14th International Congress on Civil Engineering, 21-23 October 2025
Sharif University of Technology, Tehran, Iran

Integrating Machine Learning and Ground-Based Data to Reconstruct ERA5 Precipitation Maps in the Urmia Lake Basin, Iran

Mohsen Moghaddas[1], Massoud Tajrishy[2]

1- M.Sc. Student, Department of Civil Engineering, Sharif University of Technology, Tehran, Iran.

2- Professor, Department of Civil Engineering, Sharif University of Technology, Tehran, Iran.

mohsen.moghadas1996@sharif.edu

Abstract

Accurate precipitation data is essential for hydrological modeling and water resources management, particularly in environmentally sensitive regions like Iran's Urmia Lake Basin. This study improves the accuracy of ERA5 precipitation estimates by integrating ground-based station data with advanced machine learning (ML) techniques. Five ML models (XGBRegressor, RandomForestRegressor, SupportVectorRegression, KNeighborsRegressor, and GradientBoostingRegressor) were assessed using daily precipitation data from 24 stations (March 2019–March 2022) and corresponding ERA5 data from Google Earth Engine. After preprocessing, RF and GBR outperformed others, reducing the Root Mean Square Error (RMSE) from 2.928 mm to 2.698 mm on average, with some stations achieving RMSEs as low as 2 mm. These models also significantly improved the detection of heavy precipitation events, which is critical for managing extreme weather impacts. The enhanced precipitation accuracy can improve hydrological modeling and inform more effective water management strategies in the Urmia Lake Basin. This research highlights the benefits of combining satellite and ground-based observations with ML techniques. Future studies could further refine these models by incorporating additional spatial features, such as elevation and land cover data, to capture regional topography. Additionally, integrating temporal dynamics and multi-source satellite data may yield even more precise precipitation estimates.

Keywords: Precipitation, ERA5, Urmia Lake Basin, Ground-Based Data, Machine Learning.

1- INTRODUCTION

Precipitation is one of the key parameters in hydrological modeling and water resources management. Its importance is amplified in regions experiencing water scarcity and climatic stress. This is particularly true for Iran's Urmia Lake Basin, where a marked lack of in-situ precipitation data has become a critical barrier to effective hydrological assessment and is directly impacting the formulation and implementation of lake restoration policies [1].

However, obtaining precise precipitation measurements remains a significant challenge. Precipitation exhibits high variability across spatial and temporal scales [2]. While ground-based stations provide accurate point measurements, they are often sparsely or unevenly distributed, especially across vast and topographically complex regions. Conversely, satellite and reanalysis datasets—such as ERA5—offer broad spatial coverage but may suffer from systematic biases and reduced accuracy, particularly during heavy precipitation events or in mountainous and semi-arid areas.

The Urmia Lake Basin exemplifies these challenges. As one of the most environmentally sensitive and water-stressed regions in Iran, it has experienced a dramatic decline in water levels due to prolonged droughts and poor water management [3]. Accurate precipitation data is essential for restoring and sustaining this basin, making it an ideal case for improving existing data sources.

Recent advances in machine learning (ML) have opened new pathways to enhance the accuracy of reanalysis precipitation data. ML techniques can learn complex nonlinear relationships between ground-based observations and satellite-derived estimates, potentially reducing bias and improving event detection. Several studies have demonstrated the success of ML models in correcting satellite precipitation estimates in diverse environments. For instance, a random forest algorithm was used to correct ERA5 precipitation across high-mountain basins in the Tibetan Plateau, reducing overestimation by up to 50% and significantly improving hydrological simulation accuracy [4]. Another study applied a random forest model to adjust ERA5-derived rainfall erosivity on the Chinese Loess Plateau, yielding better spatial estimates and lower bias and RMSE values on annual scales [5]. More recent approaches have taken advantage of probabilistic models such as multi-fidelity Gaussian processes (MFGPs) to downscale and bias-correct ERA5 data in mountainous regions, enabling uncertainty quantification and applicability to extreme events [6]. Additionally, deep generative