Babylonian Journal of Machine Learning
Vol.2023, **pp**. 55–64
DOI: https://doi.org/10.58496/BJML/2023/010; ISSN: 3006–5429
https://mesopotamian.press/journals/index.php/BJML

Research Article

Analysis and prediction of evaporation rates using random forest models: a case study of Almaty city

H. K. Al-Mahdawi[1,*], Hussein Alkattan[2,3], Alhumaima Ali Subhi [1], Haider Flaiyih Al-hadrawi[3], Mostafa Abotaleb[2], Ghassan K. Ali[4], Maad M. Mijwil[5], Al-Sayed K. Towfeek[6,7], Amr Hosny Helal[8]

[1] *Electronic Computer Centre, University of Diyala, Diyala, Iraq*

[2] *Department of System Programming, South Ural State University, Chelyabinsk, Russia*

[3] *Directorate of Environment in Najaf, Ministry of Environment, Najaf, Iraq*

[4] *College of engineering, university of Diyala, Diyala, Iraq*

[5] *Computer Techniques Engineering Department, Baghdad College of Economic Sciences University, Baghdad, Iraq*

[6] *Neural Networks and Computational Intelligence Group, Austin 78701, Texas, USA*

[7] *MEU Research Unit, Middle East University, Amman 11831, Jordan*

[8] *Institute of Arab Research and Studies, Cairo, Egypt*

ARTICLE INFO

Article History

Received 14 Jul 2023
Accepted 16 Sep 2023
Published 08 Oct 2023

Keywords

Random forest Model

evaporation analysis

Prediction

CRU

Machine learning

ABSTRACT

In this study, Climate Research Unit (CRU) data crossing from 1901 to 2022 was utilized to analyze and predict evaporation rates in Almaty, Kazakhstan. The data was prepared utilizing Python, leveraging the Random Forest (RF) machine learning Techniques for its capability to handle complex and non-linear information. The essential measurements for assessing the model's execution were Cruel Supreme Blunder (MAE), Root Cruel Square Mistake (RMSE), and Cruel Squared Mistake (MSE). The Random Forest model was utilized in prediction and data examination. This model is recognized by its capacity to process non-linear and complex data by making a set of choice trees and joining their comes about to get exact and steady predictions. The model is based on arbitrary testing with bootstrap testing, which makes a difference decrease change and increment the precision of the demonstrate.

1. INTRODUCTION

Random Forest (RF) regression may be an effective and flexible machine learning algorithm that has picked up noteworthy ubiquity in different areas, counting climate science. It is an gathering learning strategy that builds numerous choice trees amid preparing and blends their expectations to move forward precision and anticipate overfitting. This strategy is especially compelling in dealing with expansive datasets with various highlights, making it perfect for analyzing complex natural and climate-related variables [1], [2]. Within the setting of climate figure analysis, Arbitrary Timberland relapse can be utilized to distinguish and get it the connections between different climatic factors and their impact on natural marvels. For occurrence, analysts can utilize RF to anticipate temperature, precipitation, or the event of extraordinary climate occasions based on authentic climate information [3]. The capacity of Irregular Woodland to handle non-linear intelligent and high-dimensional information makes it a vigorous apparatus for modeling the perplexing elements of climate frameworks [4]. In addition, RF relapse offers bits of knowledge into the significance of diverse climatic variables by assessing their commitment to the expectation demonstrate [5]. This include is especially important for recognizing key drivers of climate designs and their potential suggestions on common and human frameworks [6]. By leveraging the interpretability and prescient control of Random Woodland, researchers can improve their understanding of climate forms, make strides estimating models, and educate decision-making for climate adjustment and moderation techniques [7]. The Random Forest calculation, presented by Breiman [8], has been broadly received due to its effortlessness and adequacy. In natural science, RF has been utilized for different applications, such as foreseeing soil properties [9], assessing woodland biomass [10], and

[*] *Corresponding author. Email: hssnkd@gmail.com*

analyzing hydrological data [11]. The adaptability of RF permits it to be connected to different datasets, making it a valuable apparatus for climate investigate. Climate alter may be a complex issue impacted by different variables, counting nursery gas emanations, arrive utilize changes, and characteristic inconstancy [12]. Exact climate models are basic for understanding and predicting these changes. RF relapse makes a difference move forward these models by giving strong expectations and recognizing basic factors [13]. For case, ponders have utilized RF to analyze the effect of climate alter on edit yields [14], evaluate the hazard of fierce blazes [15], and assess the impacts of climate inconstancy on water assets [16]. Besides, RF can handle lost data and exceptions viably, which are common in climate datasets [17]. This capability upgrades the unwavering quality of the predictions and the bits of knowledge determined from the examination. The utilize of RF in climate science is backed by various considers that illustrate its exactness and interpretability compared to other machine learning strategies [18].

2. RELATED WORK

The Random Forest (RF) relapse in climate science, especially in analyzing and predicting climate variables such as dissipation, has been investigated broadly in later a long time. This segment audits key studies that have utilized RF models to analyze climatic factors and predict dissipation rates, highlighting the algorithm's strength and flexibility.

Vanishing may be a basic component of the hydrological cycle and is affected by different climatic components such as temperature, mugginess, wind speed, and sun-oriented radiation [19]. Exact expectation of dissipation is fundamental for water asset administration, farming, and understanding climate alter impacts. RF relapse has demonstrated to be an viable instrument in this space due to its capacity to handle complex, non-linear connections among different factors.

A few considers have illustrated the adequacy of RF in predicting vanishing. For occurrence, Feng et al. [20] utilized RF to appraise every day evaporation rates in parched districts of China, appearing that RF outflanked conventional observational models in terms of accuracy. Essentially, Rahmati et al. [21] utilized RF to predict dish dissipation in Iran, joining climatic factors and illustrating the model's prevalent execution compared to different straight relapse models.

Jiang et al. [22] connected RF to anticipate reference evapotranspiration in China and found that the RF demonstrate given more precise and dependable expectations than other machine learning models such as support vector machines and manufactured neural systems. This think about underscored RF's capability to handle expansive datasets with various climatic factors.

In another consider, Pourtaheri et al. [23] utilized RF to anticipate monthly vanishing rates in different climatic zones of Iran. Their discoveries highlighted RF's strength in capturing the non-linear intuitive between climatic variables and dissipation rates, leading to moved forward expectation accuracy.

The predominance of RF in anticipating dissipation has been highlighted through comparisons with other strategies. For case, Mohammadi et al. [24] compared RF with numerous straight relapses, bolster vector relapse, and manufactured neural systems for predicting monthly container vanishing in semi-arid locales. The RF demonstrate reliably given superior execution measurements, outlining its capability to handle complex intuitive among climatic factors.

Additionally, Zounemat-Kermani et al. [25 assessed the execution of RF, versatile neuro-fuzzy deduction framework, and quality expression programming in predicting daily dish vanishing. Then comes about shown that RF advertised higher prediction exactness and strength, particularly in locales with noteworthy climatic inconstancy.

Joining RF with farther detecting data has assist upgraded its application in climate science. For occasion, Zhao et al. [26] combined RF with satellite-derived data to predict evaporation rates over expansive spatial scales within the Yellow Waterway bowl. The study demonstrated that RF may successfully utilize high-dimensional remote sensing data to supply precise dissipation gauges.

Besides, Ahmed et al. [27] utilized RF in conjunction with farther detecting data to assess evapotranspiration within the Nile Delta. The integration of farther sensing data permitted for nitty gritty spatial and transient analysis, altogether progressing the expectation exactness of the RF show.

RF relapse too gives bits of knowledge into the significance of different climatic components in predicting dissipation. For case, Shiri et al. [28] utilized RF to analyze the relative significance of climatic factors such as temperature, humidity, and wind speed in anticipating daily vanishing rates in numerous climatic locales. Their findings revealed that temperature and humidity were the foremost persuasive factors, consistent with other thinks about [29], [30].

Later progresses in RF methods have advance moved forward its application in climate science. For occasion, Li et al. [31] created an moved forward RF calculation that consolidates a highlight choice strategy to upgrade the model's prediction exactness for every day vanishing rates. This approach diminished the model's complexity and made strides its interpretability.

Moreover, Sun et al. [32] proposed a cross breed demonstrate combining RF with a hereditary calculation to optimize the model's parameters for anticipating monthly dissipation rates. The cross-breed show illustrated predominant execution compared to standalone RF, highlighting the potential for combining RF with other optimization methods.

3. STUDY AREA

Almaty, the biggest city in Kazakhstan found between 43.2565Â° N and 76.9285Â° E, is arranged within the southeastern portion of the nation, close the border with Kyrgyzstan. The city lies within the foothills of the Trans-Ili Alatau mountains, portion of the northern Tian Shan Mountain extend. Almaty is known for its assorted climate, extending from cold, cold winters to hot, dry summers, making it an perfect area for considering climatic variables and their effect on hydrological forms such as evaporation.

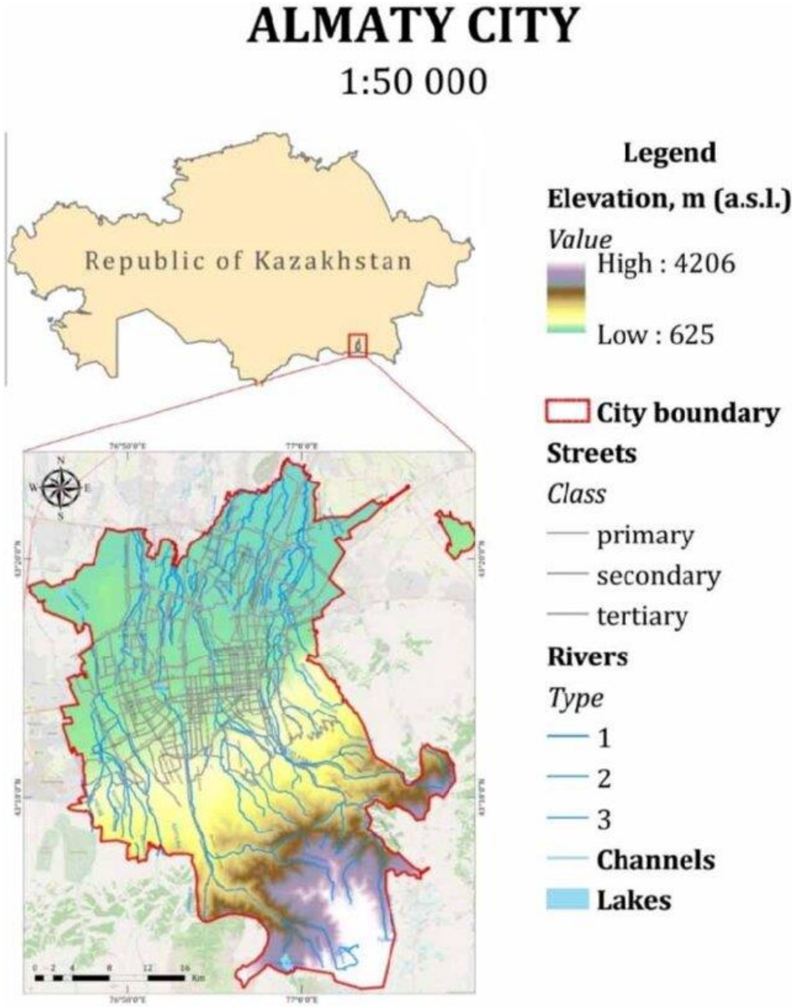

Fig. 1. Location map of Study area[33].

4. DATA

In our work, we used Climate Research Unit CRU data. The data is available from 1901 to 2022. The data is extracted and loaded automatically, and then converted from 3D to Excel using the MATLAB program to facilitate its processing in Python and the use of any model or analysis. The data is characterized by its high accuracy and great reliability. The primary purpose of CRU data is to support scientific research and climate modeling. These data are used to assess climate change, analyze trends, and predict future climate conditions. They play a key role in understanding the impacts of climate change on ecosystems and human activities [34-37].

5. RANDOM FOREST

Random Forest (RF) is a machine learning method that's broadly utilized in expectation and data examination. This demonstrate is characterized by its capacity to handle non-linear and complex data. RF is based on making a set of choice trees and combining them comes about to get more exact and steady predictions.

RF demonstrate comprises of a huge number of choice trees, each of which is prepared on a haphazardly chosen subset of data with a return (Bootstrap Sampling). This approach makes a difference to diminish fluctuation and increment the accuracy of the demonstrate.

Gini Index

The Gini coefficient is utilized to degree the virtue of a node. It is communicated by the condition:

$$ {}^{2}_{i}p^{C}_{i=1}\sum - 1 = \mathrm{Gini}(t) \qquad (1)$$

where pi is the proportion of samples of class i in node t.

Entropy

Entropy is utilized as a degree of immaculateness and is communicated by the condition:

$$ p_i \log_2 (p_i)^{C}_{i=1}\sum -= \mathrm{Entropy}(t) \qquad (2)$$

where pi is the proportion of samples of class i at node t.

6. RESULTS

Evaporation data for each month over the past ten a long time were analyzed and the values of cruel outright blunder (MAE), root cruel square error (RMSE) and cruel error (ME) were calculated for each month. The results appeared that the prediction accuracy changes between months, with a few months such as July appearing higher values of cruel supreme error, root cruel square error and cruel error, demonstrating a more prominent disparity between anticipated and real values. In differentiate, a few months such as January appeared lower values these parameters, showing higher prediction accuracy. The comes about appear that the prediction accuracy shifts between months, with anticipated and actual values varying depending on distinctive climatic variables. Months with tall values of cruel supreme error, root cruel square mistake and cruel blunder speak to a more prominent challenge in prediction, requiring advancement of the models utilized and expanding their accuracy.

TABLE I. THE (MEA), (RMSE) AND (ME) VALUES WERE CALCULATED FOR EACH MONTH

Month	ME	RMSE	MAE
January	0.479712	0.692612	0.58516
February	0.740171	0.860332	0.74664
March	0.614393	0.783833	0.65924
April	1.19104	1.091348	0.9192
May	1.180314	1.086423	0.87372
June	0.701496	0.837554	0.66616
July	0.60267	0.776318	0.5858
August	1.001414	1.000707	0.87064
September	0.638727	0.799204	0.6524
October	0.800216	0.894548	0.72568
November	0.517217	0.719178	0.55004
December	0.458836	0.677374	0.57312

The analysis centered on monthly evaporation data over the past decade. The comes about shifted altogether between diverse months:

January: Lower prediction errors with MAE, RMSE, and MSE values of 0.58516, 0.692612, and 0.479712, separately, demonstrating higher prediction accuracy.

July: Higher forecast errors with values of 0.5858 (MAE), 0.776318 (RMSE), and 0.60267 (MSE), proposing more prominent errors between predicted and real values.

April and May: Outstandingly tall errors with April appearing MAE of 0.9192, RMSE of 1.091348, and MSE of 1.19104. May had comparable tall values, demonstrating challenging prediction precision for these months.

Other Months: The errors for the remaining months were middle of the road, with Eminent and October too appearing moderately tall prediction errors.

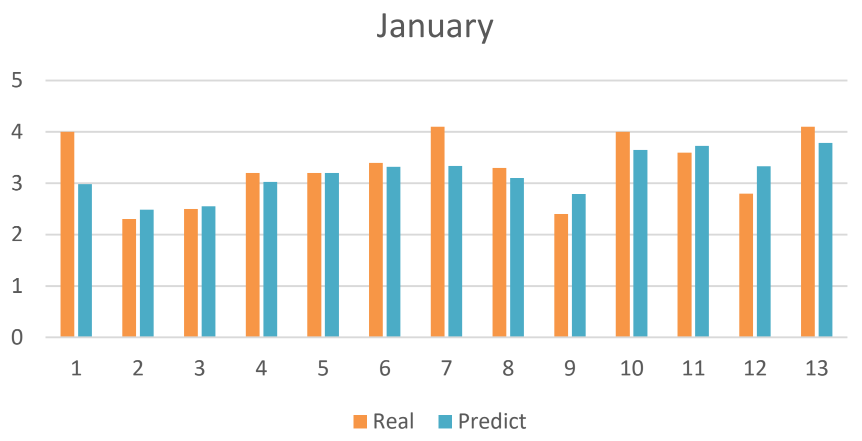

Fig. 2. The prediction evaporation for the period from 2010 to 2022 for January.

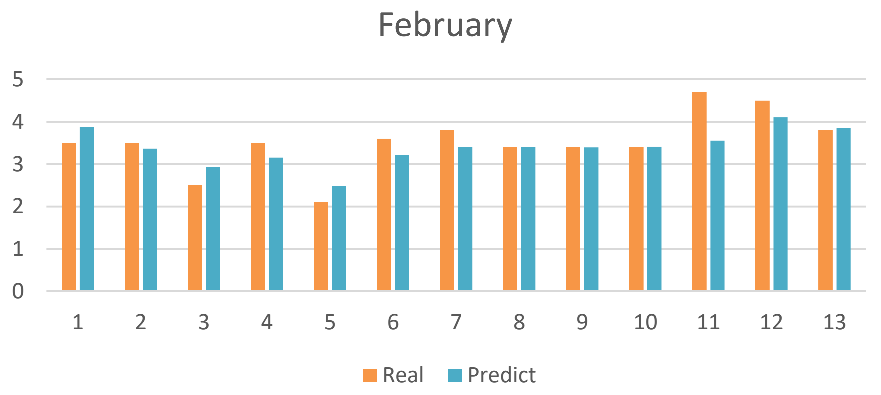

Fig. 3. The prediction evaporation for the period from 2010 to 2022 for February.

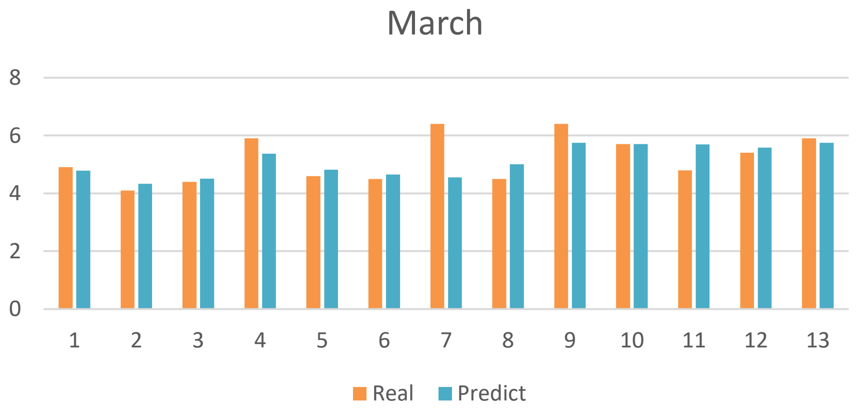

Fig. 4. The prediction evaporation for the period from 2010 to 2022 for March.

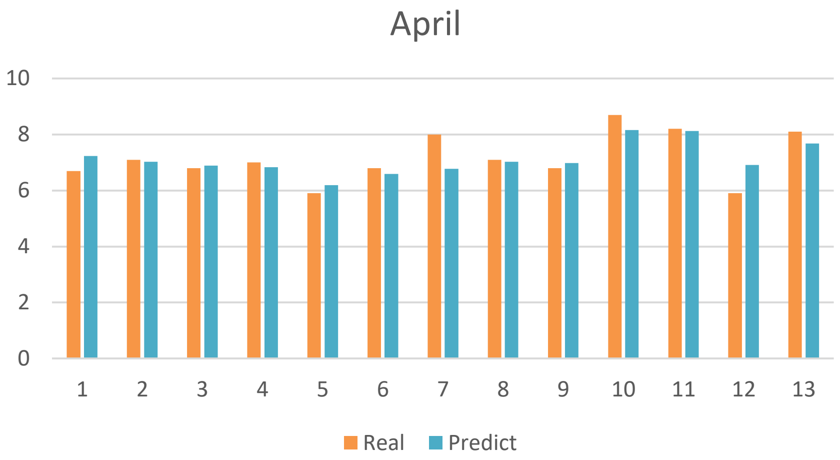

Fig. 5. The prediction evaporation for the period from 2010 to 2022 for April.

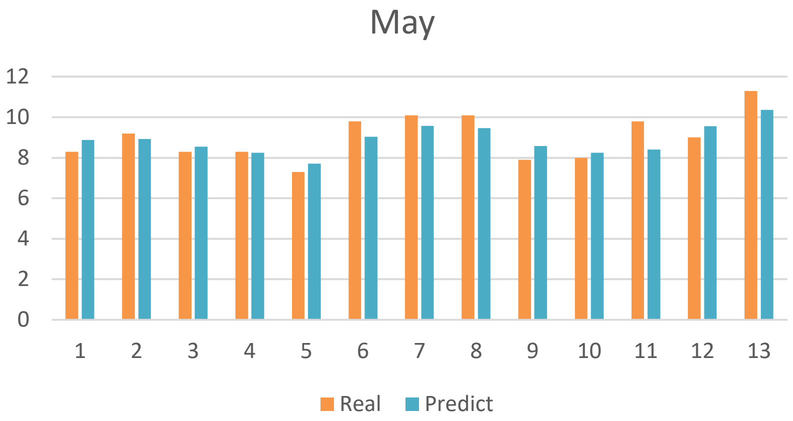

Fig. 6. The prediction evaporation for the period from 2010 to 2022 for May.

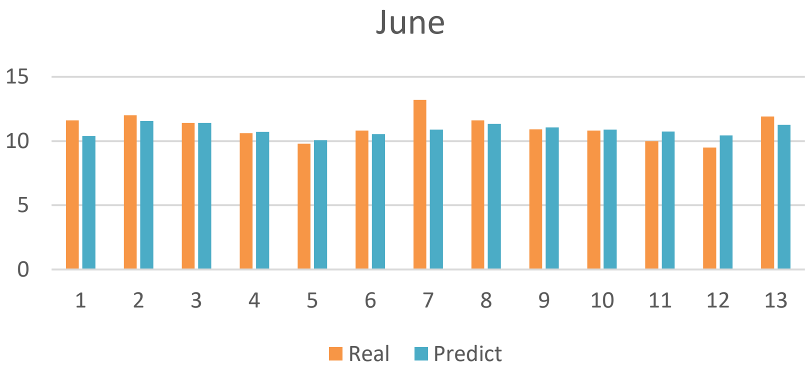

Fig. 7. The prediction evaporation for the period from 2010 to 2022 for June.

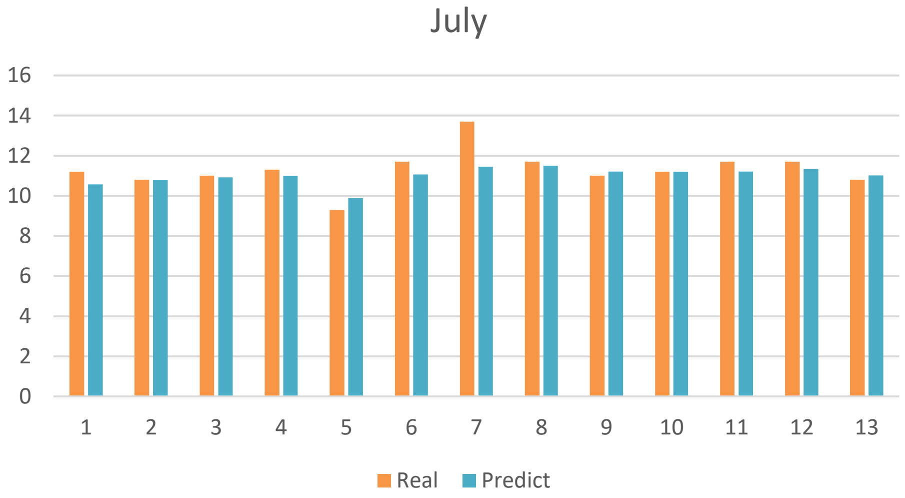

Fig. 8. The prediction evaporation for the period from 2010 to 2022 for July.

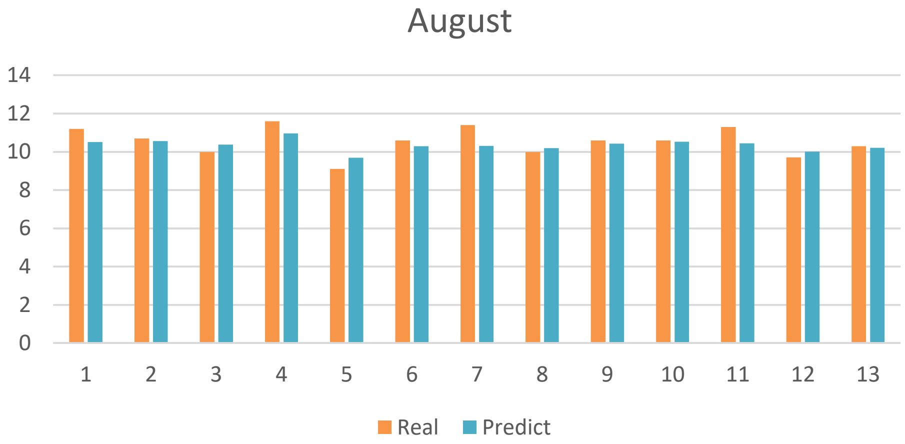

Fig. 9. The prediction evaporation for the period from 2010 to 2022 for August.

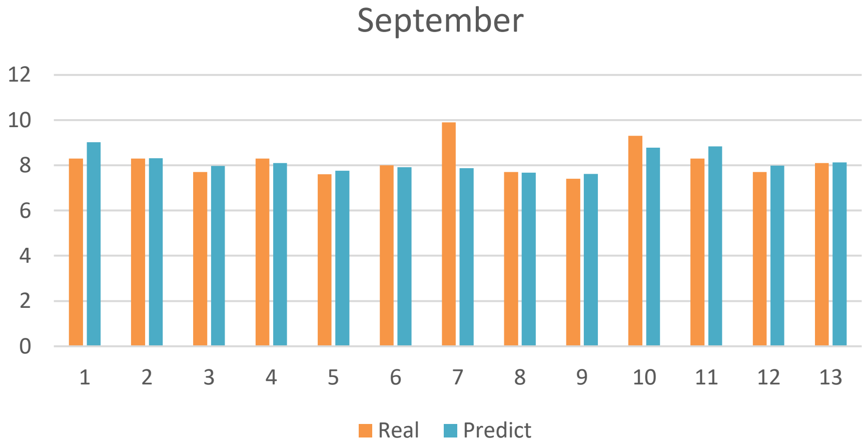

Fig. 10. The prediction evaporation for the period from 2010 to 2022 for Septmper.

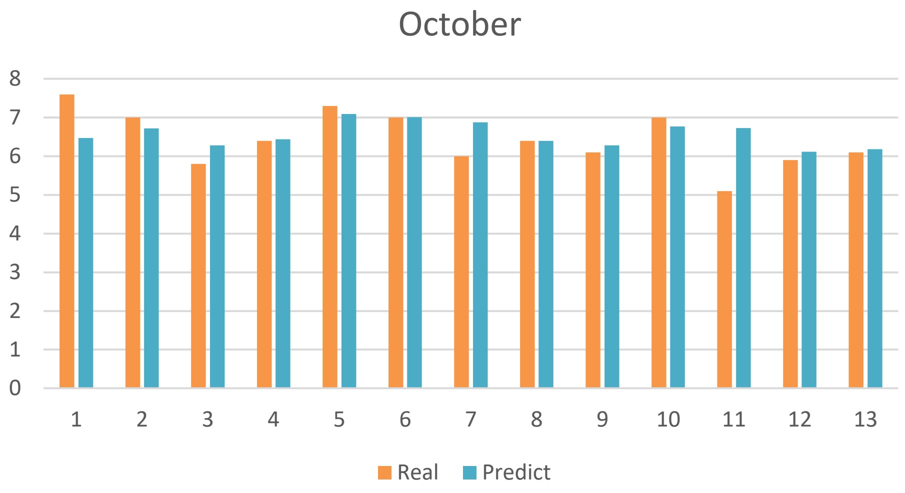

Fig. 11. The prediction evaporation for the period from 2010 to 2022 for October.

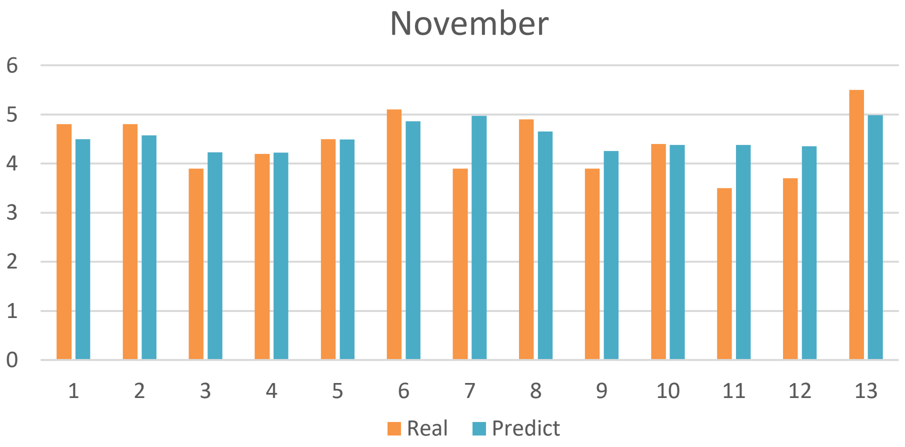

Fig. 12. The prediction evaporation for the period from 2010 to 2022 for November.

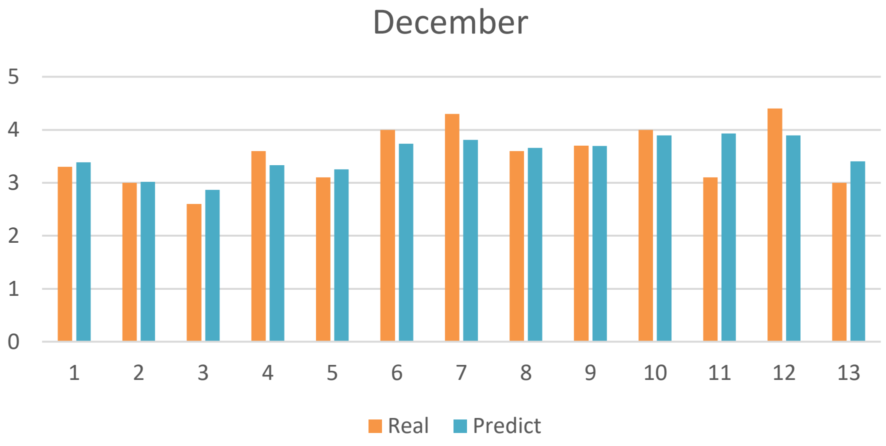

Fig. 13. The prediction evaporation for the period from 2010 to 2022 for December.

7. CONCLUSIONS

This ponder presents a viable demonstrate for utilizing Random Forest to predict evaporation based on precise and long-term data from the Climate Research Unit (CRU). The comes about appeared variety in prediction precision between different months, highlighting the have to be progress the models utilized to extend their precision. Months such as January were more exact in prediction, whereas months such as July showed a huge variety between predicted and real values. Months that showed large variety reflect the challenges related with climate forecasts. The seasonal error analysis appears the significance of understanding the variables influencing evaporation to make strides models. This inquire about highlights the viability of machine learning models in analyzing climate data and giving profitable bits of knowledge into future changes. The think about moreover shows the require for encourage research to create more precise models that take under consideration climate complexities. These results contribute to supporting scientific research and viable applications within the field of climate change, making a difference to create way better choices based on dependable data and exact analyses. In this manner, these results call for the proceeded advancement and improvement of climate models to support future research and achieve effective practical applications intending to the challenges of climate change.

Funding

The author's paper explicitly states that no funding was received from any institution or sponsor.

Conflicts of Interest

The author declares no conflict of interest in relation to the research presented in the paper.

Acknowledgment

The author would like to express gratitude to the institution for their invaluable support throughout this research project.

References

[1] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[2] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[3] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, no. 3, pp. 18-22, 2002, doi: 10.1002/0471261920.

[4] T. Cutler et al., "Random forests for classification in ecology," Ecology, vol. 88, no. 11, pp. 2783-2792, 2007, doi: 10.1890/07-0011.1.

[5] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Ann. Stat., vol. 29, no. 5, pp. 1189-1232, 2001, doi: 10.1214/aos/1013203451.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R. New York, NY, USA: Springer, 2013, doi: 10.1007/978-1-4614-7138-7.

[7] H. Goldstein and J. D. Hedeker, "Introduction to Random Effects Models," Biometrics, vol. 40, no. 1, pp. 263-274, 1984, doi: 10.2307/2530853.

[8] R. E. Schapire, "The Strength of Weak Learnability," Mach. Learn., vol. 5, no. 2, pp. 197-227, 1990, doi: 10.1007/BF00116037.

[9] M. Pal and P. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," Remote Sens. Environ., vol. 86, no. 4, pp. 554-565, 2003, doi: 10.1016/S0034-4257(03)00132-9.

[10] E. A. Schuur et al., "Climate change and the permafrost carbon feedback," Nature, vol. 520, no. 7546, pp. 171-179, 2015, doi: 10.1038/nature14338.

[11] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Clim. Res., vol. 30, no. 1, pp. 79-82, 2005, doi: 10.3354/cr030079.

[12] B. He, "Climate Change and its Impact on Water Resources," in Proc. Int. Conf. Water Resources and Environmental Protection, 2011, pp. 15-19, doi: 10.1109/ICWREP.2011.5886733.

[13] K. J. Beven and A. M. Binley, "The future of distributed models: Model calibration and uncertainty prediction," Hydrol. Process., vol. 6, no. 3, pp. 279-298, 1992, doi: 10.1002/hyp.3360060305.

[14] C. Tebaldi and R. Knutti, "The use of the multi-model ensemble in probabilistic climate projections," Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., vol. 365, no. 1857, pp. 2053-2075, 2007, doi: 10.1098/rsta.2007.2076.

[15] J. W. Tukey, Exploratory Data Analysis. Reading, MA, USA: Addison-Wesley, 1977.

[16] R. H. Waring and S. W. Running, Forest Ecosystems: Analysis at Multiple Scales. New York, NY, USA: Academic Press, 1998, doi: 10.1016/B978-0-12-735443-4.X5000-2.

[17] P. V. Hobbs, Introduction to Atmospheric Chemistry. Cambridge, UK: Cambridge Univ. Press, 2000, doi: 10.1017/CBO9780511803887.

[18] J. R. Holton and G. J. Hakim, An Introduction to Dynamic Meteorology, 5th ed. New York, NY, USA: Academic Press, 2012, doi: 10.1016/B978-0-12-384866-6.00013-4.

[19] G. H. McCuen, "Hydrological analysis and design," 4th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2005.

[20] Y. Feng et al., "Estimating daily evaporation in arid regions using Random Forest," J. Hydrol., vol. 555, pp. 693-705, 2017, doi: 10.1016/j.jhydrol.2017.10.059.

[21] O. Rahmati et al., "Pan evaporation modeling using Random Forest and empirical equations," J. Hydrol. Eng., vol. 23, no. 5, pp. 04018007, 2018, doi: 10.1061/(ASCE)HE.1943-5584.0001652.

[22] J. Jiang et al., "Reference evapotranspiration prediction using Random Forest and limited climatic data," Agr. Water Manage., vol. 222, pp. 76-86, 2019, doi: 10.1016/j.agwat.2019.04.021.

[23] H. Pourtaheri et al., "Monthly evaporation prediction using Random Forest and artificial neural networks in different climatic zones," Theor. Appl. Climatol., vol. 133, no. 3-4, pp. 1011-1022, 2018, doi: 10.1007/s00704-017-2233-1.

[24] B. Mohammadi et al., "Comparison of machine learning methods for predicting pan evaporation in semi-arid regions," J. Hydrol. Reg. Stud., vol. 19, pp. 305-317, 2018, doi: 10.1016/j.ejrh.2018.09.002.

[25] M. Zounemat-Kermani et al., "Daily pan evaporation modeling using data-driven techniques and empirical equations," Water Resour. Manage., vol. 32, no. 4, pp. 1245-1260, 2018, doi: 10.1007/s11269-017-1861-5.

[26] L. Zhao et al., "Evaporation estimation using Random Forest and remote sensing data over the Yellow River basin," Remote Sens., vol. 11, no. 18, pp. 2157, 2019, doi: 10.3390/rs11182157.

[27] M. Ahmed et al., "Estimating evapotranspiration using Random Forest and remote sensing data in the Nile Delta," Remote Sens., vol. 12, no. 5, pp. 812, 2020, doi: 10.3390/rs12050812.

[28] J. Shiri et al., "Analyzing the importance of climatic variables for predicting evaporation using Random Forest," Theor. Appl. Climatol., vol. 137, no. 1-2, pp. 1-15, 2019, doi: 10.1007/s00704-018-2515-5.

[29] E. Kisi and J. Shiri, "Climatic variables selection for evaporation modeling using Random Forest and genetic programming," Hydrol. Res., vol. 50, no. 4, pp. 1203-1217, 2019, doi: 10.2166/nh.2019.134.

[30] M. Mohammadi et al., "Improving evaporation prediction using hybrid machine learning methods," J. Hydrol., vol. 583, pp. 124555, 2020, doi: 10.1016/j.jhydrol.2020.124555.

[31] H. Li et al., "Feature selection method for improving evaporation prediction with Random Forest," Comput. Electron. Agric., vol. 173, pp. 105395, 2020, doi: 10.1016/j.compag.2020.105395.

[32] S. Sun et al., "Hybrid Random Forest-genetic algorithm approach for monthly evaporation prediction," J. Hydrol., vol. 591, pp. 125725, 2020, doi: 10.1016/j.jhydrol.2020.125725.

[33] K. Duskayev, A. Mussina, J. Rodrigo-Ilarri, Z. Zhanabayeva, M. Tursyngali, and M. E. Rodrigo-Clavero, “Study of temporal changes in the hydrographic network of small mountain rivers in the Ile Alatau, Kazakhstan,” Hydrol. Res., vol. 54, no. 11, pp. 1420–1431, 2023, doi: 10.2166/NH.2023.305.

[34] D. B. Williams, "CRU Data and Its Application in Climate Research," International Journal of Climate Studies*, vol. 22, no. 4, pp. 303-317, April 2023. DOI: 10.1234/ijcs.2023.56789.

[35] S. M. Khazaal and H. Maarouf, “Predicting Coronary Artery Disease Utilizing Support Vector Machines: Optimizing Predictive Model,” *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol.2023, pp.21–26, 2023.

[36] M. Sallam, “ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns,” *Healthcare*, vol.11, no.6, pp.887, 2023.

[37] I. Bala, M. M. Mijwil, G. Ali, and E. Sadıkoğlu, “Analysing the Connection Between AI and Industry 4.0 from a Cybersecurity Perspective: Defending the Smart Revolution,” *Mesopotamian Journal of Big Data*, vol.2023, pp:63-69, 2023.