (W/m^2), barometric pressure (mb), pan water temperature ($^{\circ}\text{C}$), and rainfall recorded at five-minute intervals. The target variable was the absolute Class A open-pan water level (mm) at each timestamp, cross-validated using both manual observer readings and the automatic water-level sensor. Thus, this study estimated pan water level rather than daily accumulated evaporation or water-level change. The sample comprised the nominal complete time grid for 2025, yielding 105,120 five-minute timestamps before preprocessing. To avoid temporal leakage, the 70/15/15 partition was chronological and performed without shuffling, with the earliest observations assigned to training, followed by validation and testing observations. Located in Bengkulu on the south-western coast of Sumatra, the station experiences a humid tropical climate with distinct wet and dry seasons. The nominal partition corresponded to approximately 73,584 training timestamps, 15,768 validation timestamps, and 15,768 test timestamps. 15,768 test observations.

2.1 DATA PRE-PROCESSING AND FEATURE ENGINEERING

Raw sensor data were cleaned of missing values, invalid entries, duplicate timestamps, and inconsistent formats, and converted to a numerical format suitable for deep learning. Standardisation parameters were fitted using the training data only and subsequently applied to the validation and test data, preventing information from the later periods from influencing model training. Cyclical time encoding using sine and cosine of the hour was applied so the models could recognise periodic daily boundaries (for example, that 23:55 is physically adjacent to 00:00). Lag features and rolling statistics were also derived, providing historical context that benefits the non-sequential MLP in particular. The choice of input variables was guided by their theoretical physical relationship with the water level: thermal variables (air temperature and pan water temperature) and solar radiation are expected to be negatively related to water level because higher energy input accelerates evaporation, whereas relative humidity is expected to be positively related because a moister atmosphere suppresses evaporation.

2.2 LOOKBACK WINDOW SELECTION

The temporal lookback window for the sequential LSTM was determined using the Autocorrelation Function (ACF) of the target variable. As shown in Figure 1, autocorrelation is very high at early lags and decays gradually with increasing time distance. Applying a threshold of 0.5, the autocorrelation curve crosses the boundary at lag 12. With a five-minute resolution, this fixes the memory window at 12 lags, equivalent to 60 minutes. The threshold was treated as an initial heuristic and did not establish that 60 minutes was the optimal predictive window.

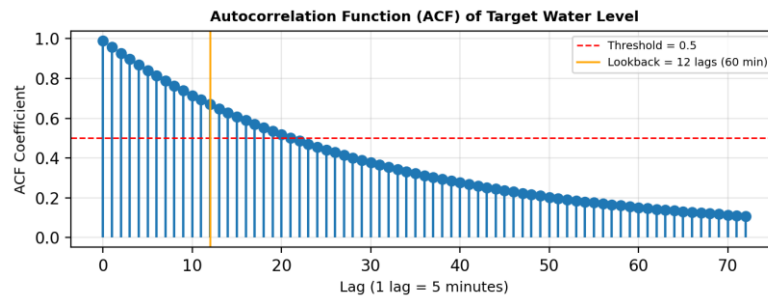


Figure 1. Autocorrelation Function (ACF) of the Class A pan water-level target

2.3 MODEL ARCHITECTURES AND TRAINING

Both models were trained under an identical comparative scheme, with the data divided chronologically into 70% for training, 15% for validation, and 15% for testing. To compare the architectures fairly, both models should receive equivalent historical information; the same 12-timestep predictor history was flattened for the MLP and retained as a sequence for the LSTM, without using previous pan water-level values as predictors. The MLP baseline is a fully connected network with a first dense layer of 256 neurons using ReLU activation, followed by 128- and 64-neuron layers, each combined with Batch Normalisation and graded dropout (0.3, 0.2, 0.1), and a single linear output neuron. The LSTM is a two-layer stacked recurrent network that receives a three-dimensional tensor with a 12-lag timestep dimension; the first LSTM layer has 128 memory units with the `return_sequences` parameter enabled, feeding a second 64-unit LSTM layer, with a 0.2 dropout between them, a 32-neuron ReLU dense layer, and a single linear output neuron.

To ensure a fair comparison, both models were compiled with Mean Squared Error as the loss function and optimised using the Adam optimiser with an initial learning rate of 0.001. Training was controlled by early stopping (patience of 10 epochs) and ReduceLROnPlateau (patience of 5 epochs, factor 0.5). Implementation used Python 3.10 with TensorFlow 2.14 and Keras; pre-processing used Pandas and NumPy; evaluation used Scikit-learn; and visualisation used Matplotlib and Seaborn.

Model performance was evaluated on the test set using three statistical metrics. The Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values, treating all errors linearly (Equation 1). The Root Mean Square Error (RMSE) is the square root of

the mean squared error and is highly sensitive to large errors because of the squaring operation (Equation 2). The coefficient of determination (R^2) measures the proportion of proportion of observed variance accounted for by the fitted model within the evaluated dataset (Equation 3), with values closer to 1 indicating a better fit. In these equations, n is the number of observations, y_i is the actual pan water-level value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values.

$$MAE = (1/n) \sum |y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{[(1/n) \sum (y_i - \hat{y}_i)^2]} \quad (2)$$

$$R^2 = 1 - [\sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2] \quad (3)$$

3. RESULTS AND DISCUSSION

3.1 QUANTITATIVE PERFORMANCE COMPARISON

Both architectures achieved strong predictive performance on the test set, with R^2 values above 0.96, indicating that the fitted models accounted for more than 96% of the observed variability in absolute Class A pan water level within this dataset. Table 1 summarises the comparative quantitative performance, and Figure 2 visualises the three metrics side by side. RMSE and MAE are expressed in millimetres.

Table 1. Comparative performance of MLP and LSTM on the test set

Model	RMSE(mm)	MAE (mm)	R^2
MLP	2.9112	1.5563	0.9742
LSTM	3.6215	1.1048	0.9600

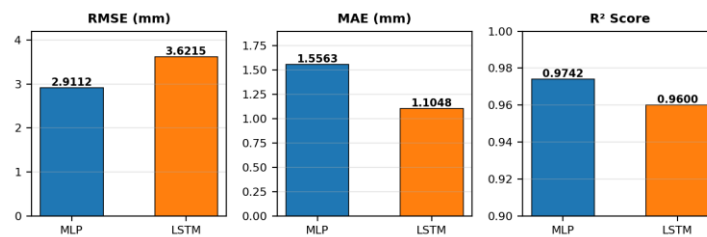


Figure 2. Comparison of RMSE, MAE, and R^2 for MLP and LSTM

The two models display distinct error characteristics. The lower RMSE suggests that the MLP produced fewer or smaller large-magnitude errors than the LSTM on the test set. Although the MLP has no temporal memory, the engineered features provided indirect temporal context, enabling it to map the non-linear relationship between inputs and the target effectively. The LSTM achieved the lowest MAE (1.1048 mm), indicating smallest average absolute prediction error across the test observations, consistent with its design as a sequential model that reads historical patterns through its memory mechanism.

However, the LSTM also produced the highest RMSE (3.6215 mm). Because RMSE penalises extreme errors heavily, this may indicate that the LSTM occasionally generated large errors. Such deviations may be associated with abrupt changes such as heavy rainfall, refilling, sensor spikes, or anomalous water-level shifts. Because these conditions were not separately tested, the causes and any time-lag effect cannot be confirmed from the current analysis. The MLP had a higher MAE (1.5563 mm), indicating a larger average absolute error across the test observations.

3.2 TIME-SERIES TRACKING

Figure 3 compares the actual and predicted water level over time. Both models follow the main pattern of decline during evaporation and the sharp rise in water level. The MLP predictions are relatively stable but less responsive to short-term fluctuations and may smooth some peaks and troughs. The LSTM follows the gradual pattern more closely, consistent with its lower MAE, but appears to respond later at the abrupt rise. This visual pattern may indicate a delayed response, although event-specific analysis is required before such an effect can be established.

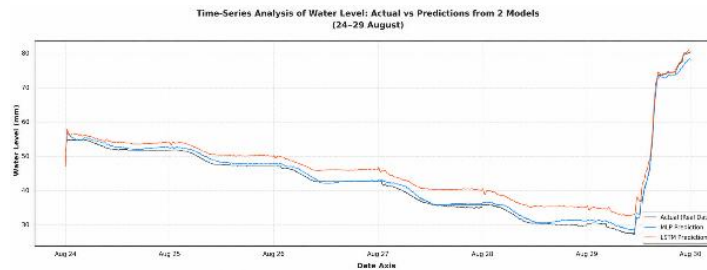


Figure 3. Time-series tracking of actual versus predicted Class A Pan water level

3.3 PREDICTION AGREEMENT AND RESIDUALS

Figure 4 presents scatter plots of predicted versus actual values, with the ideal fit line ($y = x$) shown. For the MLP, the points lie close to the diagonal, consistent with its highest R^2 . For the LSTM, most points also closely follow the diagonal. However, a limited number deviate further, which is consistent with its higher RMSE. The LSTM therefore produced small errors for many observations but larger deviations for some observations; condition-specific analysis is required to determine whether these occurred during extreme events.

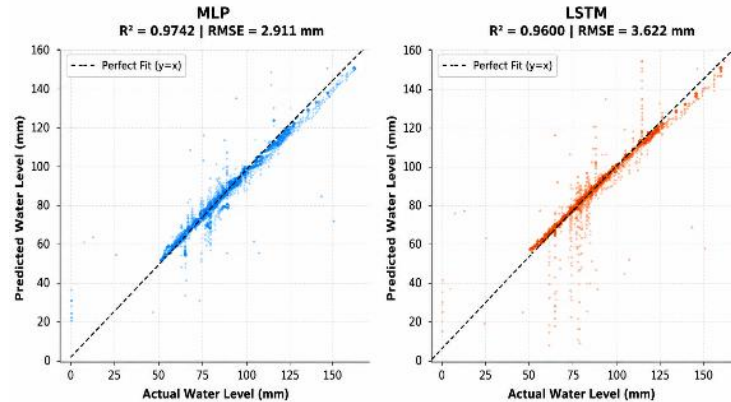


Figure 4. Scatter plots of prediction agreement for MLP and LSTM

Figure 5 shows the residual distributions. The MLP residuals concentrate near zero but show a slightly wider spread across many observations, consistent with its low RMSE yet higher MAE. The LSTM residual distribution is sharper around zero, which is consistent with its lowest MAE, but it exhibits a longer tail that may raise its RMSE. Additional residual statistics are required to support this interpretation more completely.

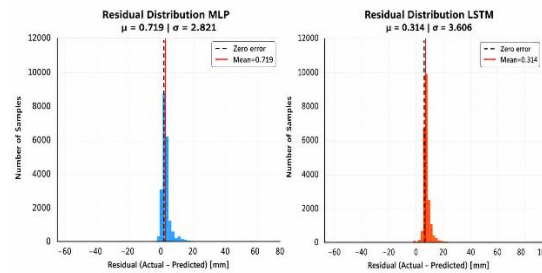


Figure 5. Residual distribution histograms of MLP and LSTM predictions

3.4 COMPARISON WITH PREVIOUS STUDIES

The differences in rankings across metrics show that selecting the best model cannot rely on a single metric. The MLP leads on RMSE and R^2 , while the LSTM leads on MAE. This finding partly diverges from studies such as Samii et al. (2023), which found LSTM to be consistently dominant for pan evaporation. Here, the LSTM is the most accurate on average but less stable against high-magnitude errors, indicating that its advantage depends on data characteristics and the selected evaluation metric. The competitiveness of the MLP also reinforces Malik et al. (2022) findings that feed-forward deep learning remains a strong baseline, especially when supported by feature engineering that provides historical context. Nevertheless, direct numerical comparison across studies should be interpreted cautiously because their target variables, temporal resolutions, climatic conditions, input features, training procedures, and test designs differ.

From an operational standpoint, both models show preliminary potential as soft sensors for maintaining continuity of pan water-level data when the direct sensor fails. The MLP may be preferred when lower RMSE and higher R^2 are prioritised, whereas the LSTM may be preferred when lower MAE is more important. However, the current results do not establish operational readiness or suitability for resource-constrained deployment. Such claims require computational benchmarking, conventional baselines, repeated training, event-specific analysis, and simulated continuous sensor gaps of different durations.

4. CONCLUSION

This study built and compared two deep-learning soft-sensor models, MLP and LSTM, to estimate the absolute Class A pan water level from high-resolution AWS data. Both architectures performed strongly, each accounting for more than 96% of the observed variance in the evaluated dataset. The MLP achieved the lowest RMSE (2.9112 mm) and the highest R^2 (0.9742), indicating fewer or smaller high-magnitude errors. In comparison, the LSTM achieved the lowest MAE (1.1048 mm), indicating the smallest average absolute prediction error, but it also had the highest RMSE (3.6215 mm). The latter result may reflect occasional large deviations; however, a time-lag effect was not established by the present analysis. The principal contribution of this work is a direct comparison of the two architectures within a single experimental framework using tropical five-minute AWS data. However, this study was limited to a single station, a single year, and the reported model runs, without multi-station validation, persistence and conventional baselines, repeated-run uncertainty, event-specific testing, computational benchmarking, or simulated continuous sensor gaps. Therefore, the results indicate preliminary feasibility rather than operational readiness. Future work should extend the historical and spatial coverage, compare alternative lookback windows and baseline models, repeat model training, and evaluate simulated sensor outages before real-time deployment.

5. ACKNOWLEDGEMENT

The authors thank the Graduate Program of Informatics Engineering, Universitas Pamulang, and the staff of the Fatmawati Bengkulu Meteorological Station for providing the data and support that made this research possible.

6. REFERENCES

- Abed, M., Imteaz, M. A., Ahmed, A. N., & Huang, Y. F. (2021). Application of long short-term memory neural network technique for predicting monthly pan evaporation. *Scientific Reports*, 11, 20742. <https://doi.org/10.1038/s41598-021-99999-y>
- Abed, M., Imteaz, M. A., Ahmed, A. N., & Huang, Y. F. (2022). Modelling monthly pan evaporation utilising random forest and deep learning algorithms. *Scientific Reports*, 12, 13132. <https://doi.org/10.1038/s41598-022-17263-3>
- Albeladi, K., Zafar, B., & Mueen, A. (2023). Time series forecasting using LSTM and ARIMA. *International Journal of Advanced Computer Science and Applications*, 14(1), 313–320.
- Al-Mukhtar, M. (2021). Modeling of pan evaporation based on the development of machine learning methods. *Theoretical and Applied Climatology*, 146(3–4), 961–979. <https://doi.org/10.1007/s00704-021-03760-4>
- Bhattarai, A., Qadir, D., Sunusi, A. M., Getachew, B., & Mallah, A. R. (2023). Dynamic sliding window-based long short-term memory model development for pan evaporation forecasting. *Knowledge-Based Engineering and Sciences*, 4(1), 37–54.
- Chivers, B. D., Wallbank, J., Cole, S. J., Sebek, O., Stanley, S., Fry, M., & Leontidis, G. (2020). Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. *Journal of Hydrology*, 588, 125126. <https://doi.org/10.1016/j.jhydrol.2020.125126>
- Fu, T., & Li, X. (2023). Estimating the monthly pan evaporation with limited climatic data in dryland based on the extended long short-term memory model enhanced with meta-heuristic algorithms. *Scientific Reports*, 13, 5960. <https://doi.org/10.1038/s41598-023-32838-4>
- Ghorbani, M. A., Kazempour, R., Chau, K. W., Shamshirband, S., & Ghazvinei, P. T. (2018). Forecasting pan evaporation with an integrated artificial neural network quantum-behaved particle swarm optimization model: A case study in Talesh, Northern Iran. *Engineering Applications of Computational Fluid Mechanics*, 12(1), 724–737.
- Kisi, O., Mirboluki, A., Naganna, S. R., Malik, A., Kuriqi, A., & Mehraein, M. (2022). Comparative evaluation of deep learning and machine learning in modelling pan evaporation using limited inputs. *Hydrological Sciences Journal*, 67(9), 1309–1327. <https://doi.org/10.1080/02626667.2022.2063724>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Malik, A., Saggi, M. K., Rehman, S., Sajjad, H., Inyurt, S., Bhatia, A. S., Farooque, A. A., Oudah, A. Y., & Yaseen, Z. M. (2022). Deep learning versus gradient boosting machine for pan evaporation prediction. *Engineering Applications of Computational Fluid Mechanics*, 16(1), 570–587. <https://doi.org/10.1080/19942060.2022.2027273>
- Nsabagwa, M., Byamukama, M., Kondela, E., & Otim, J. S. (2019). Towards a robust and affordable automatic weather station. *Development Engineering*, 4, 100040.
- Pusparini, N., Sopaheluwakan, A., & Hidayat, R. (2019). Development of an automatic weather station (AWS) network for climate monitoring in Indonesia. *Journal of Physics: Conference Series*, 1351, 012073. <https://doi.org/10.1088/1742-6596/1351/1/012073>
- Saggi, M. K., & Jain, S. (2019). Reference evapotranspiration estimation and modeling of the Punjab Northern India using deep learning. *Computers and Electronics in Agriculture*, 156, 387–398.
- Samii, A., Karami, H., Ghazvinian, H., Safari, A., & Dadrasajirlou, Y. (2023). Comparison of deep-LSTM and MLP models in estimation of evaporation pan for arid regions. *Journal of Soft Computing in Civil Engineering*, 7(2), 155–175. <https://doi.org/10.22115/SCCE.2023.367948.1550>
- Wang, H., Sun, F., Liu, F., Wang, T., Liu, W., & Feng, Y. (2023). Reconstruction of the pan evaporation based on meteorological factors with machine learning method over China. *Agricultural Water Management*, 287, 108416. <https://doi.org/10.1016/j.agwat.2023.108416>
- Wang, J., & Wang, J. (2017). A review of data quality control for automatic weather station observations. *Atmospheric and Oceanic Science Letters*, 10(6), 481–487. <https://doi.org/10.1080/16742834.2017.1369325>
- World Meteorological Organization. (2010). Guide to meteorological instruments and methods of observation (WMO-No. 8).
- Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. *Wiley Interdisciplinary Reviews: Water*, 8(5), e1533. <https://doi.org/10.1002/wat2.1533>