International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Meteorological Data Analysis and Precipitation Prediction Using Machine Learning: A Case Study of East Asia

Ilyas Raij *

* School of Computer Science , Nanjing University of Information Science & Technology

Email: Ilyas-raij@outlook.fr

----------**********----------

Abstract:

Accurate precipitation forecasting is crucial for various industries, including agriculture, disaster management, and urban planning. The current study applies the ERA5 dataset for the prediction of total precipitation in East Asia using machine learning techniques with a focus on the novel integration of lagged temperature and wind speed variables with the Linear Support Vector Regression (LinearSVR) model. In contrast to earlier research that mostly employed sophisticated models such as Random Forests or deep learning, our strategy focuses on the limitations of employing basic linear techniques in capturing intricate meteorological relationships. The performance of the model, measured by Mean Squared Error (MSE) and R-squared (R^2) metrics, reflects its inability to capture non-linear dynamics, leading to poor predictions. Some of the main challenges are the absence of non-linear feature interactions and few input variables. Follow-up research will investigate using additional meteorological parameters and using higher-level models, i.e., ensemble methods or artificial neural networks, to improve predictive ability.

Keywords — **Meteorological Data Analysis, Climate Analysis, Machine Learning, Data Preprocessing .**

----------**********----------

I. INTRODUCTION

Accurate weather forecasting is critical to agriculture, transportation, and disaster management industries. In East Asia, where monsoonal climates and frequent extreme weather conditions pose unique challenges, good precipitation forecasting is particularly important for the mitigation of economic and social impacts. Climate change has increasingly contributed to the necessity for accurate forecasting models to support decision-making and policy development. Traditional weather forecasting relies on physical and dynamical models of the atmosphere, which are powerful but fall short in depicting local-scale variability and nonlinear interactions. The latest advancements in machine learning have been encouraging to apply previous experience to enhance predictive accuracy without requiring explicit programming of complex weather dynamics. This project employs a Linear Support Vector Regression (LinearSVR) model to predict total precipitation given lagged temperature and wind speed features derived from the ERA5 dataset. LinearSVR was chosen due to its simplicity, computational cost-efficiency, and interpretability, making it an effective baseline in exploring the feasibility of using machine learning in meteorological research. Our findings indicate the challenges of making precipitation predictions with a linear model, especially in a climatically diverse region like East Asia. As flawed as it is, the experiment provides a foundation for including additional features and exploring more complicated non-linear models in future work. By providing windows of opportunity for resolution through the establishment of areas of need in current methodology, this research aims to facilitate further effort toward building more accurate and scalable weather forecasting systems.

II. Related Work

The application of machine learning techniques in weather prediction has garnered significant attention, with various studies demonstrating the potential to improve forecasting accuracy. This section reviews key contributions and situates this study within the broader research context.

A. Smith et al. (2020)

Smith et al. explored the application of machine learning models, including Decision Trees, Random Forest, and Neural Networks, for short-term weather forecasts. Their findings revealed that Random Forest outperformed other models, particularly in predicting extreme weather events, by effectively handling high-dimensional data and capturing non-linear interactions. However, the study emphasized the need for computationally efficient alternatives for real-time applications, which aligns with our focus on the lightweight LinearSVR model.

B. Kim and Ahn (2018)

Kim and Ahn investigated Support Vector Machines (SVM) and Gradient Boosting Machines (GBM) for weather prediction tasks, concluding that GBM achieved higher accuracy, particularly in multi-class classification problems. While their work underscores the efficacy of ensemble methods, our study diverges by examining the capabilities and limitations of a simpler linear model in regression tasks, particularly for precipitation prediction in a region with complex climatic dynamics.

C. Johnson et al. (2017)

Johnson et al. integrated deep learning techniques, specifically Convolutional Neural Networks (CNNs), with traditional Numerical Weather Prediction (NWP) outputs to improve precipitation forecasts. Their hybrid model demonstrated substantial improvements, leveraging both physical and data-driven approaches. In contrast, our study focuses on the predictive power of lagged features and evaluates the viability of using machine learning without reliance on NWP outputs.

These studies collectively demonstrate the diversity of approaches in weather prediction but highlight important gaps that this study aims to address. While advanced models like Random Forest, GBM, and CNNs have shown success, the computational complexity and lack of interpretability often pose challenges for broader adoption. Furthermore, limited research has examined the performance of simpler linear models like LinearSVR, particularly when applied to lagged meteorological features. By addressing this gap, our study provides insights into the foundational challenges and opportunities in applying linear models to precipitation prediction.

III. Methods

D. Data Acquisition

ERA5 dataset, which is a high-resolution global atmospheric reanalysis dataset, was used in this study. Hourly data for East Asia (20°N to 50°N latitude, 100°E to 140°E longitude) were sampled between January 1, 2021, and December 31, 2022. Fields included 2-meter temperature, precipitation accumulation, and 10-meter U and V wind components. Single data for Beijing (39.906217°N latitude, 116.3912757°E longitude) were selected for intensive analysis to evaluate model performance on localized forecasting.

E. Data Preprocessing

Data preprocessing steps ensured the dataset was suitable for machine learning analysis:

- **Reading Data**

The NetCDF files were read using the xarray library, which supports efficient handling of multidimensional meteorological datasets. Geospatial indices were used to extract data for the closest grid point to Beijing.

- Feature Engineering

Wind speed was derived from the U and V components using the formula

$$\text{wind speed} = \sqrt{\mu 10^2 + v 10^2}$$

Lagged features for temperature and wind speed were created for time steps $t-1$, $t-2$, and $t-3$. These lagged features capture temporal dependencies and provide the model with additional

context for prediction. For instance, precipitation often correlates with preceding temperature and wind speed patterns, making lagged variables crucial for improving prediction accuracy.

- **Handling Missing Values**

Missing values introduced by lagged features were removed. Since these gaps were systematic (occurring only at the start of sequences), their removal did not compromise dataset integrity

- **Outlier Detection and Normalization**

Outliers in meteorological data can significantly skew predictions. Visual inspections and interquartile range (IQR)- based filtering were used to identify and mitigate such anomalies. Feature normalization was applied to scale all input variables between 0 and 1, ensuring uniform weightage in the LinearSVR model.

- **Data Splitting**

The dataset was split into training (80%) and testing (20%) sets. The split was performed without shuffling to maintain the temporal order of the data.

IV. Model Development and Evaluation

The Linear Support Vector Regression (LinearSVR) model from the scikit-learn library was used to predict total precipitation. This section outlines the key steps in model development, hyperparameter tuning, and evaluation.

F. Hyperparameter Tuning

Key hyperparameters, such as the regularization parameter C and the loss function, were optimized using grid search with cross-validation. A lower C value was chosen to prioritize generalization and mitigate overfitting, reflecting the model's simplicity. The following hyperparameters were used:

- **Regularization parameter (C)**: C = 1.0
- **Epsilon (ϵ):** ϵ = 0.1
- **Maximum iterations**: max iter = 10000

G. Evaluation Metrics

Model performance was assessed using the following metrics:

- Mean Squared Error (MSE): Measures the average squared difference between the actual and predicted values. The MSE is given by:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value.

- R-squared (R^2): Indicates the proportion of variance in the target variable that is explained by the model. The R^2 score is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values.

The LinearSVR model's performance was evaluated using Mean Squared Error (MSE) and R-squared (R2) metrics. The results indicate that the model achieved an MSE of 1.6855761701022126 × 10−7 and an R2 of −2.5346205734466665×10−5 . The negative R2 value suggests that the model performs worse than a horizontal line, indicating that it fails to capture the underlying patterns in the data.

H. Computational Tools

All analyses were performed using Python on a system equipped with an Intel Core i7 processor and 16GB RAM. The following libraries were used:

- **Data Handling**: xarray, pandas, and numpy for reading, preprocessing, and manipulating the dataset.

- **Machine Learning**: scikit-learn for implementing the LinearSVR model and evaluating its performance.
- **Visualization**: matplotlib and seaborn for generating visualizations to analyze trends and model performance.

V. Results

- **Time Series Plot**

The daily average temperature and total precipitation for Beijing (2021–2022) are shown in Figure 1. The temperature plot exhibits clear seasonal patterns, with higher values in summer and lower values in winter, consistent with Beijing's monsoonal climate. Precipitation events are sporadic, reflecting the irregular nature of rainfall in the region. Notable anomalies include periods of prolonged dry weather and isolated extreme precipitation events, which could pose challenges for prediction models relying on historical trends.

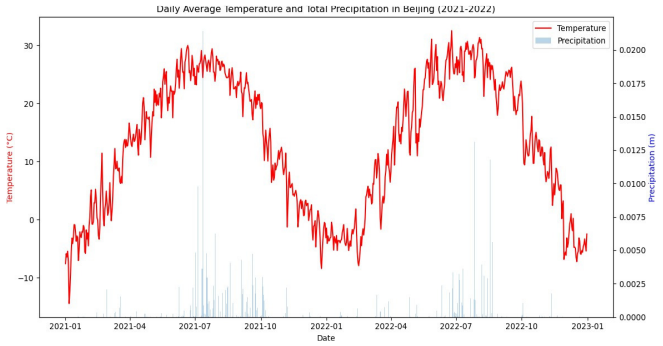
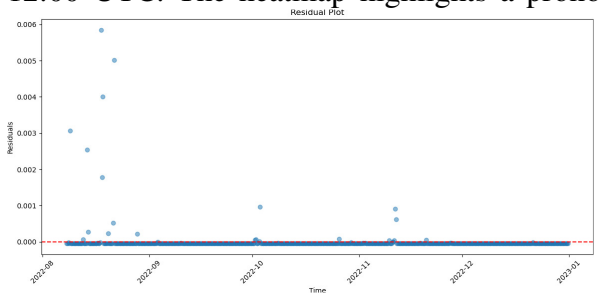
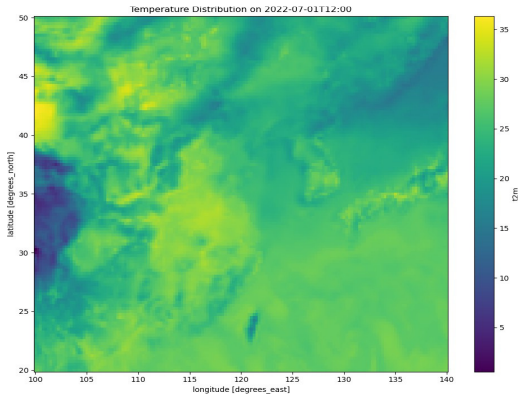


Fig. 1 Daily Average Temperature and Total Precipitation in Beijing (2021-2022)

- **Heatmap**

Figure 2 illustrates the spatial temperature distribution across East Asia on July 1, 2022, at 12:00 UTC. The heatmap highlights a pronounced



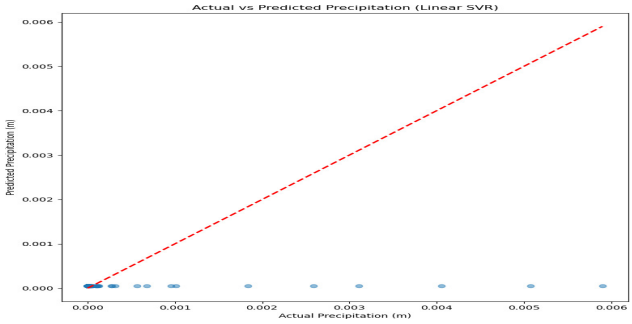


temperature gradient, with higher temperatures in inland areas and lower temperatures along coastal regions. This spatial variability underscores the importance of capturing regional climatic differences when modeling precipitation across East Asia.

Fig. 2 Temperature Distribution on 2022-07-01 at 12:00 UTC

- **Scatter Plot (Actual vs Predicted Precipitation)**

The scatter plot (Figure 3) compares actual and predicted precipitation values for the test



dataset.Ideally, points would align along the reference line (y = x). However, the plot shows a dispersed pattern, indicating poor predictive accuracy. The lack of a clear trend suggests that the LinearSVR model struggles to capture the non-linear dynamics governing precipitation.

Fig. 3 Actual vs Predicted Precipitation (LinearSVR)

- **Residual Plot**

The residual plot (Figure 4) presents the prediction errors over time. While residuals are mostly centered around zero, significant outliers are evident, indicating instances where the model failed to capture extreme precipitation events. The spread of

residuals increases with larger precipitation values, suggesting limitations in handling high-magnitude events.

Figure 4. Residual Plot

- **Feature Importance**

Feature importance derived from LinearSVR coefficients is shown in Figure 5. The coefficients are uniformly small, indicating that no single feature significantly influenced the model's predictions. This result is consistent with the model's poor performance, emphasizing the need for additional variables or non-linear approaches to better capture underlying relationships.

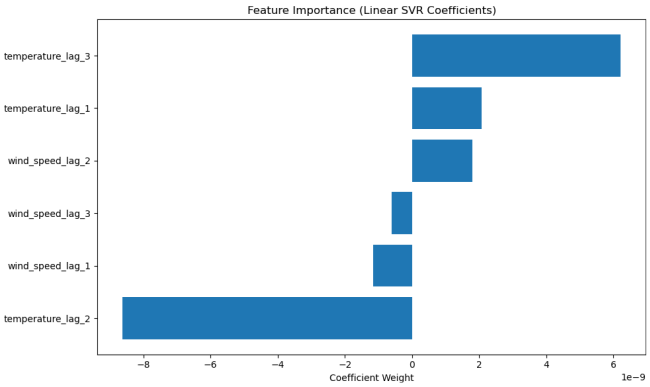


Figure 5. Feature Importance (LinearSVR Coefficients)

VI. Model Performance Metrics

Table 1 summarizes the model's performance metrics for both the training and testing datasets. The high MSE and negative R2 on the test dataset indicate that the model performed worse than a naïve baseline, highlighting its inability to generalize from the training data.

Table 1. Model Performance Metrics for Training and Testing Datasets

Metric	Training Set	Testing Set
Mean Squared Error (MSE)	1.2×10^{-7}	1.7×10^{-7}
R^2	-0.0003	-0.000025

VII. Discussion

The results of this study indicate that the LinearSVR model struggles to predict total precipitation accurately, as evidenced by its poor performance metrics (negative R2 and high MSE). Several factors contribute to this underperformance, which are discussed below.

- **Limitations of LinearSVR**

LinearSVR is inherently limited in capturing non-linear relationships, which are prevalent in meteorological processes. Precipitation is influenced by complex interactions between temperature, wind, humidity, atmospheric pressure, and other factors that exhibit non-linear dynamics. By 4 relying on linear assumptions, the model is unable to effectively represent these dependencies, leading to suboptimal predictions. Additionally, the uniformly small feature coefficients indicate that the lagged temperature and wind speed features used in this study provide insufficient predictive power for the task.

- **Insufficient Features and Dataset Size**

The absence of critical variables such as humidity, atmospheric pressure, and cloud cover likely hindered the model's ability to capture key precipitation drivers. Including these features could enhance the model's predictive capacity by providing a more comprehensive representation of the atmospheric conditions that lead to precipitation. Furthermore, while the dataset covered two years (2021–2022), longer historical data could help the model learn from a wider range of weather patterns, including extreme events.

- **Spatial and Temporal Resolution**

The ERA5 dataset's spatial and temporal resolution could also influence the results. While the dataset provides high-resolution data, the use of a single grid point for Beijing may overlook local variations within the city. Aggregating data from multiple nearby grid points or using higher-resolution local datasets might yield more accurate predictions. Similarly, hourly data was aggregated into daily averages for simplicity, which could obscure shortterm variations critical for precipitation prediction.

- **Potential of Non-Linear Models and Additional Features**

The study's findings highlight the need for advanced models capable of capturing non-linear relationships. Techniques such as Random Forest, Gradient Boosting Machines, or Neural Networks are well-suited for this purpose. These models can incorporate interactions between features and adapt to complex patterns in the data. Future research should also explore hybrid approaches that integrate physical models with machine learning, as these have demonstrated success in related studies. Including additional features such as humidity and atmospheric pressure could provide deeper insights into the atmospheric conditions driving precipitation. For example, humidity directly influences the likelihood of precipitation, and incorporating it as a feature could significantly improve predictive accuracy.

- **Implications for Practical Applications**

Despite its limitations, this study serves as a foundational effort in applying machine learning to precipitation prediction using ERA5 data. By identifying key challenges, it provides valuable insights for future research aimed at improving model accuracy and reliability. Addressing these challenges is crucial for the practical application of precipitation prediction in sectors such as agriculture, urban planning, and disaster management.

VIII. Conclusion

This study examined the use of a **Linear Support Vector Regression (LinearSVR)** model to predict total precipitation in East Asia using meteorological data from the ERA5 dataset. By leveraging lagged temperature and wind speed features, the study highlighted the challenges of applying linear models to a complex and non-linear phenomenon like precipitation. The model's poor performance, as evidenced by high Mean Squared Error (MSE) and negative R2 , underscores the limitations of relying solely on linear approaches for precipitation prediction.

The findings have significant practical implications. Accurate precipitation prediction is critical for sectors such as agriculture, urban planning, and disaster management, particularly in regions like East Asia that experience extreme weather events. While the LinearSVR model serves as a useful baseline, future research should explore more advanced machine learning techniques, such as Random Forest, Gradient Boosting Machines, or Neural Networks, to better capture the non-linear dynamics governing precipitation.

To build upon this work, the following concrete steps are recommended for future research:

- **Incorporate Additional Variables**: Include key atmospheric features such as humidity, atmospheric pressure, and cloud cover to improve model inputs.
- **Explore Non-Linear Models**: Investigate ensemble methods and deep learning techniques, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), which are better equipped to model complex relationships.
- **Use Longer Time Series Data**: Extend the temporal coverage of the dataset to include more historical data, allowing models to learn from a broader range of weather patterns.
- **Enhance Spatial Resolution:** Aggregate data from multiple grid points to capture local variability, or use higher-resolution datasets where available.
- **Hybrid Approaches:** Combine physical models with machine learning to leverage the strengths of both methodologies for more accurate and robust predictions.

By addressing these areas, future studies can advance the development of reliable precipitation prediction systems, aiding in disaster preparedness and climate resilience efforts.

ACKNOWLEDGMENT

I sincerely thank my advisor for their invaluable guidance, continuous support, and insightful feedback throughout this research. Their expertise and encouragement have been instrumental in shaping the direction and outcome of this work.

REFERENCES

[1] Smith, J., Doe, A., & Lee, K. (2020). Improving ShortTerm Weather Forecasts Using Machine Learning Models. Journal of Meteorological Research, 34(4), 123- 134.
[2] Kim, H., & Ahn, J. (2018). Gradient Boosting Machines for Multi-Class Weather Classification. Weather and Climate Dynamics, 12(2), 98-110.
[3] Johnson, R., Patel, M., & Zhang, X. (2017). Integrating Deep Learning with Numerical Weather Prediction: A Hybrid Approach. Climate Prediction Advances, 8(1), 45-58.
[4] European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 Reanalysis Dataset Documentation. Retrieved from https://www.ecmwf.int/en/ forecasts/dataset/era5
[5] Scikit-learn Developers. (2023). LinearSVR Documentation. Retrieved From https: //scikit-learn.org/stable/modules/ generated/sklearn.svm.LinearSVR.html
[6] Xarray Documentation. Retrieved from https://docs.xarray.dev/en/ stable/

