Progress Report: Machine Learning–Based Nowcasting of Extreme Rainfall over Aceh Using ERA5 Reanalysis

Muhammad Raza Adzani
Department of Informatics
Syiah Kuala University
Banda Aceh, Indonesia
email: raza.a22@mhs.usk.ac.id

Ahmad Siddiq
Department of Informatics
Syiah Kuala University
Banda Aceh, Indonesia
email: m.siddiq2@mhs.usk.ac.id

***Abstract*—Short-term prediction of extreme rainfall is essential for flood early warning in regions such as Aceh, Indonesia, where intense convective storms frequently cause flash floods. This progress report presents the current status of a machine learning–based framework for three-hour-ahead extreme rainfall nowcasting using ERA5 single-level reanalysis data. The problem is formulated as a grid-based binary classification task on a $0.25^\circ \times 0.25^\circ$ domain. At this stage, data preprocessing, feature engineering, and baseline model implementation have been completed. Preliminary experiments using logistic regression and random forest classifiers indicate that non-linear ensemble models provide improved discrimination of rare extreme rainfall events compared with linear baselines. Ongoing work focuses on refining threshold selection, extending evaluation, and developing visualization tools to support early-warning applications.**

***Index Terms*—extreme rainfall, nowcasting, machine learning, random forest, ERA5 reanalysis, progress report**

I. Introduction

Extreme rainfall remains one of the main drivers of floods and landslides in Indonesia, with Aceh Province being particularly vulnerable due to intense monsoonal precipitation and limited local monitoring infrastructure. Short-lived but high-intensity rainfall events often escalate rapidly into disasters, highlighting the need for timely early-warning information.

Existing early-warning systems rely primarily on numerical weather prediction models and sparse observational networks, which may show limited skill at very short lead times and fine spatial scales. As discussed in the proposal stage of this project, reanalysis datasets such as ERA5 offer an opportunity to explore data-driven approaches that learn empirical relationships between atmospheric conditions and subsequent rainfall extremes. This progress report summarizes the work completed to date toward developing a machine learning–based nowcasting framework for extreme rainfall over Aceh.

II. Related Work

Machine learning approaches for precipitation forecasting and extreme rainfall prediction have been widely explored in recent years. Random forest classifiers have demonstrated improved performance over logistic regression for predicting exceedance of heavy precipitation thresholds and for distinguishing extreme from non-extreme events [1], [8]. Other studies have applied machine learning to correct biases in reanalysis precipitation, leading to improved hydrological simulations [2].

Beyond tree-based methods, deep learning architectures such as convolutional and recurrent neural networks have been used for precipitation nowcasting with radar and satellite data [5]–[7], [9]. While these approaches achieve high skill, they typically require dense observational coverage. For data-scarce regions such as Aceh, relatively simple but robust models driven by reanalysis variables remain an attractive alternative [3], [4], [10]. This project builds on these ideas by focusing on short-lead extreme rainfall classification using ERA5 data.

III. Methodology

A. Data and Study Area

ERA5 single-level reanalysis data from 2020 to 2024 are used over a small domain surrounding Banda Aceh (5.0°–6.0°N, 95.0°–96.0°E). The spatial resolution of 0.25° yields a 5×5 grid covering coastal and inland areas. Hourly fields are aggregated to three-hourly intervals, resulting in a dataset of 365,400 grid-time samples.

B. Problem Formulation and Features

The task is formulated as a binary classification problem, where the goal is to predict whether three-hour-ahead accumulated precipitation exceeds the empirical 95th percentile. Features implemented at this stage include near-surface meteorological variables, lagged precipitation values, rolling means, and cyclical encodings of diurnal and seasonal time scales.

C. Models and Data Splitting

Two models have been implemented: logistic regression as a baseline and random forest as a non-linear ensemble model. Data are split chronologically by year to emulate an operational setting, with earlier years used for training and later years reserved for validation and testing.

IV. Preliminary Experiments

Initial experiments have been conducted to assess the feasibility of the proposed approach. Using default probability thresholds, the logistic regression model achieves high overall accuracy but shows limited recall for rare extreme rainfall events. In contrast, the random forest classifier demonstrates substantially higher recall and better discrimination of extreme events, as reflected by higher values of the area under the ROC curve on the validation data.

Given the strong class imbalance, preliminary threshold tuning has been performed on the validation set to explore the trade-off between recall and precision for extreme events. These initial results suggest that operating the random forest at a lower probability threshold improves detection of extreme rainfall at the expense of additional false alarms, which may be acceptable in an early-warning context.

V. Next Steps

Several tasks remain to be completed in the final phase of the project. First, further evaluation will be performed on an independent test year to assess generalization performance. Second, additional analysis of threshold selection and error characteristics will be conducted to better understand model behavior. Third, visualization tools, including grid-based risk maps and a simple web interface, will be developed to support interpretation and communication of results. Finally, potential extensions such as alternative machine learning models and incorporation of additional predictors will be explored.

References

[1] G. R. Herman and R. S. Schumacher, "Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests," *Monthly Weather Review*, vol. 146, no. 5, pp. 1571–1600, 2018.

[2] H. Sun *et al.*, "Corrected ERA5 precipitation by machine learning significantly improved flow simulations," *Journal of Hydrometeorology*, vol. 23, no. 10, pp. 1663–1679, 2022.

[3] B. Yang *et al.*, "A method for monthly extreme precipitation forecasting," *Water*, vol. 15, no. 8, 2023.

[4] S. Chkeir *et al.*, "Nowcasting extreme rain with machine learning techniques," *Atmospheric Research*, vol. 282, 2023.

[5] S. Ravuri *et al.*, "Skillful precipitation nowcasting using deep generative models," *Nature*, vol. 597, pp. 672–677, 2021.

[6] G. Ayzel *et al.*, "RainNet: A CNN for radar-based precipitation nowcasting," *Geoscientific Model Development*, vol. 13, pp. 2631–2644, 2020.

[7] F. Gamboa-Villafruela, "Immediate precipitation forecasting using deep learning," *Proceedings*, vol. 8, no. 1, 2021.

[8] D. Wolfensberger and A. Feinberg, "RainForest: A random forest algorithm for precipitation estimation," *Atmospheric Measurement Techniques*, vol. 14, pp. 3169–3193, 2021.

[9] X. Shi *et al.*, "Convolutional LSTM network for precipitation nowcasting," *NeurIPS*, 2015.

[10] X. Shi, "Smart dynamical downscaling of extreme precipitation," *Geophysical Research Letters*, vol. 47, 2020.