The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

Advancing Ocean Level Prediction with Machine Learning

Nilufar Makky [1*] Ahmad Moradi [2]

Abstract

Effectively monitoring sea surface height (SSH) presents a considerable contemporary challenge. By proposing an integrated approach, this paper indicates the specific complexities associated with SSH monitoring in the Oman SeaThe methodology involves using radar altimetry data from Sentinel-3A, along with meteorological parameters such as sea surface pressure, temperature, precipitation, and seawater vapor obtained from Google Earth Engine cloud-based platform (GEE). The acquired data from Sentinel-3A is meticulously corrected and analyzed to provide accurate and insightful insights into SSH within the specified region, focusing on the 4-month year 2023. This integrated approach enhances the precision and reliability of sea surface height monitoring in the challenging maritime environment of the Oman Sea. The research compares their efficacy in predicting SSH using machine learning algorithms, including Multilayer Perceptron, Support Vector Regression, Random Forest, Gradient Boosting, and K-Nearest Neighbors. Among these, the Random Forest model presents the better Mean Absolute Error of 0.0554, R-squared value of 0.9818, Root Mean Squared Error of 0.1101, and Mean Squared Error of 0.0121. This affirms the model's exceptional accuracy in capturing sea surface height dynamics, emphasizing the significance of incorporating meteorological parameters for a comprehensive understanding and accurate prediction of SSH in the Oman Sea. The findings suggest potential applications in improving operational oceanographic forecasting and advancing our knowledge of the intricate interactions between meteorological conditions and sea surface height variations.

Keywords: Sea surface height, Radar altimetry data, Oman, GEE, Sentinel-3A.

1 Introduction

Global warming is now a major concern worldwide [1]. Over the past ten years, the average global surface temperature has risen by more than 1.2 °C compared to the pre-industrial era [2]. Climate change [3] has been associated with rising global temperatures, leading to more frequent and intense extreme weather occurrences [4]. Coastal areas are facing numerous challenges due to climate warming and the rapid rise in sea levels. These challenges include the risk of flooding, erosion, heightened ecological vulnerability, and other harmful transformations. To effectively protect and minimize these impacts, it is crucial to gain a deeper understanding of how coasts respond to changes in sea levels [5]. Satellite altimetry (SA) faces a significant obstacle in obtaining precise sea surface information by detecting radar echoes along the Earth's surface, also known as track height data near the shoreline, particularly in sea areas with complex landforms [6]. The fundamental principle of radar altimetry involves the measurement of the time it takes for a microwave radar pulse to travel from the antenna to the surface of the Earth. Valuable information can be obtained by determining this range, which represents the distance between the satellite and the Earth's surface. For over three decades, this method has been employed to investigate various aspects of physical oceanography, including ocean surface topography, marine gravity, ocean circulation, bathymetry, and sea-level rise [7]. Additionally, in contrast to imaging sensors, altimeters serve as profiling systems, gathering information by detecting radar echoes along the Earth's surface, also known as tracks. These radar echoes are captured as waveforms, which represent a histogram of energy reflected by the ground surface over time. While satellite altimetry has been extensively utilized in hydrological and cryosphere studies, limited research has focused on lake ice, particularly in comparison to sea ice and the measurement of lake water levels [8]. The altimeters on Sentinel-3A and -3B are special because they are the first ones globally to use delay-doppler (SAR) mode everywhere on Earth. This helps get a clearer view of the Earth's surface and reduces noise through a technique called multi-looking. The radiometer, which looks straight down (nadir), has two channels. Regular checks on the instruments involve looking at their telemetered data, calibrating them at specific places, and comparing their readings with

[1] Remote sensing and GIS Department, College of Environmental Engineering, University of Tabriz (rsgis.nilufarmakki73@gmail.com)
[2] Geodesy Department, College of Surveying, KNT University (amoradi@email.kntu.ac.ir)

models, on-site measurements, and other satellites. All of this is done to keep an eye on daily data production, figure out uncertainties and errors, and understand how well they work in the long run for climate science [9]. Regarding the need for Sea Surface Height (SSH) Monitoring the Sentinel-3 mission consists of two satellites, namely Sentinel-3A and Sentinel-3B. Both satellites have the same instruments, including a dual-frequency synthetic aperture radar altimeter (SRAL) and a dual-band passive MWR [10]. The Sentinel-3 payload includes a synthetic aperture radar altimeter

(SRAL) that operates at both C and Ku-band frequencies. The SRAL instrument has two measurement modes: low-resolution mode (LRM) and SAR mode. In SAR mode, the SRAL instrument emits 64 bursts of Ku-band pulses flanked by two C-band pulses, offering an along-track resolution of approximately 300 meters. This study exclusively utilized data obtained at the Ku-band frequency [11]. Furthermore, by employing remote sensing techniques, we can integrate GIS technology [12,13] to improve our capacity for establishing connections between features and data. Weather information obtained from Google Earth Engine (GEE) is utilized to analyze the variations[14].

2 Material and Methods

European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) provided Radar altimetry data for sea surface monitoring [15]. In this study, the SRAL Level 2 Altimetry Global - Sentinel-3A, available from 10 March 2023 (Ku band), was employed for monitoring. SRAL level-2 data includes standard altimetry measurements like altimeter range, sea surface height, wind speed, and significant wave height. More of these data are in net CDF (network Common Data Form) format [16]. Furthermore, data were extracted from April and July 2023, additionally, they were processed to obtain reliable data for monitoring SSH. Additionally, environmental parameters such as water evaporation, precipitation, sea surface temperature, and sea surface air pressure were examined to indicate their performance in monitoring. The data mentioned previously were used to model SSH in the Oman Sea. Moreover, environmental parameters were prepared for April and July via the Google Earth Engine cloud-based platform which is a powerful platform for remote sensing analysis especially for water body monitoring [17,18] in the same period. Recent advancements in machine learning (ML) techniques [19] especially Random Forest (RF) is a powerful machine learning algorithm that can assess the significance of predictor variables and produce highly precise results without encountering the overfitting issue [20]. KNN, Gradient Boosting, SVR, MLP, and RF models were employed to model water height based on these parameters which are presented in Table 1.

Table 1: Environmental Parameters which were used in this study

Sea Surface Meteorological Parameters	Missions
Sea Surface Height	SRAL Level 2 Altimetry Global - Sentinel-3
Sea Surface Temperature	MODIS/006/MOD11A2
Sea Surface Pressure	NCEP_RE/sea_level_pressure
Sea Surface Precipitation	chirps
Sea Surface Water Vapor	NCEP_RE/surface_wv

2.1 Study area

The Sultanate of Oman, boasting a coastal stretch of over 3000 km along the Arabian Sea, the Sea of Oman, and the Arabian Gulf, faces environmental challenges due to the escalating intensity of storms attributed to global climate change. Coastal inundation has become inevitable, and the extent of flooding is influenced by the geomorphologic and oceanographic characteristics of the coastal zone. Despite the complex coastal ecology shaped by two monsoon seasons and substantial nutrient injection, Oman is undergoing rapid development fueled by fossil hydrocarbon exploitation. Economic diversification efforts in tourism, fisheries, aquaculture, and port services necessitate extensive shore infrastructure development, posing a threat to the coastal environment. While regulations are in place, enforcement is challenging due to the lengthy coastline and swift economic development. Unique features, including a robust summer thermocline, may offer some protection against excessive temperatures, but the rise of the thermocline poses a risk of hypoxic water intrusion, potentially leading to increased mortalities in shallow communities. Global changes are anticipated to impact Oman, albeit with distinctive features making it less vulnerable compared to other parts of the Indian Ocean. [21,22].

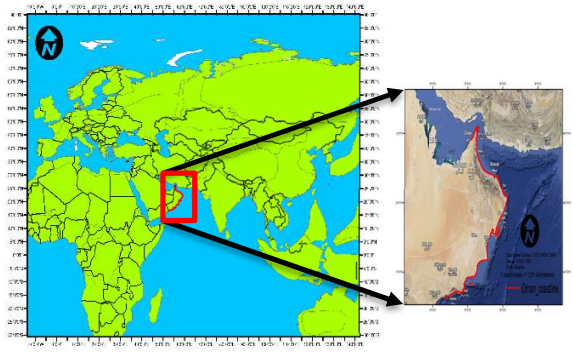

Figure 1: Study area

2.2 Methodology

The methodology employed in this paper is elucidated in the following sections, providing a comprehensive and detailed insight into the approaches, techniques, and procedures adopted for the research. The step-by-step explanation covers various aspects, including data collection, processing, analysis, and modeling, ensuring a thorough understanding of the methodology's intricacies. This comprehensive exposition aims to facilitate clarity and transparency, allowing readers to grasp the nuances of the research methodology employed in this study.

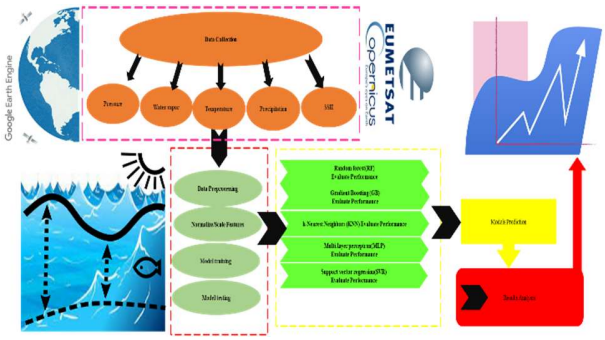

Figure 2: Methodology

3 Results and discussion

K-NN, SVR, MLP, RF, and Gradient Boosting (GB) algorithms in our study have opted to indicate that machine learning methods can be successfully applied for predictions [23,24].

By preprocessing altimetry SRAL Level2 data and providing meteorological parameters using satellite imagery [25] which was prepared by GEE, we gathered the required data for modeling them with the mentioned algorithms. Moreover, all provided data were put into ArcMap 10.8 [26] so that it's all the same size and the algorithms can understand it better. 80% of data is used for training and 20% for testing [27]. Regarding the compatibility between the real data and the predicted data for All, Train, and Test data which is indicated in Figure 3, The length of the graphs represents the actual or observed value and the width represents the predicted value.

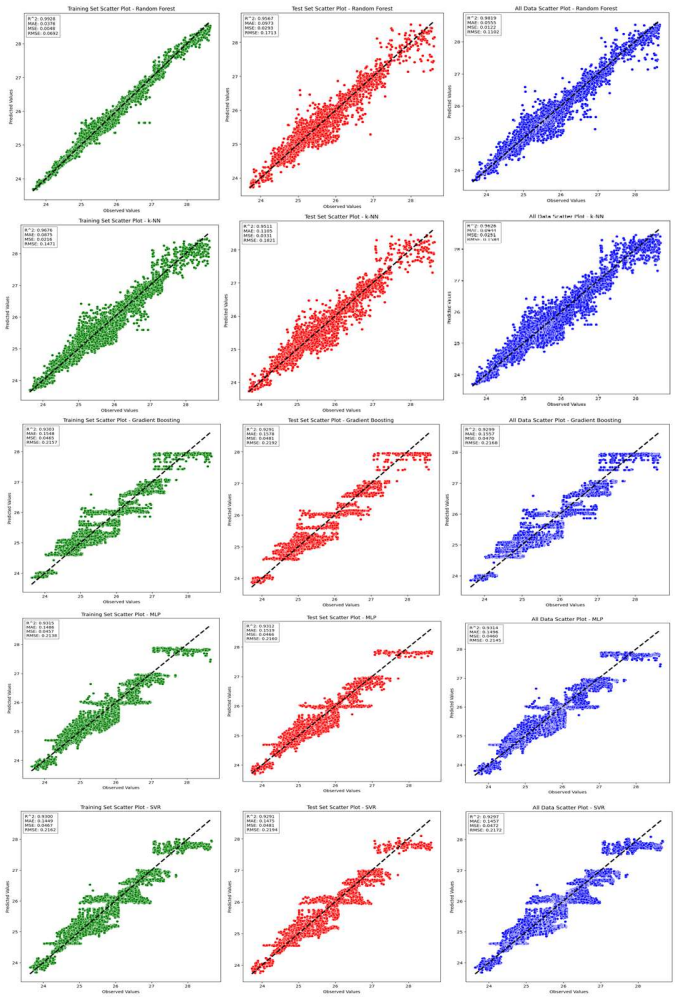

Figure 3: Compatibility between the real data and the predicted for All, Train, and Test data of KNN, RF, Gradient Boosting, MLP, and SVR models.

3.1 Histograms of residuals

Histograms of residuals were generated for the Random Forest (RF), MLP, Gradient Boosting, SVR, and K-NN models to visually represent the distribution of differences between predicted and actual values. In a rigorous scientific analysis, a well-centered and symmetric histogram is indicative of accurate predictions, suggesting that the models align closely with observed values. Conversely, the presence of patterns or skewness in the histogram may signal potential limitations or biases within the models, warranting further scrutiny and refinement.

These histograms serve as a valuable tool for assessing the predictive performance and reliability of the machine-learning algorithms employed in the study [28]. Residuals in the proportional hazards model proposed by Cox (1972) are employed for testing the assumption of proportional hazards. These residuals can be graphically plotted against time to assess the validity of the proportional hazards assumption. Furthermore, histograms of these residuals serve as a diagnostic tool to scrutinize the model fit and identify any potential outliers in covariate values[29].

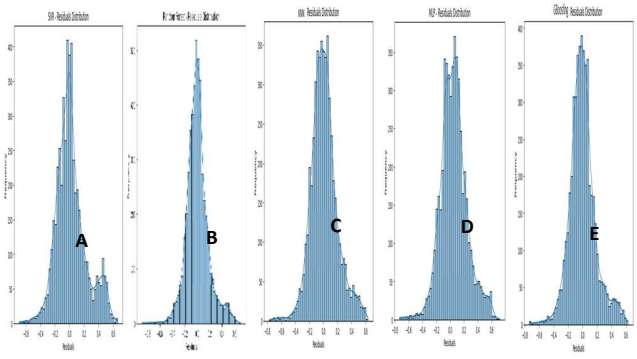

Figure 4: Residual of distribution for SVR(A), RF (B), KNN (C), MLP (D), and GB (E) models.

3.2 Accuracy Assessment

The obtained results demonstrate commendable performance from all algorithms, with R2 values exceeding 0.92 across models. Notably, the Random Forest (RF) model outshone others with a MAE of 0.0554, an impressive R2 value of 0.9818, an RMSE of 0.1101, and an MSE of 0.0121. In contrast, SVR exhibited the lowest accuracy, achieving an R2 of 0.9297, a MAE of 0.1456, an MSE of 0.0471, and an RMSE of 0.2171. The evaluation of predictive capabilities involved the utilization of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R-squared (R2), and Mean Squared Error (MSE) [30].

- ✔ $MAE\ (\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y|$
- ✔ $MSE\ (\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i}, y_i\ )^2$
- ✔ $RMSE(\hat{y}, y) = R\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i}, y_i\ )^2}$
- ✔ $R^2\ (\hat{y}, y) = 1 - \frac{SSres\ (\hat{y}, y)}{SStot\ (\hat{y}, y)}$

In the following formulas, $\hat{y}$ and y represent the average values of the actual output and predicted output respectively. $\hat{y_i}$ and y_i represent the true value and predicted value at a specific moment. n denotes the length of the training samples. Res SS refers to the sum of squares of the residuals and tot SS represents the total sum of squares of the real data [31]. The variable n signifies the length of the training samples. Res SS corresponds to the sum of squares of the residuals, and tot SS represents the total sum of squares of the real data, as detailed in reference. This explanation provides an expanded elucidation of the symbols and terms used in the formulas, enhancing comprehension of their significance within the context of the study.

Table 2: Illustrating R^2, MSE, RMSE, and MAE for each model (cm).

Model	Dataset	MAE	MSE	RMSE	R-squared
SVR	All Data	0.145	0.047	0.217	0.929
SVR	Train	0.144	0.046	0.216	0.93
SVR	Test	0.147	0.048	0.219	0.929
Random Forest	All Data	0.055	0.012	0.11	0.981
Random Forest	Train	0.037	0.004	0.069	0.992
Random Forest	Test	0.097	0.029	0.171	0.956
MLP	All Data	0.154	0.048	0.22	0.927
MLP	Train	0.153	0.047	0.219	0.928
MLP	Test	0.157	0.049	0.223	0.926
k-NN	All Data	0.094	0.025	0.158	0.962
k-NN	Train	0.087	0.021	0.147	0.967
k-NN	Test	0.11	0.033	0.182	0.951
Gradient Boosting	All Data	0.155	0.047	0.216	0.929
Gradient Boosting	Train	0.154	0.046	0.215	0.93
Gradient Boosting	Test	0.157	0.048	0.219	0.929

All applied machine learning algorithms, including Support Vector Regression (SVR), Random Forest, Multi-layer Perceptron (MLP), k-Nearest Neighbors (k-NN), and Gradient Boosting, demonstrated exceptional performance, as outlined in Table 2. The performance metrics, encompassing Mean Absolute Error (MAE),

Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared, were thoroughly assessed for each model across the entire dataset, training subset, and testing subset. Notably, the Random Forest model exhibited superior accuracy with lower MAE, MSE, and RMSE, coupled with a notably high R-squared value of 0.981 on the entire dataset, affirming its effectiveness in predicting sea surface height. This underscores the influence of meteorological parameters on the precision of sea surface height predictions, emphasizing the reliability of SRAL Level 2 data. Consequently, the study strongly recommends continuous monitoring of sea surface height under diverse spatiotemporal conditions, acknowledging the potential impact of climate change.

4. CONCLUSIONS AND RECOMMENDATIONS

SSH monitoring has become increasingly important due to global warming. Moreover, various missions have provided Altimetry data for monitoring sea surface height or water level since 1978. This study used SRAL Level 2 Altimetry - Sentinel-3 in 2023 to monitor SSH in the Oman Sea from April to July. Machine learning algorithms were used to monitor SSH based on four meteorological parameters: precipitation, temperature, water vapor, and air pressure. The algorithms used in this study included MLP, RF, Gradient Boosting, K-NN, and SVR. The results showed that all models performed well, with RF producing the best R^2 of 0.9819 and generating histograms of residuals, for all models, to visualize the distribution of differences between predicted and actual values. The residual histograms for each model based on altimetry data were also presented. This study has important implications for monitoring SSH and preventing potential crises in coastal areas throughout the year.

Acknowledgment

We extend our sincere gratitude to EUMETSAT for their generous provision of access to altimetry data, a crucial contribution that significantly enhanced the success of our research endeavors. Our special appreciation is also directed towards the Google Earth Engine cloud-based platform.

References

[1] N. Makky and Z. Eghrari, "Forecasting atmospheric air pollution in Tehran using random forest model," International Conference of Environmental Remote Sensing and GIS, vol. 2024, pp. 57–60, 2024.

[2] WMO, "State of the global climate 2020," World Meteorological Organization, 2021.

[3] Z. Eghrari, M. Delavar, M. Zare, A. Beitollahi, and B. Nazari, "Land subsidence susceptibility mapping using machine learning algorithms," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. X-4/W1-2022, pp. 129–136, 2023.

[4] A. AghaKouchak, F. Chiang, L. S. Huning, C. A. Love, O. Mallakpour, H. Mazdiyasni, S. Moftakhari, M. Papalexiou, E. Ragno, and M. Sadegh, "Climate extremes and compound hazards in a warming world," Annual Review of Earth and Planetary Sciences, vol. 48, pp. 519–548, 2020.

[5] F. Yu, A. D. Switzer, Z. Zheng, B. Chen, J. Pile, H. Jol, Z. Huang, and A. Lau, "Holocene geomorphological evolution of a sediment-starved coastal embayment in response to sea level change," Palaeogeography, Palaeoclimatology, Palaeoecology, vol. 633, p. 111895, 2024.

[6] M. Mostafavi, N. Delpeche-Ellmann, and A. Ellmann, "Accurate sea surface heights from Sentinel-3A and Jason-3 trackers by incorporating high-resolution marine geoid and hydrodynamic models," Journal of Geodetic Science, 2020.

[7] O. B. Andersen, "Global ocean tides from ERS-1 and TOPEX/POSEIDON altimetry," Journal of Geophysical Research: Oceans, vol. 100, no. C12, pp. 25,249–25,271, 1995.

[8] Q. Gao, E. Makhoul, M. J. Escorihuela, M. Zribi, P. Quintana Seguí, P. García, and M. Roca, "Analysis of trackers performances and water level retrieval over the Ebro river basin using Sentinel-3," Remote Sensing, vol. 11, no. 6, p. 718, 2019.

[9] G. D. Quartly et al., "The roles of the S3MPC: Monitoring, validation, and evolution of Sentinel-3 altimetry observations," Remote Sensing, vol. 12, no. 11, p. 1763, 2020.

[10] M. J. Fernandes, C. Lázaro, and T. Vieira, "On the role of the troposphere in satellite altimetry," Remote Sensing of Environment, vol. 252, p. 112149, 2021.

[11] J. S. Mugunthan, "Evaluation of machine learning algorithms for the classification of lake ice and open water from Sentinel-3 SAR altimetry waveforms," UWSpace, 2023.

[13] F. Grimaccia, S. Leva, A. Niccolai, B. Ranjgar, and S. Trimarchi, "Social network optimization for electric vehicle charging stations deployment," 2023 IEEE International Conference on Artificial Intelligence & Green Energy, pp. 1–6, 2023.

[14] K. V. Kamran, N. Makky, and N. K. Charandabi, "Investigating the flooded area of Bangladesh by Sentinel-1 and CHIRPS images in the GEE system," Intercontinental Geoinformation Days, vol. 6, pp. 83–88, 2023.

[15] M. Srinivasan and V. Tsontos, "Satellite altimetry for ocean and coastal applications: A review," Remote Sensing, vol. 15, no. 16, p. 3939, 2023.

[16] Y. Cheng, X. Zhang, and Z. Yao, "On the performance of Sentinel-3 altimetry over high mountain and cascade reservoir basins," Remote Sensing, vol. 15, no. 7, p. 1769, 2023.

[17] N. Makky and Z. Eghrari, "Forecasting atmospheric air pollution in Tehran using random forest model," International Conference of Environmental Remote Sensing and GIS, vol. 2024, pp. 57–60, 2024.

[18] N. Makky and K. V. Kamran, "Declining sea surface height in the Caspian Sea based on Sentinel-3A data," International Conference of Environmental Remote Sensing and GIS, vol. 2024, pp. 125–127, 2024.

[19] H. Long and P. Goethals, "Machine learning applications in river research: Trends, opportunities, and challenges," Methods in Ecology and Evolution, vol. 13, no. 11, pp. 2603–2621, 2022.

[20] H. Chen et al., "Application of machine learning to evaluating and remediating models for energy and environmental engineering," Applied Energy, vol. 320, p. 119286, 2022.

[21] M. R. Claereboudt, "Oman," World Seas: An Environmental Evaluation, pp. 25–47, 2019.

[22] M. Hereher et al., "Assessment of the coastal vulnerability to sea level rise: Sultanate of Oman," Environmental Earth Sciences, vol. 79, p. 369, 2020.

[23] N. Makky and K. V. Kamran, "Declining sea surface height in the Caspian Sea based on Sentinel-3A data," International Conference of Environmental Remote Sensing and GIS, vol. 2024, pp. 125–127, 2024.

[24] R. M. Adnan et al., "Short-term probabilistic prediction of significant wave height using Bayesian model averaging," Ocean Engineering, vol. 272, p. 113887, 2023.

[25] N. Makky, K. V. Kamran, and S. Karimzadeh, "Impact of global warming on water height using XGBoost and MLP algorithms," Environmental Science Proceedings, vol. 29, p. 83, 2024.

[26] Q. Qian et al., "GIS partial discharge data enhancement method based on self-attention mechanism VAE-GAN," Global Energy Interconnection, vol. 6, no. 5, pp. 601–613, 2023.

[27] Z. Wu et al., "Predicting groundwater level based on machine learning," Water, vol. 15, no. 4, p. 823, 2023.

[28] D. Schoenfeld, "Partial residuals for the proportional hazards regression model," Biometrika, vol. 69, no. 1, pp. 239–241, 1982.

[29] X. Zhao et al., "Outlier detection based on residual histogram preference for geometric multi-model fitting," Sensors, vol. 20, no. 11, p. 3037, 2020.

[30] Z. Eghrari et al., "Groundwater level prediction using deep recurrent neural networks and uncertainty assessment," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. X-1/W1-2023, pp. 493–500, 2023.

[31] C. Tian et al., "Photovoltaic power prediction based on dilated causal convolutional network and stacked LSTM," Mathematical Biosciences and Engineering, vol. 21, no. 1, pp. 1167–1185, 2024.