# Leveraging Machine Learning for Precipitation Prediction: Enhancing Weather Forecast Accuracy

Shunyi Rao*

Xiamen Foreign Language School, Xiamen, 361000, China
*1912100328@mail.sit.edu.cn

**Abstract.** With the rapid development of science and technology, computers have been able to calculate faster than humans. For huge databases, machine learning can organize and analyze them faster than humans, and calculate them quickly through mathematical and physical models. Therefore, this study examines how machine learning can predict precipitation more accurately and efficiently in weather forecasting. In this paper, we first collected the temperature and precipitation data of Xiamen, China in the past 20 years from the National Environmental Information Center, and then studied the potential relationship between the data and how to interact with each other. We also did data visualization processing and used a logistic regression model for prediction. The final research result is a prediction of the precipitation in Xiamen, China in the next year. In the data visualization image after prediction, it can be found that the overall trend of the precipitation is similar to the historical data, but there are also differences, which proves that the prediction obtained by machine learning after analyzing a large number of data includes the historical trend and possible variables. The main research conclusion is that machine learning can make weather prediction more accurate and efficient after analyzing a large amount of data, and the use of efficient prediction models can also make weather prediction more accurate, which is beneficial to the social economy and the natural environment.

**Keywords:** Machine learning, Precipitation Prediction, Logistic regression.

## 1 Introduction

Weather forecasting is an important science, it has positive social significance, and the public is a major user of the weather forecast. There are weather forecast services for the public on TV, radio, newspapers, and the Internet. The weather has a great impact on aviation. Weather has a great influence on electricity consumption, so power companies use weather forecasts to predict electricity consumption. Other private businesses can also use weather forecasts to adjust their demand and supply, for example, supermarkets can serve more drinks on hot days because it can help people plan their daily activities, and it can also help geo-environmental monitoring. Historically, there have been many natural disasters caused by excessive precipitation, which have brought huge losses to the economy and the environment, and too little


B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_8

precipitation has affected human agricultural production and life. It can be seen that precipitation is an important weather factor in human history. Therefore, the role of weather forecast is very important. Accurate prediction of future precipitation can enable people to do preventive work and reduce the inconvenience and loss caused by rainfall. At the same time, monitoring and understanding precipitation at any time can also provide a scientific basis for drought resistance.

Ancient weather forecasting methods are usually based on empirical summaries of observed weather characteristics. In ancient Chinese meteorology, there are abundant records obtained through observations made by astronomical and meteorological instruments. Astronomical and meteorological instruments are an extension of the human senses and are also important tools and means for studying the sun, moon, stars, clouds, rain, wind, and thunder. From feeling the wind, rain, cold, and heat, to trying to understand these natural phenomena that occur in the atmosphere, many dynasties have invented many instruments for observation. Although scientific instruments have been used to observe the weather centuries ago, before the invention of the telegraph, people could only observe the local weather conditions, because the weather conditions are changeable, and the absence of the telegraph means that the data cannot be transferred. Delivered where needed in real-time. So in the 19th century, the invention of the telegraph enabled people to predict the weather through weather data from different regions[1]. With the emergence of computers in modern times, the efficiency of weather forecasting has been greatly improved. Numerical weather forecasting uses computers to simulate the atmosphere. Lewis Fry Richardson proposed in 1922 how to calculate its evolution from a set of meteorological observations[2]. In this way, machine learning algorithms, with their unique data processing capabilities and classification practices, can help improve the accuracy of weather forecasts. Machine learning algorithms can model historical weather data to extract underlying highly relevant structures and patterns to predict future weather conditions. Finally, machine learning algorithms can also delve into the factors that affect weather changes to make more accurate predictions about them in the future.

Section 2 is the related work that includes a review of the past research results on weather items and precipitation prediction. Analyze the challenges and limitations of traditional methods in precipitation prediction, and how machine learning can solve these problems. Section 3 explains the fundamentals of machine learning in weather forecasting, including data processing and classification methods. Data preprocessing steps such as data cleaning, feature extraction, and normalization are described in detail. Introduce the selected machine learning algorithms (this experiment will use the Prophet function as a model) and their application in precipitation prediction. Section 4 describes the datasets used in the experiment, including historical weather data and precipitation observations. Describe the data preprocessing, including handling missing values, outliers, etc. Use Python for data processing, and draw a line chart according to different data relationships to reflect the relationship between the data. Section 5 discusses the strengths of machine learning models, such as their ability to handle complex relationships, adaptive learning, and more. Discuss the limitations of the model and future development directions, such as the impact of data

collection, model parameter tuning, etc. The application of machine learning to precipitation forecasting is summarized, emphasizing its potential for improving the accuracy of weather forecasts. Emphasizes the importance of further research to improve model performance and address challenges such as future climate change. Encourage the promotion of machine learning results in weather forecasting in practical applications, thereby benefiting more social and environmental aspects.

## 2 Related Research

Since the 20th century, data models have been able to learn and make predictions about changing situations. But a century ago, weather predictions were contingent and imprecise, and weather forecasters could only use their crude reasoning techniques, based on local climatological knowledge and intuition-based guesses. Meteorologists at that time plotted the observations of air pressure and other variables on a weather map in the form of symbols and drew lines on points where the air pressure was equal to judge whether there was a certain relationship between the data, whether the air pressure was related to other weather factors have a certain relationship, and then based on one's own experience, memory, and various empirical laws, a map that can predict future weather can be drawn. This method is very inefficient and very unreliable[3]. A simulation-based approach can also be used to form an ensemble. AnEn is the result of 12-15 months of training in forecasting and observations by Environment Canada (EC). It is of equal or greater skill and value than other models producing predictions, and it is also less computationally expensive than other models [4]. Decision trees represent a frequently used decision support tool due to their ease of understanding and interpretation. It emphasizes how data stored about past events can be used to predict future events. But it can also improve the accuracy of building a decision tree and forecasting weather based on it[5]. There is also a way to predict—observation networks: While developing tighter observation networks may be crucial, it only makes sense within the constraints of observability. Furthermore, predictable intrinsic limitations may ultimately lead to tissue expansion[6].

Traditionally, weather forecasting is done through a physical model of the atmosphere, but this model is very unstable due to many other uncertain factors, so its predictions for the future are often inaccurate and unreliable. However, machine learning can take into account the interference of other factors to combine more factors to make accurate predictions for future data. Historically, the tribal people of Mizoram have predicted the weather by citing traditional ecological knowledge that has been passed down through the ages. In the study, the tribe's people recorded 16 different bioindicators of weather forecasts. This is based on the fact that these creatures all have unique behavioral traits that respond to different weather conditions[7]. Different regions divide the year into four or five seasons, and different agricultural activities are carried out in each season according to the climate[8]. Nowadays, with the emergence of big data, machine learning algorithms play an important role in atmospheric science. There is a technique called supervised learning where given some labeled data is available, it can be used as a training dataset[9].

Linear regression models and functional regression models are commonly used in machine learning, and they can learn to detect trends in the weather. Both models have applications in professional weather forecasting[10].

## 3 Method

Machine learning is a computer-based algorithm that enables computer systems to automatically learn to improve performance. The purpose of machine learning is to create a system that can continuously adapt through experience. Machine learning can be used in various fields. In this paper, machine learning can be used to forecast meteorological data, making forecasts more accurate and precise. Machine learning algorithms are a class of algorithms used to derive models from data. Logistic regression is used to predict discrete outcomes, such as whether the weather will be sunny or not. These algorithms provide a reliable and efficient framework for weather forecasting. The linear regression model is the most basic algorithm for machine learning to predict the weather. The linear regression model is very effective and can be used for numericalization (such as classification) and prediction (such as predicting the weather). The linear regression algorithm also has limitations, and it requires a linear relationship between the data. For scenarios with nonlinear relationships, there are countless algorithms for data mining and machine learning. Neural network algorithms are capable of both linear and nonlinear learning algorithms. Weather forecasting is a data-dependent task where machine learning algorithms can predict future weather conditions based on past weather data and existing sensors. For example, meteorological data includes indicators such as temperature, humidity, air pressure, and wind speed. By using machine learning algorithms, it is possible to build a model and infer from this model the relationship of meteorological data to predict the possible future weather conditions. In addition, weather forecasting can also utilize various data sources, such as weather satellites, weather radars, APIs, etc. for weather forecasting research.

### 3.1 Model

Prophet will serve as a model for data prediction. The algorithm in the Prophet function is suitable for data with seasonal changes and regularities, it is an algorithm created by Facebook in 2007[11]. The Prophet model is a model with nonlinear regression like this

$$y_t = g(t) + s(t) + h(t) + \varepsilon(t) \quad (1)$$

g(t) is a trend of different segments, s(t) is different seasonal methods, h(t) is the influence of holidays, and ε(t) describes a white noise item. Among them, the standard Fourier series is used, which is a periodic change. Holidays and other events are a difficult task for machine learning to learn and predict because they are not periodic. The Prophet algorithm is to obtain the predicted value of the time series by fitting these items and finally adding them up.

This paper finds the precipitation and temperature of Xiamen, China from 2000 to 2023 from the system. Considering that this paper studies the prediction of precipitation, and the northern areas often have little precipitation, a huge amount of data is required for machine learning. To study to discover the laws that exist in it, this article chooses Xiamen, China, a southern city with more precipitation. In the process of data preprocessing, the air pressure, AQI, wind direction, and other data irrelevant to the experiment were deleted, and the pandas function was used to process the data. The processed data became more intuitive and orderly. It is beneficial to draw an intuitive line chart in the later stage to find the relationship between the data and predict the final data. In the data processing, the extreme data caused by the influence of extreme weather are deleted, such as the excessive precipitation caused by typhoons and other weather (because this paper does not predict the sharp increase in precipitation caused by extreme weather, even the occurrence of extreme weather often has a certain time law but not included in the research object in this paper).

The Prophet model used in this article uses a time series forecasting algorithm. The data received at different times and describing the change of one or more characteristics over time are called time series data (precipitation in this article is the data that changes over time, so in this article, precipitation is time series data). What time series forecasting does is use historical data to predict the future through the intrinsic and time-related characteristics of multi-dimensional data itself. From the perspective of implementation principles, it can be divided into traditional statistics and machine learning (also divided into deep learning and non-deep learning).

The Prophet algorithm is suitable for situations that are closely related to seasonal and temporal changes (for example, in this article, the Prophet function will be used to predict precipitation, which has strong seasonal changes, usually more in summer, and more in winter usually less precipitation). At the same time, the powerful algorithm of the Prohet function can make up for unrecorded data, and can accurately learn the trend of the data.

## 4 Experiment

The use of machine learning for precipitation prediction requires the machine to learn the historical precipitation situation to find the precipitation law contained therein and make predictions for future precipitation in line with the historical precipitation law (regardless of the extreme data caused by special weather). A huge database can make machine learning more accurate. The website for obtaining the database in this paper is the National Centers For Environmental Information (NOAA)

### 4.1 Data Selection

On the website of the National Environmental Information Center, people can check the precipitation in various regions from the 20th century to the 21st century (different regions have different records). However, cities in the northern region do not have a large amount of precipitation in a year. Machine learning requires a large amount of

data for learning to discover potential patterns between data. Too little precipitation will lead to too small data sets, which will affect the final prediction of precipitation. accuracy. Therefore, the precipitation data collected in this paper are the precipitation data from 2000 to 2023 in Xiamen City, Fujian Province, China. Xiamen was chosen as the data collection point because Xiamen is a coastal city in the south (118°04'04" east longitude, 24°26'46" north latitude), which is greatly affected by the southeast Pacific monsoon and has relatively high temperature and rainy summer. The winter is mild and rainless, and the precipitation season is strong, which is conducive to the data analysis and regular analysis of the machine. There are three analysis features in the database: date (every day from 2000 to 2023), temperature (daily average temperature for each day and the maximum and minimum air temperature for that day), and precipitation (daily average precipitation for each day). This article will explore the relationship between precipitation and time, the relationship between precipitation and temperature, and the relationship between time and temperature so that the machine can learn the potential laws between the data and make the most accurate prediction for the final precipitation.

### 4.2 Data Input

Pandas is one of the main data analysis libraries in Python. It provides a lot of functions and methods to efficiently process and analyze data. What makes pandas so popular is its concise, flexible, and powerful syntax. Pandas have many functions, models, and methods to efficiently complete the organization of the database[12]. This article downloads the file containing the data in CSV format and imports the pandas function to process the data, and then it can get the following table. The data in the obtained table has been sorted in an orderly and intuitive way. The air pressure, wind direction, and other weather factors irrelevant to the experiment were deleted from the obtained data. The date is left, as well as the location where the data was obtained (since the data is obtained at the same location, the factors of different locations are not considered), the average daily precipitation, the average temperature, the maximum temperature, and the minimum temperature per day (drawing the temperature The relationship between precipitation and precipitation can be found in the potential relationship between precipitation and the maximum and minimum temperature of the day). Figure 1 is the processed data.

[3]:

| DATE | STATION | NAME | PRCP | TAVG | TMAX | TMIN |
|---|---|---|---|---|---|---|
| 2000/1/1 | CHM00059134 | XIAMEN, CH | 0.00 | 60 | 72.0 | 52.0 |
| 2000/1/2 | CHM00059134 | XIAMEN, CH | 0.00 | 62 | 76.0 | 53.0 |
| 2000/1/3 | CHM00059134 | XIAMEN, CH | 0.00 | 61 | 68.0 | 56.0 |
| 2000/1/4 | CHM00059134 | XIAMEN, CH | 0.00 | 59 | 71.0 | 53.0 |
| 2000/1/5 | CHM00059134 | XIAMEN, CH | 0.00 | 64 | 77.0 | 55.0 |
| ... | ... | ... | ... | ... | ... | ... |
| 2023/7/24 | CHM00059134 | XIAMEN, CH | 0.00 | 88 | NaN | 79.0 |
| 2023/7/25 | CHM00059134 | XIAMEN, CH | 0.00 | 90 | NaN | 81.0 |
| 2023/7/26 | CHM00059134 | XIAMEN, CH | 0.00 | 89 | 96.0 | 84.0 |
| 2023/7/27 | CHM00059134 | XIAMEN, CH | 0.02 | 88 | NaN | NaN |
| 2023/7/28 | CHM00059134 | XIAMEN, CH | 6.04 | 80 | 94.0 | 75.0 |

8608 rows × 6 columns

Figure 1: Processed data. (Picture credit: Original)

### 4.3 Relationship between data

In the above code, using the Plotly function, which is a Python visualization library that can be used to create interactive charts, graphs, and visualizations. Define data as the imported file, and in Figure 2, the x-axis is the Date, the y-axis is the Average Temperature, and the Title is the Relationship between temperature and date.

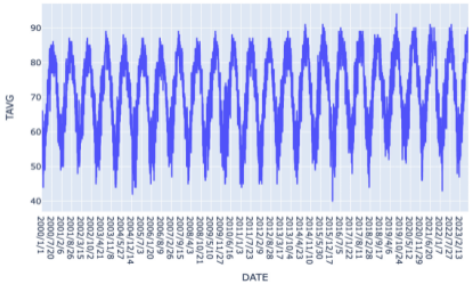


Figure 2: Relationship between temperature and date. (Picture credit: Original)

This is a line chart of the relationship between temperature and date drawn by Python. It can be seen from the line chart that the temperature has strong seasonality, with the highest value appearing in summer and the lowest value appearing in winter. Overall, there is no great change with the change of time, but the annual maximum temperature shows a rising trend.

In Figure 3, the x-axis is the Date, the y-axis is Precipitation, and the Title is the Relationship between precipitation and date.

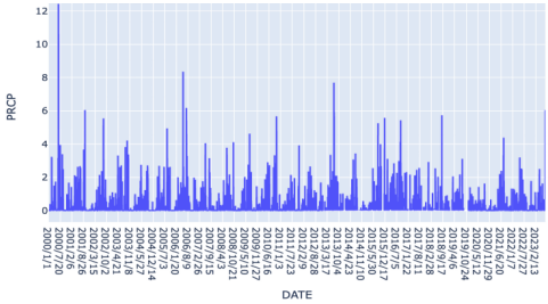


Figure 3: Relationship between precipitation and date. (Picture credit: Original)

This is a line chart of the relationship between precipitation and date drawn by Python. It can be seen from the line chart that the highest value of precipitation occurs in June, July, and August (summer), which also indicates that precipitation has a strong seasonality. With the change of time, the maximum precipitation showed a decreasing trend.

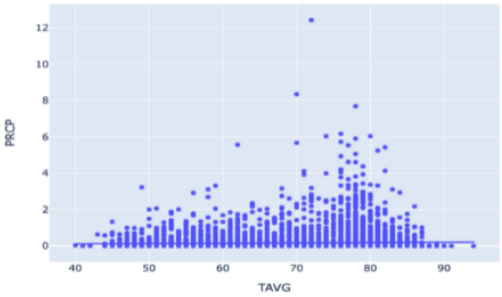


Figure 4: Relationship between temperature and precipitation. (Picture credit: Original)

In Figure 4, the x-axis is the Date, the y-axis is Precipitation, and the Title is the Relationship between precipitation and date. This is a line chart of the relationship between precipitation and temperature drawn by Python. It can be seen from the line chart that precipitation also increases with the increase in temperature. The highest value of precipitation occurs around 70 degrees Fahrenheit (which may be due to extreme weather caused by special circumstances, such as typhoons), and the precipitation is high and stable around 75-80 degrees Fahrenheit. From this, it can be judged that the maximum annual precipitation should occur between 75 and 80 degrees Fahrenheit, and the maximum value in a year at 70 degrees Fahrenheit is due to special circumstances.

### 4.4 Prediction

Once enough data sets have been collected and relationships between the data explored, machine learning is used to predict precipitation. The prophet model is introduced in the code, which determines the forecast time as the future year (365 days). Figure 5, can derive historical data, and the blue lines in the figure are the prediction of the precipitation for the coming year. The variation of precipitation throughout the year is small, with higher precipitation in July and August in 2024, which is consistent with the relationship between the analysis of historical data.

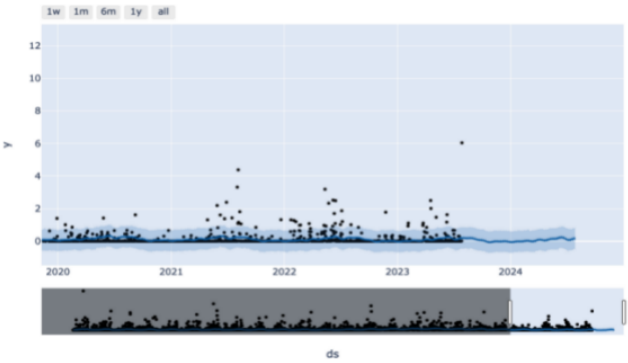


Figure 5: The final predicted data. (Picture credit: Original)

## 5 Conclusion

Weather forecasts are closely related to our lives, but weather factors are full of uncertainties. Weather forecasts are sometimes inaccurate and untimely, which will cause certain losses to human society and the economy. Using machine learning to learn more contextual data and use more accurate models is an important way to improve forecast accuracy. Compared to traditional methods of weather forecasting and data analysis, machine learning models' ability to identify trends in a wide range of data means that machine learning can continuously improve its performance over time with more data, enabling it to make better Previously more accurate predictions or decisions. This experiment uses the Prophet model to identify the precipitation and temperature data of the past 23 years, find out the trend, and make a forecast for the precipitation in the coming year. However, the Prophet model used in this experiment requires a large amount of data to make accurate predictions, which is a challenge for organizations with limited datasets. In this experiment, the high precision of the Prophet model is its advantage, allowing it to continue to be used in future research. In real life, machine learning has become an important way for people to predict the weather. Further research based on existing models and algorithms in the future can make predictions more accurate and improve model performance to cope with

challenges such as future climate change diversity. It is hoped that machine learning can be used to predict more weather factors in the future, and some initiatives can be made to protect the earth's ecological environment. All in all, using machine learning to benefit society and the environment is the ultimate goal of future research studies.

## References


1. Namias, J. (1968). long range weather forecasting—history, current status and outlook1. Bulletin of the American Meteorological Society, 49(5–1), 438–470.
2. RICHARDSON LF. (1922), Weather prediction by natural process, Cambridge University Press, https://archive.org/details/weatherpredictio00richrich.
3. Lynch, P. (2008). The origins of computer weather prediction and climate modeling. Journal of Computational Physics, 227(7), 3431–3444. https://doi.org/10.1016/j.jcp.2007.02.034
4. Monache, L. D., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic Weather Prediction with an Analog Ensemble. Monthly Weather Review, 141(10), 3498–3516. https://doi.org/10.1175/mwr-d-12-00281.1
5. Petre, Elia Georgiana. "A decision tree for weather prediction." Bul. Univ. Pet.–Gaze din Ploies, ti 61.1 (2009): 77-82.
6. Yano, J., Ziemiański, M. Z., Cullen, M. J. P., Termonia, P., Onvlee, J., Bengtsson, L., Carrassi, A., Davy, R., Deluca, A., Gray, S. L., Homar, V., Kohler, M., Krichak, S. O., Michaelides, S., Phillips, V. T. J., Soares, P. M. M., & Wyszogrodzki, A. A. (2018). Scientific Challenges of Convective-Scale Numerical Weather Prediction. Bulletin of the American Meteorological Society, 99(4), 699–710. https://doi.org/10.1175/bams-d-17-0125.1
7. Chinlampianga, M. (2011, January 1). Traditional knowledge, weather prediction and bioindicators : A case study in Mizoram, Northeastern India. http://nopr.niscpr.res.in/handle/123456789/11083
8. Sanni, S. , Oluwasemire, K. and Nnoli, N. (2012) Traditional capacity for weather prediction, variability and coping strategies in the front line states of nigeria. Agricultural Sciences, 3, 625-630. doi: 10.4236/as.2012.34075.
9. Bochenek, B., & Ustrnul, Z. (2022). Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. Atmosphere, 13(2), 180. https://doi.org/10.3390/atmos13020180
10. Holmstrom, Mark, Dylan Liu, and Christopher Vo. "Machine learning applied to weather forecasting." Meteorol. Appl 10 (2016): 1-5.
11. Zhao, Daren, et al. "Prediction of global omicron pandemic using ARIMA, MLR, and Prophet models." Scientific reports 12.1 (2022): 18138.
12. McKinney, Wes. "pandas: a foundational Python library for data analysis and statistics." Python for high performance and scientific computing 14.9 (2011): 1-9.