International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Predicting Evapotranspiration in the Semi-Arid Region of Indore Using AI models

Adnan Barwaniwala

Student, Daly College

Abstract

Indore is facing increasing water scarcity and inefficient irrigation is a major contributor to it. Hence, this research explores the use of Artificial Intelligence models like Artificial Neural Networks (ANNs) and Light Gradient Boosting Machine (LGBM) in predicting reference evapotranspiration (ET_0) using limited and sufficient data for Indore's semi-arid climate. In places where water is scarce, accurate prediction of ET_0 plays a vital role in efficient irrigation planning. Based on historical meteorological data from 1985 to 2022, the models were trained and tested, with ANN generally outperforming LGBM especially when an extensive set of input variables was used. Furthermore, wind speed and net radiation were found to be crucial factors for ET_0 estimation. Nonetheless, even though the accuracy of ANN was higher than that of LGBM, its computational efficiency was higher and it proved to be more useful in certain scenarios where data is limited.

1. Introduction

Indore, a city in Central India, is facing severe water scarcity due to the rapid depletion of groundwater resources. The groundwater level in Indore has dropped from 150 metres in 2012 to a staggering 560 feet (about 170 metres) in 2023 [1]. This alarming decline is a result of overexploitation and unsustainable water use practices. Agriculture plays a significant role in the water crisis, with around 48% [2] of the total cultivable area in Indore under irrigation. Unfortunately, inefficient irrigation practices contribute to substantial water wastage with a significant portion of water being lost through evaporation, runoff, and inefficient application methods. Hence, there is a very urgent need to implement efficient irrigation practices in Indore to tackle the increasing water scarcity, decreasing crop yields and higher water costs farmers are facing.

To resolve this issue, accurate prediction of reference evapotranspiration (ET_0) is essential as it is used very commonly for efficient irrigation planning and sustainable agriculture. Evapotranspiration is the combined process of water evaporation from soil and plant surfaces and transpiration from plants (refer to Section 2.2). ET_0 measures the potential evapotranspiration from a well-watered reference surface and is a key indicator for determining crop water requirements. Traditionally, to accurately calculate ET_0, the FAO Penman-Monteith equation is widely used, but it relies on the availability of various meteorological data, which may not always be accurate or readily available in data-scarce regions like Indore [3]. Existing research has explored the use of empirical and other models for predicting ET_0 with only some studies addressing the challenge of limited meteorological data in Indore [4]. However, this study is the first to leverage artificial intelligence (AI) models, which have demonstrated superior performance in ET_0 prediction compared to traditional models, to enhance prediction accuracy in this context.

The primary objective of this study is to evaluate the performance of two advanced AI models—Artificial Neural Networks (ANN) and Light Gradient Boosting Machine (LGBM)—in predicting ET_0 for Indore's semi-arid climate. This research investigates how well these models perform across 6 different combinations of meteorological inputs. By testing various input combinations, the study aims to determine the most effective model and input variables for accurate ET_0 prediction in this specific climate and the performance of the models with limited and sufficient data.

To achieve this, historical meteorological data from 1985 to 2022 were collected and used to calculate ET_0 using the FAO Penman-Monteith equation. These calculated values were then used as the ground truth for training and testing the ANN and LGBM model. The models were evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), which help to quantify the accuracy of the predictions.

This study not only demonstrates the effectiveness of AI models in predicting ET_0 but also underscores the importance of specific weather factors like wind speed and net radiation in semi-arid climates like Indore. These insights can be crucial for improving agricultural water management in regions with climates similar to Indore. Additionally, it advances the field of agroclimatology by showing how AI can enhance the accuracy and reliability of ET_0 predictions in various types of climates, thereby promoting sustainable agricultural practices in water-limited environments.

2. Background

2.1 Area of Study

Indore, the largest city in the central Indian state of Madhya Pradesh, is situated on the Malwa Plateau at an average elevation of 553 metres above sea level. The city lies between 22.2° and 23.05° North latitude and 75.25° and 76.16° East longitude [5]. It has a semi-arid climate, characterised by three seasons: summer, monsoon, and winter. Summers are very hot, with average daily maximum temperatures ranging from 34.5°C in March to 40.4°C in May. The monsoon season lasts from June to September, bringing moderate rainfall ranging from 700 to 800 mm. During the rainy days, maximum temperatures can remain around 24-25°C due to cloud cover. Winters are cool, with average daily minimum temperatures between 10.3°C in January and 16.9°C in March [6].

2.2 What is Evapotranspiration (ET) and Reference Evapotranspiration (ET_0)?

Evapotranspiration (ET) is the combined process of water evaporation from soil and plant surfaces and transpiration through plant leaves. ET_0 represents the rate at which water is lost from a hypothetical reference crop—typically assumed to be a well-watered, actively growing grass with a height of 0.12 metres, a surface resistance of 70 s m^{-1} (relates to how easily water vapour moves from the plant surface to the air) and an albedo of 0.23 (indicates that 23% of incoming solar radiation is reflected). It plays a crucial role in agriculture by determining the amount of water required for crops to grow optimally. This ensures that crops receive the precise amount of water needed, reducing waste and optimising yields [7].

2.3 The FAO Penman-Monteith Equation

The FAO Penman-Monteith equation is a globally recognized method for estimating reference evapotranspiration (ET), essential for determining crop water requirements. The equation originated from the work of Howard Penman, who first introduced a method to estimate evaporation in 1948. It was later refined by John Monteith in 1965 to incorporate aerodynamic and surface resistance factors, leading to the Penman-Monteith equation. The FAO version, standardised in 1998, further simplifies the calculation by assuming a well-watered grass crop as a reference. The equation is [8]:

$$ET_0 = \frac{0.408\times\Delta\times(R_n-G)+\gamma\times\left(\frac{900}{T_{mean}+273}\right)\times u_2\times(e_s-e_a)}{\Delta+\gamma\times(1+0.34\times u_2)} \quad (1)$$

ET_0: Reference Evapotranspiration (mm/day)
Δ: Slope of the saturation vapour pressure curve (kPa/°C)
R_n: Net radiation at the crop surface (MJ/m²/day)
G: Soil heat flux density (MJ/m²/day)
γ: Psychrometric constant (kPa/°C)
T_{mean}: Mean daily air temperature at 2 m height (°C)
u_2: Wind speed at 2 m height (m/s)
e_s: Saturation vapour pressure (kPa)
e_a: Actual vapour pressure (kPa)
(e_s - e_a): Saturation vapour pressure deficit (kPa)

2.4 The Need for Estimating ET_0 through AI and ML

The FAO-56 Penman-Monteith (FAO-56 PM) model requires extensive and diverse meteorological data, which is often unavailable or unreliable, especially in developing regions such as Indore. Empirical models, which use fewer variables, have been proposed as alternatives but suffer from inconsistencies across different climatic conditions, leading to uncertainties in their predictions [3]. To address these challenges, machine learning (ML) models have emerged as a more practical and reliable approach. ML models can estimate ET_0 with greater accuracy and consistency, even in data-scarce regions like Indore, making them a favourable alternative to both the FAO-56 PM and empirical models.

2.5 Previous Work Done

In recent years, significant advancements have been made in estimating ET_0 using AI and ML models. In 2019, Yamfaç [9] performed a study where the kNN model outperformed the ANN in estimating ET_0 using various combinations of climatic data in the semi-arid environment of Turkey. Subsequently, in 2023, Babazadeh et al. [10] demonstrated the effectiveness of AI models like ANN, ANFIS, and hybrid approaches like ANN-GWO[1] in arid regions of Iran. These models were particularly successful in handling incomplete meteorological data, with ANFIS showing superior accuracy when external data from nearby weather stations were integrated. Similarly, in 2023, Yong et al. [3] explored the use of ML models, including ANN, LGBMs, and DFRs[2] in the tropical climate of Malaysia, a region with limited weather data. It found that ANNs performed best when all meteorological variables were included, while LGBM was most effective when fewer variables were available. These studies underscore the growing importance of AI and ML in enhancing the accuracy and reliability of ET_0 estimation across diverse climatic conditions.

2.6 Gaps in Existing Research and Objectives of the Current Study

Existing research in ET_0 estimation using AI has predominantly focused on various geographical regions, including regions in China [11], India [12], Iran [10], and Malaysia [3]. However, the semi-arid climate of Indore, India, has remained largely unexplored. Prior studies have identified ANNs and LGBMs as highly effective models for predicting ET_0. This study aims to evaluate the performance of ANNs and LGBMs specifically in Indore's unique climate with limited and sufficient data and hence, to also assess their ability to generalise across different geographical locations. By addressing this gap, the research aims

[1] Artificial Neural Network (ANN), Fuzzy Neural Adaptive Inference System (ANFIS), and ANN-Gray Wolf Optimization (ANN-GWO).
[2] Artificial Neural Network (ANN), Light Gradient Boosting Machines (LGBMs), Decision Forest Regression (DFRs).

to enhance the applicability of AI models in diverse climatic contexts.

3. Dataset

The dataset was derived from weather data available on this NASA website, covering Indore from January 1, 1985, to December 31, 2022. The weather factors included:

- Surface Pressure (kPa)
- Temperature at 2 Metres (°C)
- Relative Humidity at 2 Metres (%)
- Wind Speed at 2 Metres (m/s)
- Maximum Temperature at 2 Metres (°C)
- Minimum Temperature at 2 Metres (°C)
- All Sky Surface Albedo (dimensionless)
- All Sky Surface Shortwave Downward Irradiance (MJ/m²/day)
- Clear Sky Surface Shortwave Downward Irradiance (MJ/m²/day)

Net Radiation (MJ/m²/day)[3] was calculated using the last three weather factors. The daily ET_0 (mm/day) was then computed using the FAO Penman-Monteith Equation (1), utilising all variables except the last three and net radiation. Additionally, The accuracy of the calculated ET0 values was verified by comparing them with the actual ET0 values for 2016[4].

Date	Mean Temp (C)	Max Temp (C)	Min Temp (C)	R. Humidity (%)	Wind Speed (m/s)	S. Pressure (kPa)	Net Radiation (MJ/m^2/day)	ET0 (mm/day)
01-01-1985	19.35	28.92	11.85	25.38	2.39	96.66	3.06	4.38
02-01-1985	18.20	27.43	10.20	25.31	2.06	96.92	3.36	3.87
03-01-1985	18.77	28.13	10.14	28.25	2.17	96.98	5.06	4.29
04-01-1985	21.37	31.23	12.73	32.94	2.94	96.79	6.63	5.59
05-01-1985	20.34	29.41	12.38	21.38	2.23	96.68	5.21	4.88

Figure 1. First 5 Rows of the Dataset Used. This is the head of the dataset, displaying all the different columns and the types of values they contain.

The dataset comprises 13,879 data points and was split into three parts for model development:

- **Training data (1985–2011): 71.05%**

Data used to train the model

- **Validation data (2012–2019): 21.05%**

Data used to evaluate model performance to optimise training

- **Testing data (2020–2022): 7.9%**

Data used to evaluate model performance after training

The data was split sequentially, rather than randomly, so the model captured any trends in the data that were season- or year-dependent. Below are some figures to better understand the dataset. Note in Fig. 3

[3] Refer to Appendix A to see how the Net Radiation was calculated.
[4] Refer to Appendix B for more information.

that Max Temp has the highest correlation coefficient with ET_0, magnitude-wise while R. Humidity and S. Pressure have negative correlations with ET_0.

	Mean Temp	Max Temp	Min Temp	Humidity	Wind Speed	Surface Pressure	Net Radiation	ETo
mean	26.00	33.32	19.51	50.07	2.65	96.29	9.73	5.48
std	4.94	5.15	5.40	26.26	1.25	0.50	3.44	2.73
min	11.44	17.23	3.73	4.12	0.32	94.38	-0.20	0.90
50%	25.71	31.72	21.04	47.00	2.32	96.32	9.73	4.47
max	38.93	47.76	31.56	95.38	8.08	97.49	18.80	15.27

Figure 2. Summary Statistics of Input and Output Variables. The table presents the statistics for the weather variables and ET0 across the dataset. They provide an overview of the data distribution, highlighting the range and variability of each variable.

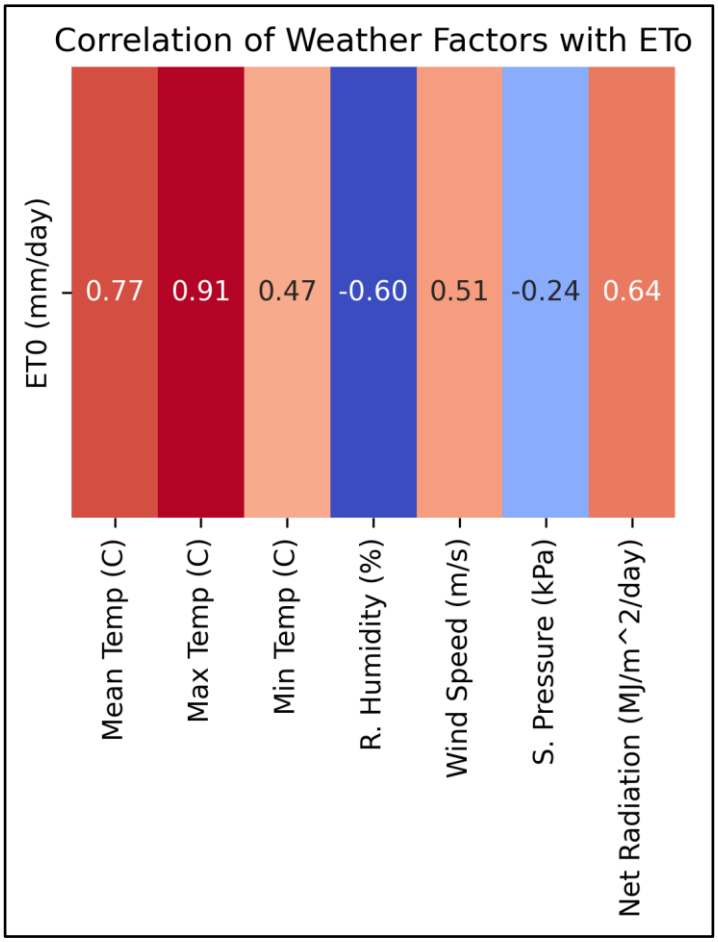


Figure 3. Correlation of Weather Factors with ET_0. It illustrates the correlation coefficients between key weather variables and ET0 for Indore. Higher values indicate stronger relationships. Negative correlations indicate an inverse relationship with ET0.

4. Models

Two models, ANNs and LGBMs, were chosen to predict ET_0 due to their proven superior performance across various climatic conditions, according to the findings of existing research.

4.1 Artificial Neural Networks (ANNs)

ANNs are inspired by the biological neural networks in the human brain, consisting of interconnected neurons (nodes) organised into layers. The ANN model used in this study had three layers: an input layer, 4 hidden layers, and an output layer. The input layer received the weather factors, including temperature, humidity, wind speed, surface pressure, and net radiation. The hidden layers employed non-linear activation functions (such as ReLU) to capture complex relationships in the data. The output layer produced the ET_0 prediction.

The model was trained using backpropagation, a method that adjusts the weights of the connections to minimise the error between the predicted and actual ET_0 values. It had the following architecture:

- Input Layer had 3,4 or 6 neurons (depends on the weather factor combination for the input variables. Refer to section 5)
- Hidden Layers: 256, 128, 64 and 32 neurons (4 layers)
- Output Layer had 1 neuron

4.2 Light Gradient Boosting Machine (LGBM)

Light Gradient Boosting Machine (LGBM) is a model that predicts outcomes using decision trees. In simple terms, decision trees make decisions by splitting data into branches based on specific conditions. LGBM builds multiple small decision trees sequentially, with each tree improving on the errors made by the previous one. Unlike other tree-based methods, LGBM grows these trees by focusing on the branches (leaves) that improve predictions the most, which makes it faster and more accurate, especially with large datasets [13].

It was trained on the same (input) weather variables as the ANN model to predict a single continuous output, ET_0. Instead of optimising architecture, the model used default settings, which often performed similarly or even better than tuned parameters. The key default values were a learning rate of 0.01, 100 estimators (trees), and 3 leaves per tree [13]. This approach balanced simplicity and performance, allowing the model to efficiently capture the complex relationships in the data while maintaining computational efficiency.

5. Methodology

5.1 Input Data Combinations

The LGBM and ANN models were trained and tested on 6 different combinations of weather factors:

1. Max, Min and Mean Temperature
2. Max, Min and Mean Temperature and Net Radiation
3. Max, Min and Mean Temperature and Relative Humidity
4. Max, Min and Mean Temperature and Surface Pressure
5. Max, Min and Mean Temperature and Wind Speed
6. Max, Min and Mean Temperature, Wind Speed, Relative Humidity and Surface Pressure

Most papers [3, 9, 10] evaluate the models on 3-4 combinations but this limits the scope of their results. Hence, this study uses 6 combinations to assess the model performance and trends in performance more comprehensively.

Combination 1, consisting of Maximum, Minimum and Mean Temperature, was selected based on the

strong correlation these variables displayed with ET_0 (as shown in Fig. 2). These temperature variables are also easily accessible in data-scarce regions like Indore, making them a practical choice. Combinations 2 through 5 explore the impact of adding individual weather factors (Net Radiation, Relative Humidity, Surface Pressure, and Wind Speed) to the core temperature variables. This approach helps assess how each factor enhances the prediction accuracy and their specific correlation with ET_0 in Indore's semi-arid climate. Additionally, combinations 1 to 5 are used to test the models' performance on limited input data. Combination 6 integrates most of the key weather variables (excluding Net Radiation, which is less accessible) to evaluate the model's performance when provided with nearly complete information. This comprehensive approach ensures a thorough examination of the influence of different weather factors on ET_0 predictions, specifically for the climate of Indore.

5.2 Performance Metrics

5.2.1 Mean Absolute Error (MAE)

MAE measures the average magnitude of errors between the predicted and observed values, without considering their direction. It is calculated using the formula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and n is the number of observations. A lower MAE indicates better model performance, with a value of 0 representing a perfect prediction. MAE is useful for understanding the average error magnitude, but it doesn't account for the variance in error magnitude.

5.2.2 Root Mean Squared Error (RMSE)

RMSE is similar to MAE but gives more weight to larger errors by squaring the differences between predicted and observed values before averaging them. The formula is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

Because RMSE squares the errors, it amplifies the effect of larger discrepancies, making it particularly useful when large errors are especially undesirable. A lower RMSE signifies better performance, with a value of 0 indicating a perfect fit.

5.2.3 R-squared (R^2)

R^2, or the coefficient of determination, represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

where $\bar{y}$ is the mean of the observed values. R^2 values range from 0 to 1, with values closer to 1 indicating that the model explains a large portion of the variance in the data, which means better model performance. An R^2 of 0 indicates that the model does not explain any of the variance.

These metrics together provide a comprehensive understanding of model performance, with lower MAE and RMSE values indicating less error and higher R^2 values indicating better explanatory power [14].

6. Results and Discussions

The results obtained from the ANN and LGBM models across six different combinations of weather variables are summarised in Table 1. The primary objective of the study was to evaluate the performance of these AI models in predicting ET0 in Indore's semi-arid climate, and the results provide significant insights into their effectiveness.

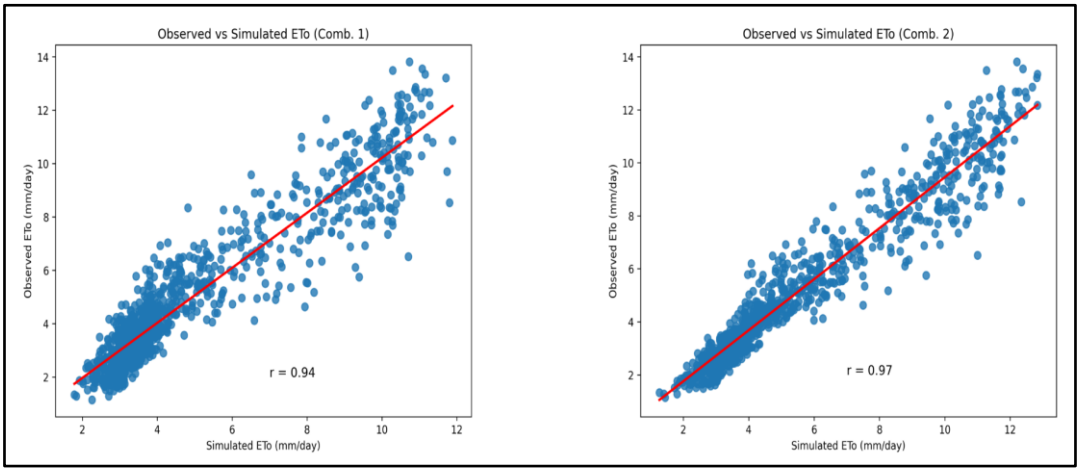
Figure 4. Results of the ANN and LGBM Model for the 6 Input Combinations on the Testing Subset.

	ANN			LGBM		
	MAE	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
Comb. 1	0.728	0.945	0.883	0.728	0.954	0.882
Comb. 2	**0.572**	**0.800**	**0.917**	**0.576**	**0.804**	**0.916**
Comb. 3	0.696	0.975	0.877	0.633	0.884	0.899
Comb. 4	0.747	0.944	0.885	0.695	0.899	0.895
Comb. 5	**0.477**	**0.602**	**0.953**	**0.482**	**0.605**	**0.953**
Comb. 6	**0.373**	**0.469**	**0.972**	**0.402**	**0.499**	**0.968**

6.1 Model Performance Overview

Both models generalise well to Indore's semi-arid climate, producing acceptable results across all input combinations, indicating their potential to predict ET_0 accurately in various climates. Although ANN generally performed better in terms of prediction accuracy, the LGBM model was significantly more efficient in terms of computational cost, processing time, and overall efficiency, highlighting LGBM's advantage in practical applications where resources and time are constrained, despite ANN's slightly superior predictive power.

Refer to Figures 5 and 6 to better understand the models' performance. They provide scatter plots for Observed vs Simulated ET_0 for the models across all the different input combinations.



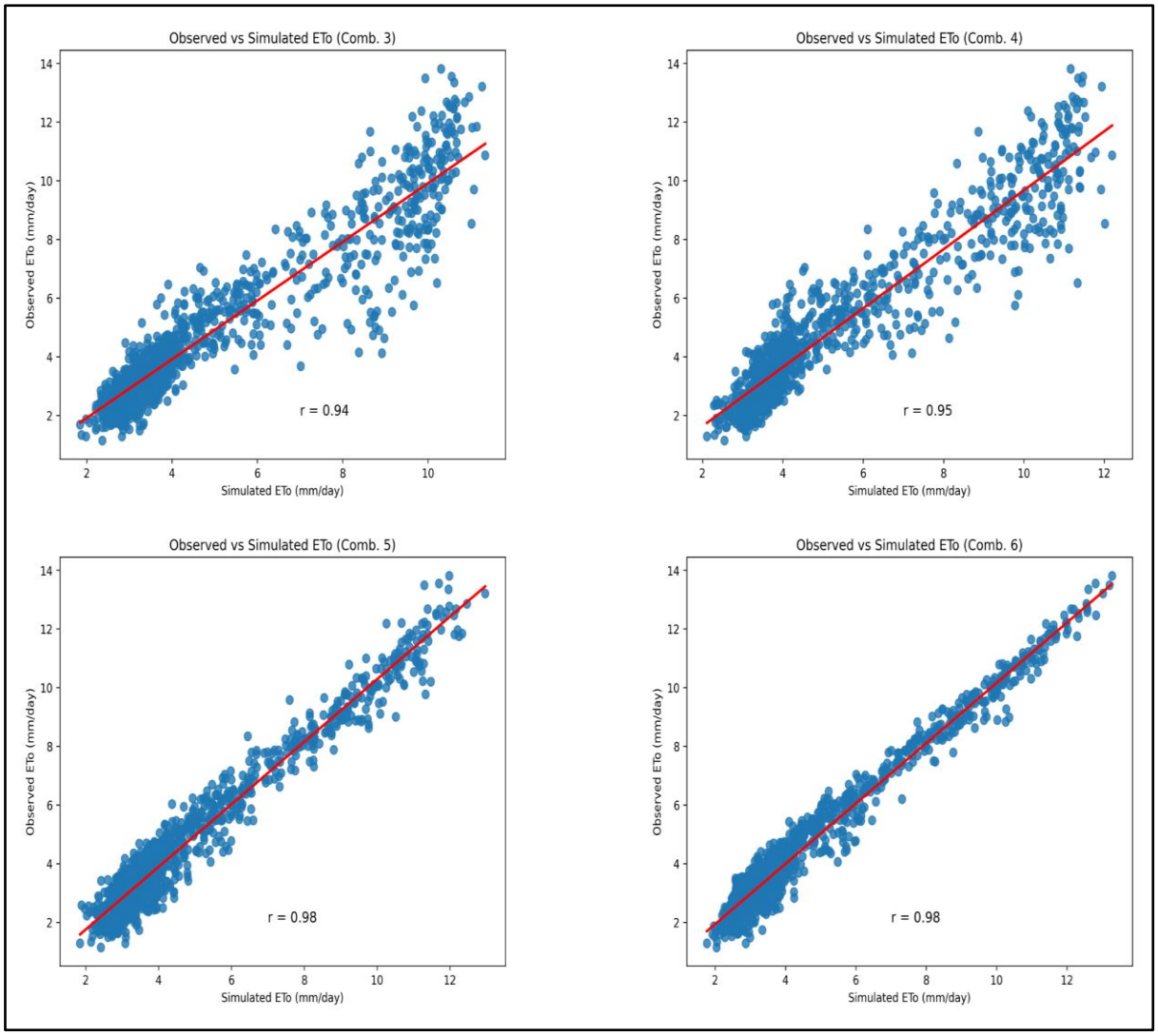
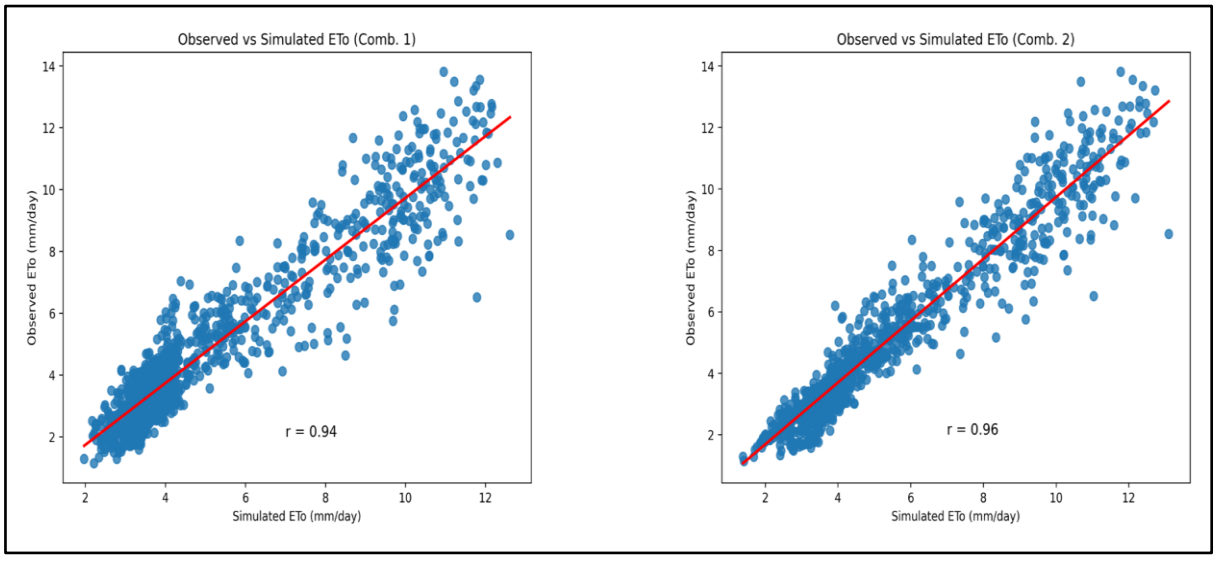


Figure 5. Observed vs Simulated ET_0 graphs for the ANN Model. The scatterplots compare the simulated and observed ET_0 for all the input combinations to the ANN model on the test data. The closer the points are to the red line and 'r' is to 1, the better the ANN is fit for the task.



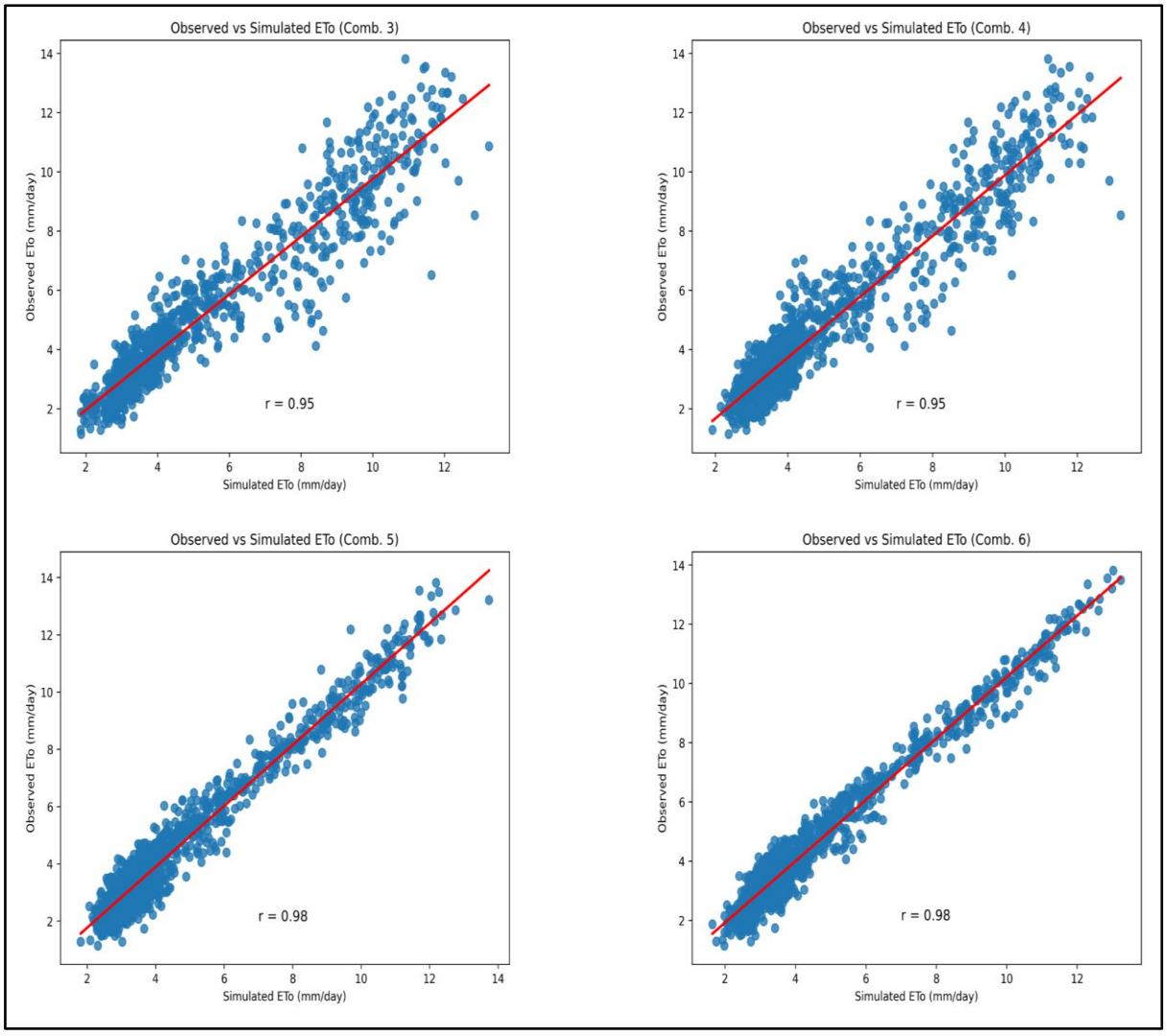


Figure 6. Simulated vs Observed ET0 Graphs for the LGBM Model. The scatterplots compare the simulated and observed ET0 for all the input combinations to the LGBM model on the test data. The closer the points are to the red line and 'r' is to 1, the better the LGBM is fit for the task.

6.2 Certain Interesting Results

To start off, combination 6, which included the most comprehensive set of weather variables, consistently delivered the best performance across both models. The ANN and LGBM model achieved their lowest MAE of 0.373 and 0.402 and RMSE of 0.469 and 0.499, along with the highest R^2 value of 0.972 and 0.968 respectively, indicating a strong correlation between the predicted and actual ET_0 values. This demonstrates the models' superior ability to generalise across different climatic conditions when provided with a rich set of input variables. This is consistent with previous research, which has shown that when a model is provided with additional weather factors originally used to calculate ET_0 via the PM equation, its accuracy in predicting ET_0 significantly improves [3, 9].

Furthermore, the findings of this study highlight the significant roles of wind speed and net radiation in accurately predicting ET_0 in semi-arid climates, aligning with previous research [3, 9]. While both factors are crucial, this study's results show that wind speed, included in Combination 5, achieved a higher R^2 of 0.953 for both models compared to net radiation's 0.917 and 0.916 for ANN and LGBM respectively in Combination 2. This challenges the conventional emphasis on radiation as the dominant factor after the

temperature variables in predicting the ET_0 [9], suggesting that wind speed is more critical for certain semi-arid regions like Indore. Hence, wind speed should not be underestimated as an input factor in predicting ET_0, as the importance of factors in accurately predicting ET_0 largely depends on local climatic conditions.

Interestingly, the ANN model struggles when additional variables like relative humidity (Combination 3) and surface pressure (Combination 4) are included alongside the core temperature variables. In fact, for Combination 3, the ANN's performance declines across all metrics compared to Combination 1, suggesting the model might be overfitting or capturing noise rather than meaningful patterns. While Combination 4 shows slight improvements in RMSE and R^2 for the ANN, the MAE increases, indicating inconsistent performance. In contrast, the LGBM model benefits from the inclusion of these additional variables, outperforming its performance in Combination 1 across all metrics. This suggests that LGBM is better equipped to leverage the extra data and use it to enhance its predictive ET_0 accuracy.

6.3 Consistency with Previous Results

When comparing this study's results to those of Yong et al. in Malaysia's tropical climate, a partial consistency is observed. Yong et al. found that ANN outperformed LGBM when all meteorological variables were available, while LGBM excelled with fewer inputs. In this study, ANN slightly outperforms LGBM in combinations 1 and 2, which use limited data. However, for combinations 3 and 4, LGBM surpasses ANN by a wider margin. In combination 5, both models perform similarly, though ANN has a slight edge in MAE and RMSE. When nearly all weather factors are included in combination 6, ANN significantly outperforms LGBM. Overall, ANN excels in combinations 1, 2, 5 (limited data), and 6 (all data), while LGBM performs better in combinations 3 and 4 (limited data). This suggests a potential general trend in model performance across different climates, warranting further research.

6.4 Limitations and Modifications

The study's major limitation is the accuracy of the calculated ET_0 values. According to the accuracy test, the calculated ET_0 values for a particular year (2016) showed an r^2 value 0.772 with the actual ET_0 values. Hence, the dataset has moderately accurate ET_0 values which might've limited the models' performance. For example, Yamfaç's 2019 [9] study performed in the semi-arid environment of Turkey had an ANN with lower MAE and RMSE and higher r^2 values. Additionally, due to inaccuracies in the dataset, the findings and analyses might contain inaccuracies. Thus, future work could delve into training the models with more accurate data, especially for net radiation and ET_0, providing greater certainty about the study's findings and analyses.

Secondly, the model types used are limited. In Yamaç's study [9], the KNN model outperformed the ANN and in Babazadeh et al. 's study ANFIS, a hybrid model, outperformed the ANN. Thus, future research could test such models in Indore's climate.

Lastly, the research lacks external validation across different climates. The study focuses only on Indore's semi-arid climate, so the findings may not apply to other semi-arid regions or different climates. For example, the discovery that wind speed is more critical than radiation in predicting ET_0 could be unique to Indore's specific conditions. Without testing the models in other regions, it's hard to know if these findings are universally valid or specific to Indore's microclimate. This reduces the broader applicability of the results. Hence, future work could test the models in more semi-arid regions to identify broader trends and patterns in model performance.

7. Conclusion

In conclusion, this study demonstrates the efficacy of AI models like ANN and LGBM in predicting ET_0 in Indore's semi-arid climate. The findings reveal that the ANN generally outperforms the LGBM model, particularly when provided with a comprehensive set of input variables. However, the LGBM model shows advantages in computational efficiency and may be better suited for scenarios with limited data as it outperforms the ANN by a high margin in combinations 3 and 4 and shows very similar performance to it across combinations 1, 2 and 5.

The study also highlights the critical role of wind speed and net radiation in ET_0 predictions, suggesting that wind speed may be more significant in regions like Indore. This insight challenges conventional wisdom and underscores the importance of considering local climatic conditions in ET_0 modelling.

While the results are promising, the study's limitations, including the accuracy of calculated ET_0 values and the focus on a single climatic region, suggest that further research is needed. Future work could involve testing these models in different semi-arid regions and exploring other AI models to enhance ET_0 prediction accuracy across diverse climates.

Hence, this study displays the massive potential of AI in solving the water crisis in Indore via efficient irrigation and simultaneously, realises the urgent need to spend, innovate and research further in this field.

8. Acknowledgements

Special thanks to Ms. Jenny Yang for her constant guidance and helpful insights throughout the entire paper-writing process.

9. References

1. Bharat, E., (2024, March 19), “Madhya Pradesh: Water Scarcity in Indore; Ground Water Level Reaches 560 Feet Below.”, ETV Bharat News, ETV Bharat. https://www.etvbharat.com/en/!bharat/madhya-pradesh-water-scarcity-in-indore-ground-water-level-reaches-560-feet-below-enn24031906854
2. Chauhan C., Panchal M., Dubey S., Sharma S., Surwade S., Dubey D., (2023), “Analysis of Groundwater Samples of Indore Region to Estimate Inorganic and Organic Particulates.” 2023 IJNRD, 8, 558. https://www.ijnrd.org/papers/IJNRD2304565.pdf
3. Yong S. L. S., Ng J. L., Huang Y. F., Ang C. K., (2023), “Estimation of Reference Crop Evapotranspiration with Three Different Machine Learning Models and Limited Meteorological Variables.”, Agronomy, 13(4), 1048. https://doi.org/10.3390/agronomy13041048
4. Kumar M., Jawaharlal A., Krishi Vishwavidyalaya N., (2019)., “ETo Tables for Districts of Madhya Pradesh (Daily, Weekly, Monthly and Annual Values)”. https://doi.org/10.13140/RG.2.2.19883.39203
5. “About District I District Indore, Government Of Madhya Pradesh | India”, (2024), Indore.nic.in. https://indore.nic.in/en/about-district/
6. “Indore climate: weather by month, temperature, rain - Climates to Travel”, (2020), Climatestotravel.com. https://www.climatestotravel.com/climate/india/indore
7. “Part A - Reference Evapotranspiration (ETo)”, (2024), Fao.org. https://www.fao.org/4/X0490E/x0490e05.htm
8. “Chapter 2 - FAO Penman-Monteith equation”, (2024), Fao.org, https://www.fao.org/4/X0490E/x0490e06.htm#fao%20penman%20monteith%20equation

9. YAMAÇ S. S, (2021), "Reference Evapotranspiration Estimation with k-Nearest Neighbour and Artificial Neural Network Models using different climate input variables in the semi-arid environment", Tarım Bilimleri Dergisi. https://doi.org/10.15832/ankutbd.630303
10. Bidabadi M., Babazadeh H., Shiri J., Saremi A., (2023), "Estimation reference crop evapotranspiration (ET0) using artificial intelligence model in an arid climate with external data", Applied Water Science, 14(1). https://doi.org/10.1007/s13201-023-02058-2
11. Feng Y., Cui N., Gong D., Zhang Q., Zhao L., (2017), "Evaluation of Random Forests and Generalized Regression Neural Networks for daily reference evapotranspiration modeling.", Agric Water Manag 193:163–173.
12. Elbeltagi A., Kushwaha N. L., Rajput J., Vishwakarma D. K., Kulimushi L. C., Kumar M., Zhang, J., Pande C. B., Choudhari P., Meshram S. G., Pandey K., Sihag P., Kumar N., Abd-Elaty I., (2022), "Modelling Daily Reference Evapotranspiration based on Stacking Hybridization of ANN with Meta-Heuristic Algorithms under Diverse Agro-Climatic Conditions", Stochastic Environmental Research and Risk Assessment. https://doi.org/10.1007/s00477-022-02196-0
13. "LightGBM (Light Gradient Boosting Machine)", (2020, July 15), GeeksforGeeks. https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/
14. Chugh A., (2020, December 8), "MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?" Medium, Analytics Vidhya. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Appendix A

All Sky Surface Albedo, All Sky Surface Shortwave Downward Irradiance and Clear Sky Surface Shortwave Downward Irradiance for each day were used to calculate the daily Net Radiation (MJ/m^2/day) as follows:

1. Firstly, the Net Shortwave Radiation (Rns) was calculated using:

$$R_{ns} = (1 - \alpha) \times R_s \quad (5)$$

Where:

- α is the surface reflection coefficient.
- R_s is the incoming solar radiation.

2. Next, the Net Longwave Radiation (Rnl) was calculated:

$$R_{nl} = \sigma \left(\frac{T_{max}+273.16}{2}\right)^4 \left(0.34 - 0.14\sqrt{e_a}\right)\left(1.35\frac{R_s}{R_{so}} - 0.35\right) \quad (6)$$

Where:

- σ is the Stefan-Boltzmann constant (4.903×10^{-9} MJ K^{-4} m^{-2} day^{-1}).
- T_{max} is the maximum temperature in Celsius.
- e_a is the actual vapour pressure, which was approximated using the humidity data.
- R_{so} is the clear sky solar radiation.

3. Thirdly, the Net Radiation (Rn) was calculated using**:**

$$R_n = R_{ns} - R_{nl} \quad (7)$$

Where:

- R_{ns} is the Net shortwave radiation (MJ/m²/day)
- R_{nl} is Net longwave radiation (MJ/m²/day)

Appendix B

To verify the accuracy of the calculated ET_0 values, the actual ET_0 data for 2016 was obtained from a research paper [4] and compared against the dataset's calculated values. The comparison was evaluated using three key metrics: MAE, RMSE, and R^2. The results showed an MAE of 0.874, an RMSE of 1.116, and an R^2 of 0.772. These metrics indicate that while the calculated ET_0 values in the dataset are moderately accurate, certain assumptions made during the calculation process likely introduced some discrepancies. Consequently, the dataset provides reasonably reliable ET_0 values, though there is room for improvement in accuracy.

Appendix C

Using the trained AI models, I've created an application called IndoreCropWaterWise that uses the ET_0 predicted by the model to calculate the Irrigation Requirements for a crop type at a certain growth stage, using further calculations and equations. Links:

- Project GitHub Repository
- App website
- YouTube video of project demonstration
- PPT for a detailed understanding of the workings of the application