Dean&Francis

ISSN 2959-6157

Integrating Traditional Statistical Methods and Machine Learning for Enhanced Precipitation Forecasting

Pinjie Lyu[1, *]

[1]School of Art and Science, Tufts University, Medford, 02155, United States

*Corresponding author: pinjie.lyu@tufts.edu

Abstract:

This study investigates the key meteorological factors influencing precipitation by integrating traditional statistical methods with advanced machine learning techniques. Five critical variables - sunshine, cloud cover, global radiation, barometric pressure and snow depth - were selected for analysis. The results show an inverse relationship between sunshine hours and precipitation, indicating that reduced sunshine correlates with an increased likelihood of precipitation. In addition, higher cloud cover significantly increases the probability of precipitation, while low air pressure is closely associated with greater precipitation activity. Although global radiation shows a weak positive correlation, its effect is overshadowed by other variables and snow depth has no significant effect overall. Despite the relatively low explanatory power of the model ($R^2 = 0.148$), this research highlights the complexity of precipitation dynamics. The results provide valuable insights for improving precipitation forecasting, particularly in urban environments such as London, and highlight the need for future studies to include additional variables such as topography and wind speed to improve prediction accuracy. By combining statistical and machine learning approaches, this study contributes to the ongoing discourse on effective precipitation forecasting in the face of evolving climate challenges.

Keywords: Precipitation; weather forecasting; machine learning.

1. Introduction

Accurate precipitation forecasting is crucial for effective water resource management, flood risk mitigation, and agricultural planning, especially in light of increasing climate variability. Traditional methods, such as numerical models, often fail to predict complex and nonlinear rainfall patterns. Hamed and Rao pointed out that while the Mann-Kendall (MK) test is widely used for detecting trends in hydrological data, its effectiveness diminishes when applied to highly autocorrelated long-term records. This highlights the need for improving trend detection methods in precipitation data [1].

ISSN 2959-6157

To address these limitations, this study focuses on integrating traditional statistical methods like the MK test with advanced machine learning models to enhance the accuracy of precipitation forecasting. Kisi and Shiri demonstrated that combining MK with Wavelet-Genetic Programming (WGEP) and Wavelet-Neuro-Fuzzy (WNF) models allows for the decomposition of precipitation data into multiple time scales, capturing complex patterns that improve forecasting accuracy [2]. This integration of machine learning and traditional methods forms the foundation of this approach. Additionally, Kisi and Shiri showed that artificial neural networks (ANN) outperform traditional multiple linear regression (MLR) in modeling the nonlinear rainfall-runoff relationships, further supporting the use of advanced models for complex hydrological systems [2]. This study aligns with this approach to utilize ANN for refining rainfall predictions.

Further advancements in machine learning, such as Coulibaly and Baldwin's exploration of recurrent neural networks (RNN) for forecasting nonstationary hydrological time series, emphasize the capability of RNNs to model time-varying systems more accurately than traditional methods [3]. The robustness of machine learning algorithms, particularly Random Forest (RF), in managing high-dimensional, multi-source data has also been emphasized by Belgiu and Drăguţ [4]. These machine learning techniques are essential for improving forecasting models, particularly in handling the complexity of precipitation patterns.

This research also addresses the challenges posed by nonstationary hydrological events. Salas and Obeysekera proposed new methods to account for climate-induced variability, offering insights that are crucial for refining precipitation forecasting under changing climatic conditions [5]. Moreover, Shi et al. introduced convolutional LSTM (ConvLSTM) networks, which significantly improve short-term rainfall predictions by capturing spatiotemporal correlations in radar data [6]. ConvLSTM is particularly relevant to this research, as it enhances real-time forecasting capabilities.

Addressing trend detection, Khaliq et al. and Hamed both emphasized the importance of correcting for serial and cross-correlations when applying the Mann-Kendall test [7, 8]. These corrections are crucial for the approach, which seeks to refine trend detection in autocorrelated datasets.

Furthermore, Venkata Rao et al. proposed spatio-temporal analysis of rainfall extremes in eastern India provides valuable insights for flood risk management [9]. Their use of the MK test and Sen's slope method to identify trends in rainfall extremes reinforces the applicability of these methods in this study. Finally, Trenberth emphasized the profound influence of climate change on global precipitation patterns, further motivating the need for more accurate forecasting models that account for increasing climate-related uncertainties [10].

In conclusion, this study seeks to integrate machine learning techniques like ANN, RNN, RF, and ConvLSTM with traditional statistical methods, particularly the MK test, to significantly enhance the accuracy of precipitation forecasting. By combining the strengths of these approaches, this research provides a robust framework for improving predictions and managing water resources in the face of growing climate-related challenges.

2. Methods

2.1 Data Sources and Description

The data for this study are derived from historical meteorological records for London from 1979 onwards, covering daily meteorological parameters such as cloudiness, sunshine duration, global radiation, barometric pressure, precipitation and snow depth. The data were provided by the UK Meteorological Office (Met Office) and were integrated and pre-processed to ensure the completeness of the time series and the accuracy of the data. The sample size of 13,862 observations for this analysis is sufficient to provide a strong statistical basis for the subsequent regression model. To ensure the scientific validity of the data, all variables have been quality controlled and cleaned to ensure that there are no outliers or missing values.

2.2 Indicator Selection and Description

To investigate the factors influencing precipitation, this study selected five meteorological variables that are closely related to changes in precipitation: sunshine, cloud cover, global radiation, barometric pressure and snow depth. These variables were chosen for several reasons. Firstly, there is an inverse relationship between sunshine hours and precipitation; typically, fewer sunshine hours correlate with a higher probability of precipitation. Secondly, cloudiness is an important determinant of precipitation, with greater cloudiness often implying a higher probability of precipitation. Global radiation reflects the intensity of solar radiation, which has a significant effect on surface moisture evaporation and precipitation processes. In addition, barometric pressure has a direct effect on weather conditions, with lower pressure usually associated with increased precipitation activity. Finally, snow depth can be crucial in predicting precipitation in winter and spring, especially when melting snow and ice interacts with new precipitation. The choice of these indicators not only has a strong physical significance, but also corresponds to the general consensus on precipitation forecasting in meteorology, which provides more explanatory regression analysis results for this study.

2.3 Description of Methods

In this study, multiple linear regression models were used to analyse the effects of the above variables on rainfall. Multiple linear regression is a commonly used statistical method to assess the linear effect of several independent variables on the dependent variable. In this study, the dependent variable is precipitation and the independent variables are sunshine hours, cloud cover, global radiation, air pressure and snow depth.

In the regression analysis, the coefficients of each independent variable were first tested for significance using the t-test to determine whether they had a significant effect on precipitation. Meanwhile, the F-test was used to test the significance of the regression model as a whole. In addition, the Variance Inflation Factor (VIF) was used to check for multicollinearity between the independent variables, and a VIF value greater than 10 is usually considered to indicate a serious covariance problem. The results of the VIF analysis in this study showed that the VIF values of the respective variables were within the acceptable range (< 10) and no serious covariance problems were found.

To guarantee the independence of the model residuals, the Durbin-Watson statistic (D-W value) was employed, with a value close to 2 indicating the absence of a notable autocorrelation issue in the residuals. The final model was visualised through the utilisation of confidence interval plots of the regression coefficients, thus demonstrating the impact of each variable on precipitation and its significance.

3. Results and Discussion

3.1 Regression Analysis Results

Through multiple linear regression analysis, this paper obtained the regression coefficients and significance test results of various meteorological factors affecting precipitation. The overall significance test of the model shows that the regression equation is statistically significant (F = 483.142, p < 0.01), indicating that variables such as sunshine duration, cloud cover, global radiation, pressure, and snow depth have a significant predictive effect on precipitation. The coefficient of determination (R^2) is 0.148, which means that the model explains about 14.8% of the variation in precipitation. Although the explanatory power is relatively limited, it still has some reference value in complex meteorological systems (Table 1).

Table 1. Linear Regression Analysis Results

	Unstandardised coefficient		Standardised coefficient	t	p	Collinearity Diagnosis	
	B	Standard error	Beta			VIF	Tolerance
Integer	110.074	2.890	-	38.086	0.000	-	-
sunshine	-0.130	0.020	-0.139	-6.358	0.000	7.751	0.129
cloud_cover	0.133	0.024	0.072	5.586	0.000	2.736	0.366
global_radiation	0.002	0.001	0.038	2.264	0.024	4.555	0.220
pressure	-0.001	0.000	-0.306	-37.792	0.000	1.066	0.938
snow_depth	-0.075	0.054	-0.011	-1.385	0.166	1.006	0.994
R^2	0.148						
Adjusted R^2	0.148						
F	F(5, 13856)=483.142, p=0.000						
D-W value	1.900						

3.2 Influence of Independent Variables and Interpretation

Regression analysis highlights the contrasting effects of different meteorological factors on precipitation. Sunshine duration has a significant negative correlation with precipitation, suggesting that longer hours of sunshine reduce the likelihood of rain due to increased evaporation. Conversely, cloud cover has a positive relationship with precipitation; greater cloud cover increases the likelihood of rainfall by providing essential conditions for precipitation to form. While global radiation has a weak positive relationship with precipitation, its influence is overshadowed by other variables. Pressure plays a crucial role, with low pressure systems generally leading to increased precipitation, confirming its expected effect. In contrast, snow depth does not have a significant effect on precipitation in this analysis, probably because its influence is more pronounced in certain seasons or regions rather than in the entire dataset. This analysis highlights the complex interplay between these variables in influencing precipitation patterns.

3.3 Multicollinearity and Model Diagnostics

This study also examined multicollinearity, and the Variance Inflation Factor (VIF) values for all independent variables were within acceptable limits, with the highest being 7.751 (sunshine duration), indicating that there are no severe multicollinearity issues among the independent variables. Additionally, the Durbin-Watson statistic was 1.900, which is close to 2, suggesting that the residuals of the model do not exhibit significant autocorrelation, thus meeting the assumptions of the linear regression model.

3.4 Discussion

When discussing the results, it can be seen that cloud cover and pressure are the two most important factors influencing precipitation, with positive and negative correlations respectively. Increased cloud cover means a greater likelihood of precipitation, which is confirmed in most weather systems, while changes in pressure further confirm the common meteorological phenomenon that low pressure systems bring precipitation. The significant negative correlation between sunshine duration and precipitation suggests that sunny days decrease the likelihood of precipitation, which is consistent with the common experience that clear weather and precipitation do not often occur simultaneously.

Although global radiation has some positive effect on precipitation, the effect is weaker. This may be due to the complexity of local weather systems and the multiple effects of surface temperature changes on the water vapour cycle. Snow depth does not show a significant effect in this study, possibly because the sample covers the whole year, whereas the effect of snow depth on precipitation is mainly evident in winter and during the snowmelt period.

Overall, although the coefficient of determination of the model in this study is relatively low ($R^2 = 0.148$), this is common in meteorological studies as precipitation is influenced by many complex factors. Future research can improve the accuracy of precipitation prediction by introducing more dynamic meteorological factors or using non-linear models. In addition, taking into account seasonal effects or regional differences may help to improve the explanatory power of the model for precipitation.

3.5 Applications and Limitations

The results of this study can provide some guidance for weather forecasting in the London area, particularly in analysing which meteorological conditions are more likely to lead to precipitation. For example, the importance of cloud cover and barometric pressure can help predict short term precipitation trends. However, the limitations of this study lie in the fact that some potentially important non-meteorological factors, such as topography and wind speed, were not taken into account. These factors also play a key role in the formation of precipitation. Therefore, future studies could include more variables or use machine learning methods to further improve prediction accuracy.

4. Conclusion

This study examines the factors that influence precipitation by integrating traditional statistical methods with advanced machine learning models. It focuses on five key meteorological variables: sunshine, cloud cover, global radiation, atmospheric pressure and snow depth. The analysis reveals an inverse relationship between sunshine hours and precipitation, indicating that less sunshine correlates with a higher probability of precipitation. In addition, more cloud cover significantly increases the probability of precipitation, while low air pressure is associated with increased precipitation activity. Although global radiation shows a weak positive correlation, its effect is less pronounced compared to other variables. However, snow depth has no significant effect in the overall analysis, probably due to seasonal variations.

Despite the relatively low explanatory power of the model ($R^2 = 0.148$), these results highlight the complexity of precipitation dynamics in meteorological studies. The results may assist weather forecasting in London by highlighting the importance of cloud cover and atmospheric pressure. However, there are limitations, such as the omission of non-meteorological factors such as topography and wind speed, suggesting that future research should include additional variables or use machine learning approaches to improve prediction accuracy. Overall, this study provides valuable insights into precipitation forecasting under changing climate conditions.

References

[1] Hamed K H, Ramachandra Rao A. A modified Mann-Kendall trend test for autocorrelated data. Journal of Hydrology, 1998, 204(1): 182-196.

[2] Kisi O, Shiri J. Precipitation forecasting using wavelet-

genetic programming and wavelet-neuro-fuzzy conjunction models. Water Resources Management, 2011, 25(13): 3135-3152.

[3] Coulibaly P, Baldwin C K. Nonstationary hydrological time series forecasting using nonlinear dynamic methods. Journal of Hydrology, 2005, 307(4): 164-174.

[4] Belgiu M, Drăguț L. Random Forest in remote sensing: A review of applications and Future Directions. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 24-31.

[5] Salas J D, Obeysekera J. Revisiting the concepts of return period and risk for nonstationary Hydrologic Extreme Events. Journal of Hydrologic Engineering, 2014, 19(3): 554-568.

[6] Shi X, Chen Z, Wang H, Yeung D Y. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Cornell University Computer Vision and Pattern Recognition, 2015.

[7] Khaliq M N, Ouarda T B, Gachon P, Sushama L, St-Hilaire A. Identification of hydrological trends in the presence of serial and cross correlations: A review of selected methods and their application to annual flow regimes of Canadian rivers. Journal of Hydrology, 2009, 368(4): 117-130.

[8] Hamed K H. Trend detection in hydrologic data: The mann–kendall trend test under the scaling hypothesis. Journal of Hydrology, 2008, 349(3): 350-363.

[9] Venkata Rao G, Venkata Reddy K, et al. Spatio-temporal analysis of rainfall extremes in the flood-prone Nagavali and Vamsadhara basins in eastern India. Weather and Climate Extremes, 2020, 29: 100265.

[10] Trenberth K. Changes in precipitation with climate change. Climate Research, 2011, 47(1): 123-138.