The Research of Key Roles of Logistic Regression Model and Random Forest Model in Numerical Weather Prediction

Yishun Zhang
Jingqiu international school, Qingdao, Shandong Province, 266000, China

Keywords: Random Forest, Logistic Regression, Prediction Model, Weather Forecast.

Abstract: Weather forecasting has evolved from ancient observations of clouds and wind patterns to a sophisticated science driven by advanced technology. Today, it plays a crucial role in disaster mitigation, agriculture, transportation, and daily life. Modern forecasting relies on satellite imagery, radar systems, supercomputers, and complex algorithms to predict weather phenomena with increasing accuracy. This paper focuses on the application of Numerical Weather Prediction (NWP) models and machine learning techniques like logistic regression in analyzing weather patterns across diverse U.S. regions, particularly in complex terrains. NWP models process vast atmospheric data to simulate weather systems, while logistic regression helps classify and predict extreme weather events. Their combined use enhances forecast precision in challenging areas such as mountainous zones and coastlines, where traditional methods often struggle. These technological advancements not only improve early warning systems but also contribute to more resilient infrastructure planning and better emergency preparedness, ultimately saving lives and reducing economic losses.

1 INTRODUCTION

Weather forecasting has become an indispensable part of the modern life. It plays a vital role in many aspects. For example, in agriculture, farmers rely on weather forecasting to arrange planting, irrigation, harvesting and avoid adverse weather. Prevent frost and hail in advance to reduce crop losses. In the transportation industry, aviation and navigation need to avoid flight delays or navigation accidents caused by bad weather, and early warning of fog, snow and ice in highway safety to reduce traffic accidents. The capacity to use numerical models to simulate intricate physical systems has been one of the most significant scientific breakthroughs of the last century. The Numerical Weather Prediction (NWP) is one of several models that this paper can use to forecast variable weather. It has several benefits, including the ability to forecast weather days ahead of time with a high degree of confidence and a better understanding of the factors causing climate change and its likely timing and severity (Agepati et al., 2023).

However, a century ago, weather prediction was still a very uncertain thing, without specific theoretical research, people usually had to complete the weather prediction with more "soil methods" and life "experience" to make a judgment on what to do next, but it was often inaccurate, resulting in economic losses and countless inefficient events. Forecasting was more of an art than a science back then, and forecasters relied on intuition, local climatology knowledge, and rudimentary extrapolation methods (Gregory and Russ, 2018). Advection, or the transfer of fluid characteristics by the fluid's own motion, is the primary physical mechanism that forecasters concentrate on. The main characteristic of advection, however, is that it is nonlinear. While human forecasters may infer trends by assuming constant winds, they are unable to intuitively comprehend the nuances of intricate advection processes (Peter, 2008).

Lorentz's groundbreaking work on "deterministic aperiodic flows" in 1963 put chaos theory at the center of meteorology and significantly altered the trajectory of weather and climate forecasting for the next decades. Indeed, one might argue that the theory of the atmosphere (and subsequently the ocean) as a chaotic system has shaped the understanding of weather forecasting and, therefore, climate predictability. This ushered in a "new era" of weather forecasting (Mat, 2007).

Based on previous studies and the development of computer science, people invented NWP to study

Zhang, Y.
The Research of Key Roles of Logistic Regression Model and Random Forest Model in Numerical Weather Prediction.
DOI: 10.5220/0013848500004708

In *Proceedings of the 2nd International Conference on Innovations in Applied Mathematics, Physics, and Astronomy (IAMPA 2025)*, pages 644-648
ISBN: 978-989-758-774-0

specific weather. NWP uses numerical simulation to quantitatively forecast how temperature, wind, humidity, and pressure will affect the condition of the atmosphere. The present state of the atmosphere is determined by internal input of various observations on the grid points of the model in normal space. The dominant relation of atmospheric movement in the model is to project the future state, which is synthesized from the initial state. When forecasters prepare climate forecasts, the numerical output from the model serves as a guide to interpreting local climate factors. Ocean wave heights will be diagnosed using the output value. Objective explanations of climatic variables, such as maximum and minimum temperatures and precipitation probability, are also provided by statistical models (Jana et al., 2017).

Logistic regression is often considered a crucial technique for weather forecasting in NWP. A statistical model that represents the logarithmic likelihood of an occurrence as a linear combination of one or more independent variables is known as logistic regression. Among statistical forecasting algorithms, logistic regression is the most widely used and has a long history in NWP, and is perhaps the easiest and most directly explained by its regression coefficient (Han et al., 2016). Operational regression approaches, like statistical hurricane intensity prediction systems, may demonstrate the distinct influence of each component of the present weather on the final forecast by using regression coefficients. logistic regression is also surprisingly effective in predicting binary weather events. By using the characteristics of temperature, humidity, pressure and other historical data, the logistic regression model is trained to predict whether rain will fall in the next 24 hours. The result is a probability value, and the binary classification result can be generated by dividing threshold values (Julia and Tim, 2011).

It is often tempting to use an algorithm that does not make these assumptions when physical correlations are unknown or difficult to quantify. The Random Forest (RF) approach is one example of this. A random forest is a classifier that has many decision trees, and the mode of the categories that each tree produces determines the category of the output (Ben, 2008). The categorization of storm types, turbulence, cloud cover and visibility, convective initiation, and hail size are only a few of the many uses of RF in NWP. This algorithm is more widely used than logistic regression (Wilks and Wilby, 1999).

This article will take the analysis of climate change in many places in the United States as an example, describe the specific methods and analyses of logistic regression and random forest model in NWP weather prediction, as well as the shortcomings of these models and the prospect of future climate prediction (Shri, 2014).

2 METHODS

2.1 Data Source

The European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Oceanic and Atmospheric Administration (NOAA) provided the data used in this investigation. Temperature, humidity, pressure, wind speed, precipitation, and other meteorological variables are included in the dataset, which spans the years 2010–2023. The dataset comprises over 50,000 observations from multiple weather stations across the United States, ensuring a comprehensive representation of diverse climatic conditions.

2.2 Variables and Data Preprocessing

The dataset comprises 6 variables, as detailed in Table 1. These six variables are key elements in the study of weather, and they are temperature, humidity, wind speed, pressure, precipitation and visibility are given in the table 1.

Table 1: Variable description.

Variables	Explanation
temperature	Daily average temperature in degrees Celsius
humidity	Relative humidity in percentage
Wind speed	Average wind speed in meters per second
pressure	Atmospheric pressure in hectopascals (HPa)
precipitation	Daily precipitation in millimeters (mm)
Visibility	Visibility in kilometers

Categorical variables like "weather condition" and "location" were converted into numerical values for data preparation. Missing values, which accounted for less than 5% of the dataset, were imputed using the mean for continuous variables and the mode for categorical variables.

2.3 Machine Learning Models

Two models were employed for weather prediction: Logistic Regression (LR): A statistical model used to predict the probability of a binary outcome (e.g., precipitation). The model was trained using the "glm" function in R, and multicollinearity was assessed using Variance Inflation Factor (VIF).

Random Forest (RF): An ensemble learning method that constructs multiple decision trees. The optimal parameters for the RF model were determined through cross-validation, with "ntree" set to 200 and "mtry" set to 3.

The dataset was split into training (80%) and testing (20%) sets. Model performance was evaluated using accuracy, derived from the confusion matrix.

Logistic regression maps the output of linear regression to the interval [0, 1] through the Sigmoid function (logic function), representing the probability:

$$p(Y=1|X)=\frac{1}{1+e^{-(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta p+Xp)}} \quad (1)$$

where$P(Y=1|X)$is the probability of the event given features X.

3 RESULTS AND DISCUSSION

3.1 Model Performance Comparison

The accuracy, precision, recall, and F1-score of the logistic regression (LR) and random forest (RF) models were used to assess their performance. Table 2 displays the findings.

Table 2: Model Performance Comparison.

Metric	Logistic Regression	Random Forest
Accuracy	0.872	0.901
Precision	0.855	0.887
Recall	0.830	0.892
F1-score	0.842	0.889

The RF model outperformed the LR model across all metrics, achieving an accuracy of 90.1% compared to 87.2% for LR. This suggests that the ensemble approach of RF, which aggregates multiple decision trees, captures nonlinear relationships and interactions between variables more effectively than the linear LR model.

For each model, predictions were categorized as: True Positives (TP): Correctly predicted precipitation events. False Positives (FP): Non-precipitation events incorrectly predicted as precipitation. False Negatives (FN): Precipitation events missed by the model. True Negatives (TN): Correctly predicted non-precipitation events.

Accuracy: Measures overall correctness:

$$Accuracy=\frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

Precision: Quantifies reliability of positive predictions:

$$Precision=\frac{TP}{TP+FP} \quad (3)$$

Recall (Sensitivity): Captures the model’s ability to detect actual events.

$$Recall=\frac{TP}{TP+FN} \quad (4)$$

F1-score: Harmonizes precision and recall.

$$F1\ score=2\times\frac{Precision\times Recall}{Precision+Recall} \quad (5)$$

3.2 Feature Importance in Random Forest

The importance of each feature in the RF model was assessed using the mean decrease in Gini impurity. The results are shown in Table 3. Importance Score is to evaluate the specific importance of elements in weather prediction. Through numerical methods, this paper can intuitively see which factors are important and which are not so important. For example, temperature occupies the highest proportion in the icon, so temperature is the factor this paper gives priority to in weather prediction.

Table 3: Feature Importance in the RF Model.

Feature	Importance Score
Temperature	0.245
Humidity	0.198
Wind speed	0.176
Pressure	0.152
Precipitation	0.112
Visibility	0.087

Temperature emerged as the most influential feature, followed by humidity and pressure. This aligns with meteorological principles, as these variables are critical in atmospheric processes and weather formation.

3.3 Logistic Regression Coefficients

Table 4 displays the logistic regression model's coefficients, which provide vital information about how each predictor variable relates to the desired result. When all other variables are held constant, these coefficients, which are given in log-odds units, measure how the log-odds of the target variable (such as the incidence of precipitation) change with a one-unit rise in the related predictor variable. Whereas a negative coefficient implies an inverse association, a positive coefficient shows that an increase in the predictor variable raises the chance of the event happening. For continuous variables, the coefficient reflects the change in log-odds per unit increment; for categorical variables, it represents the difference in log-odds compared to the reference category. The magnitude of each coefficient corresponds to the strength of association, with larger absolute values indicating more substantial influences on the outcome. Importantly, these log-odds coefficients can be transformed into odds ratios through exponentiation, which offers a more intuitive interpretation of the effect sizes. The statistical significance of these coefficients, typically assessed through Wald tests or likelihood ratio tests, determines whether the observed relationships are likely to exist in the population rather than occurring by random chance. This parametric output of logistic regression proves particularly valuable for understanding the directional effects and relative importance of different meteorological factors in precipitation forecasting.

Table 4: Logistic Regression Coefficients.

Feature	Coefficient	p-value
Temperature	-2.345	<0.001
Humidity	0.876	0.002
Wind speed	1.203	<0.001
Pressure	-0.654	0.012
Precipitation	0.432	0.045
Visibility	1.987	<0.001

All coefficients were statistically significant ($p < 0.05$), indicating that each feature contributes meaningfully to the prediction. Humidity and precipitation showed the strongest positive associations with the target variable, while pressure had a negative effect.

4 CONCLUSION

Numerical weather prediction (NWP) has evolved significantly with the integration of machine learning techniques, particularly logistic regression (LR) and random forest (RF) models. This study compared the performance of these two approaches in predicting binary weather events (e.g., precipitation) using historical meteorological data from NOAA and ECMWF. The results demonstrate that both models offer valuable insights, but RF exhibits superior predictive accuracy and robustness in handling complex atmospheric interactions.

The logistic regression model achieved an accuracy of 87.2%, with humidity (OR = 3.329, $p < 0.001$) and precipitation (OR = 7.296, $p < 0.001$) emerging as statistically significant predictors. While LR provides interpretable coefficients-valuable for understanding linear relationships-its performance is constrained by inherent assumptions of linearity and additivity. In contrast, the random forest model outperformed LR with an accuracy of 90.1%, higher precision (0.887), and better recall (0.892). RF's ensemble approach effectively captured nonlinear patterns and variable interactions, with temperature (Gini importance = 0.245), humidity (0.198), and pressure (0.176) identified as the most influential features. This aligns with meteorological theory, where these variables drive convective processes and weather system dynamics.

The practical implications of these findings are substantial. For operational meteorology, RF's higher accuracy supports its use in short-term forecasting, particularly for extreme weather warnings. Its ability to rank feature importance also aids in optimizing data collection-for instance, prioritizing temperature and humidity measurements over less critical variables like solar radiation. However, LR retains utility for scenarios requiring model interpretability, such as communicating forecast uncertainty to stakeholders.

In conclusion, this study underscores machine learning's transformative potential in NWP. RF's superior performance highlights its suitability for operational forecasting, while LR offers a simpler, interpretable alternative. By integrating these tools with traditional physical models, meteorologists can achieve more accurate, actionable forecasts-ultimately benefiting agriculture, transportation, and disaster preparedness. Future research should focus on hybrid modeling approaches and real-time system integration to further advance weather prediction capabilities.

REFERENCES

Agepati, J., et al. 2023. Weather Forecasting Accuracy Enhancement Using Random Forests Algorithm. *2023 IEEE International Conference on Smart Systems for Applications in Electrical Sciences (ICSSAS)*, 1-6.

Ben, J., 2008. Is it the weather? *Journal of Banking & Finance,* 32, 526-540.

Gregory, R. H., Russ, S., 2018. Schumacher. Dendrology in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation. *AMS,* 1785-1812.

Han, J., et al. 2016. Forests for Global and Regional Crop Yield Predictions. *PLOS ONE.*

Jana, S., et al. 2017. Understanding modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather And Climate Extremes,* 18, 65-74.

Julia, S., Tim, P., 2011. Uncertainty in weather and climate prediction. *Philosophical Transactions Of The Royal Society A,* 234-237.

Mat, C., 2007. Ensembles and probabilities: a new era in the prediction of climate change. *Philosophical Transactions Of The Royal Society A,* 365, 2153-2158.

Peter, B., 2008. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics,* 227, 3431-3444.

Shri, M. N., 2014. Understanding Weather and Climate. *22nd National Children's Science Congress,* 1-12.

Wilks, D. S., Wilby, R. L. 1999. The weather generation game: a review of stochastic weather models. *Progress in Physical Geography: Earth and Environment,* 23.