International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

A Review of Machine Learning Approaches for Rainfall Forecasting Trends, Datasets, and Challenges

Atharv Taksali

Sancta Maria International School

Abstract: ***Predicting rainfall is a must-have capability in climate science, and it has implications in other interconnected fields like agriculture, hydrology, energy, and disaster management. Traditionally, physical or statistical models were used; however, these models have faced limitations. Among these limitations are the high computational requirements, the inflexibility of assumptions, and the difficulty of capturing nonlinear atmospheric processes. Machine Learning (ML) and Deep Learning (DL) have become powerful alternatives. They are not only able to learn directly from the data, but also can model complex patterns in time and space, and provide more reliable forecasts. This survey collects the cutting-edge research works that have been done in ML and DL methods for rainfall prediction. It covers models ranging from the classical methods like Support Vector Regression and Random Forest to the highly complex architectures such as Long Short-Term Memory Networks (LSTM) and Convolutional Neural Networks. Besides that, it stresses the importance of the datasets that are obtained from satellite missions, re-analysis products, and community-driven platforms for better performance of the models. The review pinpointed existing issues, including data quality problems, missing values, limited coverage in geographical areas, high computational costs, and limited interpretability. Moreover, it presented ideas about the potential future paths of these issues. These paths involve hybrid and physics-informed models, improved benchmarking, real-time data integration, and interpretable AI. By describing the current state of research and the gaps therein, this paper sets out the essentials that constitute a viable framework for making advances in rainfall forecasting not only with higher precision, but also greater scalability and operational trust.***

Keywords: Climate resilience, Deep Learning, Machine Learning, Rainfall forecasting, Time-series prediction

1. Introduction

Rainfall forecasting remains instrumental for earth sciences and is equally significant in many practical sectors as it aligns with the fact that precipitation essentially affects the environment, the economy, and the society (Gupta et al. 2025; Latif et al. 2023). The accurate predictions streamline the agricultural calendar, enhance agricultural output, and avert agricultural losses (Latif et al. 2023). Moreover, they are very instrumental in water management and preventive measures against natural calamities, in which case they minimize the dangers posed by flooding, droughts, and landslides. Besides, the sectors like transport, energy, real estate, and tourism that rely on the safety of their operations can similarly benefit from the lessening of economic losses resulting from timely forecasts (Gupta et al. 2025; Latif et al. 2023).

Weather forecasting is a challenging task because weather is an intricate and constantly changing system which, in turn, makes it difficult for traditional statistical and numerical methodologies to effectively provide solutions in the field (Gupta et al. 2025). Machine learning (ML) tackles these problems by extracting knowledge from extensive data, mastering intricate patterns, and refining prediction. Deep learning architectures such as LSTM networks are highly effective, in fact, they can handle different climatic zones and timescales (Latif et al. 2023). As a consequence, it is possible to have early warning systems and crucial decision-making supported with timely and reliable forecasts (Gupta et al. 2025).

The present work brings together a range of research done on ML methods for rainfall forecasting, which concentrates on the model applications, datasets, and performance parameters. It identifies the issues such as limited data, changing trends, and computing restrictions, and by discussing these issues, it drills down into the viability of different sectors like agriculture, water management, disaster preparedness, and energy planning. After identifying trends and limitations, it provides an outline of improvement in ML-based rainfall prediction and points research to more accurate and practical forecasting.

2. Background

Alongside human history, physical and statistical models were employed to predict rainfall and weather. Physical methods are those that use highly detailed mathematical formulas to simulate the atmosphere, for example, General Circulation Models (GCMs) and Numerical Weather Prediction (NWP). The main problems with these types of models are that they require a lot of calculations and that they do not always have the capacity to detect the minute atmospheric phenomena such as turbulence and cloud microphysics. As a result, they often cause structural uncertainties. (Hussain et al. 2021; 2025). On the other hand, statistical models like regressions and ARIMA, that rely on linear assumptions and specific distributions of data, have lower performance when depicting complicated and nonlinear patterns and are more vulnerable to the presence of outliers (Hussain et al. 2021; 2025). The mentioned disadvantages imply the significance of developing more flexible data driven approaches. For rainfall forecasting, machine learning (ML) and deep learning (DL) have become strong, flexible substitutes. ML models are capable of learning complex nonlinear relationships from data.

Machine learning (ML) and deep learning (DL) have become efficient and flexible alternatives for precipitation prediction. However, DL with its multi-layered neural networks can automatically derive hierarchical features from raw data, hence, there is no need for manual feature engineering whereas ML models can learn complex nonlinear relationships from data (Patil et al., 2024). Coupled with the progressively available high-resolution datasets and the improved computational power, these methods panel significantly enhance the accuracy, the speed, and the scale of predictions. Therefore, accurate yet speedy forecasts are realized through this which in turn facilitates early warning systems and decision-making processes.

Studies and tests were performed on various machine learning models in order to predict rainfall. Traditional methods such as Random Forest (RF) and Support Vector Regression (SVR) which are able to handle nonlinear interactions have been reported to work well even in situations where data is insufficient. At the same time, Deep Learning frameworks are perfect for the handling of sequential and grid-based meteorological data. These are Long Short-Term Memory (LSTM) networks for learning temporal dependencies, Convolutional Neural Networks (CNNs) for representing spatial features, and Artificial Neural Networks (ANNs). Through these models, both long-term and localized precipitation predictions have been significantly improved (Obaidalla 2024; Patil et al. 2024).

3. Discussion

3.1 Comparative Analysis of ML and DL Methods

When considering time-series forecasting, machine learning (ML) vs. deep learning (DL) comparisons reveal that increasingly complex and hybrid architectures are the trend. Though they deliver good predictive performance, conventional machine learning models such as Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Networks (ANN), and k-Nearest Neighbor (KNN) are usually unable to effectively handle the temporal and nonlinear nature of rainfall dynamics (Wani et al., 2024). On the other hand, DL models like Long Short-Term Memory (LSTM), Bi-directional LSTM, Gated Recurrent Unit (GRU), and Recurrent Neural Networks (RNN) are more capable of capturing temporal dependencies and extracting hierarchical features (Wani et al. 2024; Dtissibe et al. 2024).

The feature-learning capability of deep learning along with the decision-making power of machine learning is what new hybrid methods essentially represent. As an example, the use of Particle Swarm Optimization on Multi-Layer Perceptrons (PSO-MLP) for rainfall prediction has led to a situation where results are quite promising (Hatem Abdul-Kader et al. 2018). Accurate datasets from satellite missions like TRMM and GPM, meteorological databases like IMD records for local forecasting, detailed reanalysis, and open repositories like ERA5 and Kaggle are equally important for these models.

3.2 Datasets and Data Challenges

Essentially, ML/DL forecasting demands high-quality datasets with substantial spatial and temporal resolutions. TRMM was the pioneering mission to monitor precipitation from space, and its successor GPM brought about IMERG. IMERG covers almost the entire globe from 60°N to 60°S, boasts a spatial resolution as small as 0.1°, and records updates every 30 minutes, which is a great advantage for urban forecasting (NASA GPM: IMERG). Furthermore, ERA5 operates similarly but on a different scale, providing global, hourly re-analysis at a 0.25° resolution and thus being quite instrumental in understanding long-term atmospheric patterns (Hersbach et al. 2020). Besides, Kaggle and similar platforms are great resources for model development and benchmarking. Nevertheless, the problem of satellite coverage being non-uniform persists. With TRMM, mid- and high-latitudes were not covered while IMERG experiences challenges in polar areas. Hence, there is always a need to depend on reanalysis data. The issues of missing values, retrieval biases, and limited access to ground-based observations, thus, are factors that further hamper the accuracy of the data (Patil et al. 2024).

Indeed, the pre-processing stage is a decisive factor in the outcome of the whole process. It involves data cleaning and normalizing, imputing missing values, correcting biases against ground truth, and smoothing or decomposing time series (Gupta et al. 2023). By doing so, model reliability is increased and their practical application in rainfall prediction at urban and regional scales becomes feasible.

3.3 Key Challenges

Although ML and DL-based rainfall forecasting have made considerable progress, they still encounter various operational and technical challenges. For instance, overfitting is still a significant problem in data-sparse areas with a complex terrain, like the North-Western Himalayas. A limited number of observations in this area impede the generalization of models (Wani et al. 2024; Patil et al. 2024). Models that focus on noise rather than essential spatiotemporal patterns lead to a lower level of predictive reliability, particularly when small or unbalanced datasets are used. Furthermore, the issue of scalability is still present as the process of training complex structures such as CNNs and LSTMs requires a lot of energy, need for specialized hardware, and high processing power; these requirements can limit the availability of such techniques in developing countries. Besides, the "black-box" characteristic of deep learning restricts the interpretability of the model. This factor negatively affects the building of user trust and makes the validation process more complex in essential sectors like flood forecasting and disaster management (Patil et al. 2024).

3.4 Future Perspectives

Further developments will mainly emphasize on hybrid models and Physics-Informed Machine Learning (PIML) that merges the physical laws in the learning to make the predictions not only accurate but also physically consistent (Patil et al. 2024; Rao et al. 2023). Enhancing data infrastructure is equally important. The ultimate aim is to produce more detailed, open-access datasets and standardize the benchmarks for particular forecasting tasks. Pipelines for real-time data integration that can merge satellite, ground, and NWP inputs will enable almost instant forecasts which are

imperative for flash flood warnings (Rao et al. 2023). Moreover, Interpretable AI (IAI) will be a key factor in the implementation. It will reveal the model decisions, thereby, empowering the user community, and facilitating the identification of bias. Improved interpretability is essential for turning research models into reliable operational tools (Patil et al. 2024).

4. Conclusion

The conversation of ML and DL application in agriculture, hydrology, and urban rainfall forecasting illustrates both the substantial advances and the remaining gaps. One key takeaway from the success of the models is the role of data. Deep Learning models, particularly LSTMs and their variants, have been found to be more effective in understanding the complex, non-linear patterns in rainfall time-series. These models can provide better predictive accuracy than conventional ML techniques (Hatem Abdul-Kader et al. 2018). On the other hand, problems related to data such as regional gaps due to satellite missions like GPM IMERG (NASA GPM: IMERG) and biases in reanalysis products (Hersbach et al. 2020) limit their generalization power. Furthermore, the high computational costs and the black-box nature of DL models make their practical application less attractive and also affect the trust of the users (Patil et al. 2024).

Addressing these issues requires us to work on three aspects simultaneously:

- Data quality: stricter pre-processing, imputation, and bias correction to increase the trustworthiness of the data (Patil et al. 2024).
- Efficiency: the development of hybrid and lightweight models to reduce the computational costs and increase the scalability of the models (Patil et al. 2024).
- Interpretability: building Interpretable AI (IAI) for improved openness and physical validation.

Developments in the future will depend on Physics-Informed ML (PIML), better benchmark datasets, and infrastructures for the real-time integration of data (Rao et al. 2023). Together, these enhancements can make forecasts more accurate and reliable, thus contributing to climate resilience and disaster preparedness.

References

[1] Latif, Hazrin, Koo, Ng, Chaplot, Huang, El-Shafie, Ahemd. “Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches.” Alexandria Engineering Journal, Volume 82, 1 November 2023.
[2] Patil, Rane, Desai, Rane. “Machine learning and deep learning: Methods, techniques, applications, challenges, and future research opportunities.” Deep Science Publishing, October 2024.
[3] Abd elkader, Salam, Mohamed. “Hybrid Machine Learning Model for Rainfall Forecasting.” Journal of Intelligent Systems and Internet of Things, Volume 1, January 2020.
[4] Wani, Mahdi, Yeasin, Kumar, Gagnon, Danish, Al-Ansari, Hendawy, Mattar. “Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas.” Scientific Reports, 13 November 2024.
[5] NASA. “Tropical Rainfall Measuring Mission.” 2011
[6] NASA. “Global Precipitation Measurement.” IMERG
[7] Hersbach et al. “The ERA5 global reanalysis.” Quarterly Journal of the Royal Meteorological Society, 15 June 2020
[8] Gupta, Jain, Pandey, Gupta, Saha. “Evaluation of global precipitation products for meteorological drought assessment with respect to IMD station datasets over India.” Atmospheric Research, Volume 297, January 2024
[9] Croprese, Pierantozzi, Lops, Montelpare. “DL^{2}F: A Deep Learning model for the Local Forecasting of renewable sources.” Computers and Industrial Engineering, Volume 187, January 2024
[10] Khairudin, Mustapha, Aris, Zolkepli. “Comparison of Machine Learning Models For Rainfall Forecasting.” 2020 International Conference on Computer Science and Its Application in Agriculture
[11] Carleton, Lee. “Modeling lake recovery lag times following influent phosphorus loading reduction.” Environmental Modelling & Software, Volume 162, April 2023
[12] Salcedo-Sanz et al. “Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources.” Information Fusion, Volume 63, November 2020.