K-Means and C4.5 Decision Tree Based Prediction of Long-Term Precipitation Variability in the Poyang Lake Basin, China

Dan Lou 1, Mengxi Yang 2, Dawei Shi 3,*, Guojie Wang 1✉, Waheed Ullah 1, Yuanfang Chai 4 and Yutian Chen 1

Abstract: The machine learning algorithms application in atmospheric sciences along the Earth System Models has the potential of improving prediction, forecast, and reconstruction of missing data. In the current study, a combination of two machine learning techniques namely K-means, and decision tree (C4.5) algorithms, are used to separate observed precipitation into clusters and classified the associated large-scale circulation indices. Observed precipitation from the Chinese Meteorological Agency (CMA) during 1961–2016 for 83 stations in the Poyang Lake basin (PLB) is used. The results from K-Means clusters show two precipitation clusters splitting the PLB precipitation into a northern and southern cluster, with a silhouette coefficient ~0.5. The PLB precipitation leading cluster (C1) contains 48 stations accounting for 58% of the regional station density, while Cluster 2 (C2) covers 35, accounting for 42% of the stations. The interannual variability in precipitation exhibited significant differences for both clusters. The decision tree (C4.5) is employed to explore the large-scale atmospheric indices from National Climate Center (NCC) associated with each cluster during the preceding spring season as a predictor. The C1 precipitation was linked with the location and intensity of subtropical ridgeline position over Northern Africa, whereas the C2 precipitation was suggested to be associated with the Atlantic-European Polar Vortex Area Index. The precipitation anomalies further validated the results of both algorithms. The findings are in accordance with previous studies conducted globally and hence recommend the applications of machine learning techniques in atmospheric science on a sub-regional and sub-seasonal scale. Future studies should explore the dynamics of the K-Means, and C4.5 derived indicators for a better assessment on a regional scale. This research based on machine learning methods may bring a new solution to climate forecast.

Keywords: Poyang Lake basin; K-means; C4.5 decision tree; precipitation; climate indices; dynamics

1. Introduction

China’s largest freshwater lake, the Poyang Lake, is located in the middle-lower Yangtze River Basin in a humid subtropical monsoon climate. The changes in the East Asian Monsoon have led to heterogeneous spatiotemporal changes in precipitation, with obvious seasonal and regional differences in the Poyang Lake basin [1]. Changes in precipitation have a profound impact on the conservation of the ecological environment, restoration and conservation of biodiversity [2], and water resource management and monitoring of a region [3], and flood control mitigation [4]. Thus, a more accurate prediction of the precipitation pattern in Poyang Lake has become a research hotspot during recent decades.
To date, many studies have been conducted with emphasis on predicting the future precipitation pattern from regional to global scale employing multiple techniques, including model projections [5,6], empirical relationship forecast approaches [7], and machine learning algorithms method [8–10]. The most frequently used models are the general circulation model (GCM), forced with likely future GHG emission scenarios. The GCMs simulate the historical, current, and future precipitation changes on the global scale, forming the core contents of the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report [11–13]. However, the low spatial resolutions of the simulations from the GCMs make it difficult to provide an accurate precipitation prediction in the Poyang Lake basin for policymakers. Thus, projections of GCMs cannot be readily used in catchment-scale applications such as future precipitation modeling [8]. The empirical forecast approaches, such as multiple linear regressions or canonical correlation analysis (CCA), is a form of multiple linear regression applied to multivariate pattern predictands [14]. These empirical relationships forecast approaches may be efficient tools for predicting short-term rainfall. However, some studies find that the empirical forecast approaches such as multiple linear regressions or CCA have no skill in some regions for predicting rainfall [15].

With the advances in computer hardware and increased computing memory, storage, and network capabilities [16], the application of machine learning in atmospheric sciences has increased [17]. The applications of such in-depth techniques include climate change prediction [18], and the modeling of hydrological processes [19], especially since 2000. Hartigan et al. [20] used linear regression and support vector regression (SVM) for attributing climate drivers of annual precipitation and mean maximum temperature over Canberra, Australia. The attribution factors explored for precipitation were large-scale circulation and oceanic forcing, whereas global warming was associated with changes in mean maximum temperature. The study concluded that machine learning techniques could improve our understanding of the prediction and forecast of dynamics of the atmospheric variables. Chivers et al. [21] applied a two-step analysis method for imputing missing data in ongoing precipitation datasets over the United Kingdom at an hourly scale for a better climate and weather forecast. An integrated two-step analysis technique including multiple machine learning algorithms appeared to overperform the traditional surface fitting techniques that utilize the data from the nearest rain-gauges. Teegavarapu et al. [22] evaluated multiple data-driven models, including inverse distance, correlation weighting procedure, linear weight optimization, and artificial neural networks to fill existing gaps in precipitation data. The results inferred the improved performance of the single best estimator based on multiple evaluation criteria with the rest of the methods outputs. Miao et al. [23] used a deep neural network comprised of a convolution and Long Short Term Memory (LSTM) recurrent module for improving the monsoon precipitation prediction based on well-resolved atmospheric dynamical fields. Results inferred relative improvement in comparison with ECMWF forecast and enriched downscaling as compared to quantile mapping, SVM, and Neural networks.

The spatiotemporal distribution of precipitation in Poyang Lake is uneven, with obvious seasonal and regional differences [24]. The precipitation concentration period is from early March to June, accounting for about 54.2% of the total annual precipitation [25]. At present, numerical and computational such as regression analysis, singular spectrum analysis, neural networks, and other mathematical methods are used to predict precipitation based on historical data [26,27]. Meanwhile, the climate numerical model is also another important method applied in predicting precipitation [28,29]. With the promotion of big data theories and methods, the application of data mining models gradually shows their advantages. These methods can quickly and effectively discover potential laws and key information from a large number of complex databases and use these rules and information combinations to select factor attributes reasonably for building prediction models. As a result, the model prediction accuracy rate reaches a considerable level, and it has been widely used in the study of nonlinear meteorological problems [30]. The K-Means clustering technique is widely used in meteorological studies globally with the aim to model the
multiscale precipitation variation in different climate, land, and ocean masses [31,32]. The decision tree algorithm is an important technique in data mining, which includes several methods: C4.5 algorithm, CART algorithm, PUBLIC algorithm, SLIQ algorithm, and so on. Among them, the C4.5 algorithm proposed has been well applied in the meteorological field due to its simple calculation, high data processing efficiency, and easy model interpretation [33].

In this study, we first investigated the level of the similarity in the summer precipitation over the Poyang Lake based on using the K-means [34] and evaluating clustering’s results with silhouette coefficient [35], we reasonably and objectively split the whole Poyang Lake basin into sub-regions using the classical clustering machine learning algorithm [36]. With an emphasis on the sub-regional precipitation patterns and their attribution to large-scale climate drivers, we then introduced the decision tree prediction model (C4.5 based algorithm) for predicting the summer precipitation patterns in the Poyang Lake basin. The findings of the study will help provide a practical, convenient, and effective model for precipitation prediction in the Poyang Lake basin. The rest of the paper contains a description of the study area, data, and methods in Section 2, followed by results in Section 3, discussion in Section 4, and conclusion in Section 5. The improved precipitation prediction can provide more accurate climate information for policymakers.

2. Study Area

The Poyang Lake basin is located in the middle and lower reaches of the Yangtze River over China (Figure 1a), fed by its five major tributaries, including Ganjiang, Fuhe, Xinjiang, Raohe, and Xiuhe. The total drainage area of the Poyang Lake basin is ~16,220,000 km², covering ~97% of Jiangxi province. The altitude of the Poyang lake basin is from ~191 m to ~2127 m above mean sea level, as shown in Figure 1b, shaping the Poyang Lake basin area into a typical valley. The dominant land cover classes derived from MODIS [37] include water bodies, croplands, and forests with moderate grasslands and urban areas (Figure 1c). The climate of the Poyang Lake can be characterized as subtropical warm humid, with monsoon east Asian monsoon precipitation system as the primary driver of the water cycle. The annual average temperature and rainfall are 17.6 °C and 1639.42 mm respectively. The seasonal precipitation is unevenly distributed, accounting for about 42–53% from April to June. The extreme precipitation frequently occurs from June to July while frequently drought in August. The interannual variability of precipitation is also very big. The difference between the rainy and drier years is almost doubled. This is also one of the major reasons for the frequent occurrence of droughts and floods in PLB.

During recent years, with the accelerated urbanization and increase of multiple land development projects, several challenges such as soil erosion, changes in inflow pattern into the lake, sedimentation of the lake, and enhanced pollution of the lake have threatened the regional flora and fauna [38]. Due to the increase in the global temperature and precipitation patterns with more extreme events projected to increase [39], the Poyang Lake basin needs more detailed studies exploring changes in the water and energy cycle.
3. Data and Methods

3.1. Data

The daily precipitation observations used in the study were collected from the Chinese Meteorological Agency (CMA) for a total of 87 meteorological stations, which density is shown in Figure 2a. We adopted the standard normal homogeneity test (SNHT) to ensure the consistency and consistency of the data [40]. The purpose of the SNHT technique was primarily to detect outliers or spikes in a dataset that could be attributed to non-climatic factors, which most often were used for homogeneity estimation of climate data records. The analysis period of this paper was taken from 1961 to 2016. At the same time, in order to analyze the different time-scale precipitation variations, the daily data were also converted into the monthly and annual totals precipitation. After the quality screening, the number of weather station observations used were 83. The missing data in some of the stations were less than ~3%, and hence such small missing observations will not influence the results to a greater extent [41]. Figure 2b shows the summer totals precipitation climatology employed over the 1961–2016 study period. We see that the northern part of the PLB exhibited more precipitation than the southern in summer. Most of the stations with total values being observed at >550 mm in the summer and northeastern regions have the highest precipitation with the value above 600 mm over the whole PBL, while the central and southern regions received between 450 mm and 550 mm.
The climate index dataset from National Climate Centre (NCC) were used (http: //cmdp.ncc-cma.net/Monitoring/cn_index_130.php, accessed on 28 June 2021) for a potential attribution of the large-scale climate drivers of the Poyang Lake basin. The climate signal during spring (March, April, and May) was used as a predictor to forecast the summer rainfall in sub-regions of the Poyang Lake Basin. In the land-atmosphere coupling domain, similar approaches have been widely used with a potential time-lag as a forecast indicator [42]. The 130-item climate index dataset was obtained by averaging the March, April, and May values of each index from the database, which provided a wide range of choices for selecting a potential index as a forecast indicator of Poyang Lake basin sub-regional precipitation variation.

3.2. Methods
3.2.1. K-Means

K-Means clustering is an unsupervised clustering algorithm [43–47] with the ability to automatically classify N samples of data in G-dimensional space into k number of non-overlapping clusters according to their descriptive characteristics. The similarity of data within a cluster is large, and the similarity of objects among different clusters is small. The main feature of the K-Means algorithm is to determine k initial cluster centers randomly, to classify and to divide the source points based on distance comparison, and to calculate the new cluster centroids. The next round of iteration is performed until the center position is unchanged and stops the classification. The process of classification is actually the process of minimizing errors. The K-Means minimization is to minimize the sum of the distances between all points and their associated cluster centers. The evaluation index of SSE (Sum of Squared Errors) and the calculation equation are as follow:

\[
SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2
\]  

(1)

where \( C_i \) is \( i \)th cluster; \( p \) is the sample point of \( C_i \); \( m_i \) is the cluster center of \( C_i \); \( k \) is the number of clusters based on prior knowledge; SSE is the sum of squares clustering errors for all points in the source cloud, representing the effects of clustering. In K-mean, the purpose of selecting the clustering center is to minimize the error caused by \( p \), and relocating the
clustering center purpose is to minimize the error caused by \( m_i \), for minimizing the error during each iteration.

3.2.2. C4.5

The decision tree C4.5 algorithm is a non-parametric supervised machine learning technique used to generate tree-like classification rules based on the induction of data features, usually from discrete values in nature [48–52]. In the current work, the C4.5 algorithm was used as the classifier, summarizing the classification rules from a set of random instance cases. The model building processes of the C4.5 algorithm can be defined as follows: The construction processes of the algorithm started from “which feature in the feature attribute set \( U \) will be tested at the root node of the tree”. The feature attribute with the best classification ability was selected as the root node of the tree, and then the root node with each possible value of the node feature was used to generate a branch and arrange the training sample set \( D \) under the appropriate branch; repeat the whole process, using each branch node associated training sample to select the best feature tested at the node. There were 4 feature parameters \((h, w, s, \text{ and } p)\) in the feature parameters set \( U \). We used the gain rate of the C4.5 to select the best partition feature attributes. The specific steps [53] were as follows:

Step 1: Calculating information entropy

Information entropy is the most commonly used index in measuring the purity of a sample set. It chooses the attribute with the highest information gain as the split attribute of node \( N \). This attribute minimizes the amount of information required for tuple classification in the result partition. The expected information is required to classify the tuples in \( D \), following the below formula:

\[
\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2 p_i
\]

where \( m \) refers to the number of different types of elements in the result set, and \( p_i \) is the ratio of the number of category elements of the \( i \)-th to the total number of the sample set.

Step 2: Calculating the information entropy of each attribute

Assuming the tuples in \( D \) are divided according to attribute \( A \), and attribute \( A \) divides \( D \) into \( n \) different classes. After the division, obtaining an accurate classification of the information entropy of each attribute need to be measured by the following formula:

\[
\text{Info}_A(D) = -\sum_{j=1}^{n} \left| D_j \right| \left| D \right| \times \text{Info}(D_j)
\]

where \( A \) is the attribute classification of \( D \), \( D_j \) is the number of different categories in the sample set, \( D \) is the total number of the whole sample set, \( \text{Info}(D_j) \) is the entropy of certain categories extracted from the sample set.

Step 3: Calculating information gain

The information gain is defined as the difference between the original information demand (that is based only on the class ratio) and the new demand (obtained after dividing \( A \)).

\[
\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)
\]

Step 4: Calculating attribute split information metrics

The information gain rate is equal to the information gain/intrinsic information, which will cause the importance of the attribute to decrease as the intrinsic information increases. This can be regarded as compensation for purely using information gain.

\[
\text{SplitInfo}_A(D) = -\sum_{j=1}^{n} \left| D_j \right| \left| D \right| \times \log_2 \text{Info}\left(\left| D_j \right| \left| D \right| \right)
\]

Step 5: Information gain rate
This value represents the information, which is generated by dividing the training data set $D$ into $n$ divisions corresponding to the $n$ outputs of the attribute $A$ test. Information gain rate definition is as follow:

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

(6)

In this work, precipitation observations from 83 meteorological stations were used for finding the similarity in the sub-regional precipitation with a similar temporal variation. The summer season was selected as the primary study period, which contributed 54% of the annual precipitation magnitude [54]. The K-Means algorithm was used to separate the regional precipitation based on the devised methodology into clusters, and then followed by the decision tree application. In the decision tree application, the climate indices during the spring season were used as a forecast tool to predict changes in the precipitation clusters of the basin with a time lag [55,56].

4. Results

4.1. Precipitation Clusters

Figure 3 shows the summer precipitation division derived with the application of K-Means clustering algorithm in the Poyang Lake basin. The current study used the clustering number $k$ as 2, 3, 4, 5 for the precipitation clustering of the Poyang Lake basin for a scientific division and convenience of practical application [57,58]. The closer the silhouette coefficient [59,60] is to 1, the better cluster will be developed. Comparatively, the $k = 2$, with silhouette coefficient of $\approx 0.5$ provides the relatively best clustering effect from rest of the clusters (Figure 3), therefore, we set the clustering number $k$ as 2, inferring two precipitation clusters in the Poyang Lake basin.

![Figure 3. Distribution of silhouette coefficient vs. clustering number of precipitation division in Poyang Lake basin based on K-Means algorithm.](image)

Figure 4 shows the two precipitation clusters boundaries of the PLB. From the results obtained (Figure 4), a north-south precipitation distribution can be seen, divided into two clusters. Based on the spatial pattern (Figure 4), the two clusters were expressed as cluster 1 (C1), and cluster 2 (C2), representing the northern basin and southern basin with distinct boundary obvious, respectively. The C1 covers 48 of 83 stations in the PLB, accounting for 58% of the regional station density, while Cluster2 covers 35 of 83, accounting for 42% of the station density. The region where C1(C2) was located referred to as region I (region II).
Changes in the annual precipitation in the whole basin and the two precipitation clusters over 1961–2016 in the Poyang Lake basin.

Figure 5 shows the interannual precipitation variability during the summer season averaged for the whole basin and both clusters (C1, and C2), respectively. The mean interannual variability of the whole basin average precipitation was shown in a dark black color, the mean of the C1 (green color) and C2 (brown color) was also shown. The results inferred similarities and differences in the mean interannual variation of the two clusters and basin-scale mean precipitation during the summer season. The statistical significance was assessed by the t-test and passing the significance at 0.01 level. The striking feature of the variability showed a more obvious change in the mean of the clusters from the whole basin mean precipitation in the recent decades, implying a possible change in the precipitation magnitude of C1 to be relatively higher than C2. Furthermore, significant differences between the two clusters were obvious during the extreme events during 1969, 1983, 1998, and 2011 implying a distinct response of each cluster from the mean. Further studies were suggested to explore these aspects in detail with emphasis on changing climate-induced increases in extreme events [61]. In conclusion, the K-Means approach application can obviously provide meaningful output in the atmospheric domain for better identification of precipitation patterns in a sub-regional domain.

Figure 6 shows the mean precipitation magnitude of C1 and C2 calculated from the interannual mean during 1961–2016 for each month of the monsoon season. The overall precipitation in C1 was higher than that in C2, with monthly scale differences obvious in each cluster (Figure 6a). The mean magnitude for C1(C2) during June was 306(248) mm, during July 160(130) and during August is 126(146), implying overall extra precipitation of 23.48% during June, followed by 23.07% during July in C1, whereas during August C1
had a $-14\%$ deficit relative to C2. In conclusion, apart from interannual differences, the K-Means clustering technique can further classify and show differences in sub-seasonal precipitation, implying its potential application in identifying sub-seasonal atmospheric variables clustering and differences. Indeed, the drivers of such regional-scale deviations were several factors ranging from local scale topography, convective activities to large-scale circulations and complex atmospheric modes that may further need to explore as well [62]. In the current work, exploring such drivers in relation to the techniques used may introduce further complexity in results interpretation and hence were excluded from being studied in a separate study. In the next section, the decision tree was used to predict changes in precipitation attributed to large-scale climate drivers.

![Figure 6](image_url)

**Figure 6.** Differences of the mean sub-seasonal precipitation in June, July, and August between two precipitation clusters in the Poyang Lake basin during 1961–2016. (a) the mean precipitation magnitude of C1 and C2. (b) Percentage in June, July, and August.

### 4.3. Experimental Data Pretreatment

In this section, the decision tree algorithm (C4.5) was used to predict changes in summer precipitation from the long-term precipitation observations. To do so, the training dataset required for the C4.5 algorithm accounted for 70% of the total number of samples, while the remaining 30% belonged to the test set. The training set was generally used to build the decision tree model, while the test set was applied to test the generalization and prediction capability of the model. The data for the period of 1961–2000 (40 years, 72% of the whole period) was selected as the training set of the model, and the data during 2001–2016 (16 years, accounting for 28%) was used as the test set. To follow the protocols, we defined the year with the standardized anomaly of the summer precipitation higher with standard deviation $\approx 0.5$ as the rainy year and lower as the normal year. The summer precipitation in the PLB here was abstracted to the binary classification of whether the summer precipitation was more than the normal or not. Using the criteria for defining a rainy year, region I and the region II both have 16 years, belonging to the rainy year during the summer season, as shown in Table 1.

| Table 1. Distribution of the rainy year of summer in different years of Poyang Lake. |
|-----------------------------------------------|
| **Separation of Poyang Lake** | **The Year with Heavy Rain in Summer** |
| Region I | 1969, 1970, 1973, 1977, 1980, 1983, 1993, 1994, 1995, 1997, 1998, 1999, 2010, 2011, 2014, 2015 |
| Region II | 1961, 1962, 1968, 1973, 1976, 1977, 1982, 1993, 1994, 1995, 1996, 1997, 1999, 2002, 2006, 2014 |
Furthermore, as the predictive factor, the climate indices from NCC during spring were used to predict whether the summer precipitation in the PLB will more than the normal or not. We obtained 130 climate indices data for the spring season through calculating the mean values of each climate index in March, April, and May, based on the hundreds of the climate system index from NCC, providing data support for the following establishment of summer precipitation prediction model.

4.4. Construction of the Model and Its Verification

Taking “whether the summer precipitation is more than the normal or not” as the object variable, the input variables of the model were the 130-climate signal indices during the preceding spring season of the corresponding summer season. After the pretreatment, the training data were input into the C4.5 algorithm for obtaining the decision tree (Figures 7 and 8). The main climate predictor of whether the summer precipitation was more than normal or not in region I (C1) was the location of the subtropical ridgeline location over North Africa.

![Figure 7. The classification tree model derived for cluster 1 of the Poyang Lake basin.](image1)

![Figure 8. The classification tree model derived for cluster 2 of the Poyang Lake basin.](image2)
The dominant factors about whether the summer precipitation was more than the normal or not in region II are the Atlantic-European Polar Vortex Area Index in the Atlantic and Europe. The closer of the node position to the leaf node, the significance of the node prediction variable will be less. The learning accuracy of the model in the region I was 90%. We verified the model after entering the decision tree through the test set. The test accuracy rate was up to 87.5%. The accuracy rate in region II was also up to 85.0%. After entering the decision tree through the test set for verifying the model, we found that the test accuracy rate was up to 93.8%.

It is clear that the rainy year forecast model based on decision tree C4.5 has certain common ability as well robustness in predicting the precipitation deviation with emphasis on large-scale atmospheric drivers as predictors. This model can help to provide a better comprehensible, concise, and valuable reference for the prediction about whether the summer precipitation is more than the normal or not. The decision-making tree has the advantage of concision, following the user’s logical judgment. Based on each embranchment of the decision-making tree, the rule of If ... then ... can be abstracted from root node to leaf node (T/F). As shown in Tables 2 and 3, all the above rules from the decision-making tree can be formed into the decision-making set of rules. Based on these rules, it will be convenient to use and seek the prediction in advance with a time-lag with the provision of early warning for extreme events and better awareness and preparedness. Simultaneously, there also exists learning accuracy for reference under each rule.

The following table illustrates the rules derived from the decision-making tree:

| Rule | Attributes (Factors and Levels) | Training Accuracy Rate |
|------|--------------------------------|------------------------|
| Rule A | If (North African Subtropical High Ridge Position Index > 15.473) Then less than normal precipitation | 10/12 = 83% |
| Rule B | If (North African Subtropical High Ridge Position Index ≤ 15.473 and West Pacific 850 mb Trade Wind Index ≤ 0.398) Then more than normal precipitation | 5/5 = 100% |
| Rule C | If (North African Subtropical High Ridge Position Index ≤ 15.473 and West Pacific 850 mb Trade Wind Index > 0.398 and India-Burma Trough Intensity Index ≤ 117.919) Then less than normal precipitation | 17/18 = 94% |
| Rule D | If (North African Subtropical High Ridge Position Index ≤ 15.473 and West Pacific 850 mb Trade Wind Index > 0.398 and India-Burma Trough Intensity Index > 0.389) Then more precipitation | 4/5 = 80% |

For a better description regarding the spatial distribution of the summer precipitation in the PLB under different climate scenarios of the preceding spring season. The spatial distribution of the precipitation anomaly (Figure 9) was derived using the rules for stating the factorial conditions associated with above-normal precipitation from those associated with below-normal precipitation from the decision-making tree. The regional precipitation differences owing to the set of rules derived stating whether the summer precipitation was higher than normal or not in different regions of the Poyang Lake (namely the rule B and rule D in the region I, and the rule B and rule D in the region II) in Figure 9. When
the early spring climate indices were in accord with the rule B of the decision-making tree in the region I, the precipitation was higher than the normal in region I, especially in the southern and eastern reaches of region I (\(-\sim 300\) mm). In the northern parts of region I, the precipitation in some parts was reduced by \(-\sim 50\) mm from the normal level (Figure 9a). When the early spring climate indices were in accord with the rule D of the decision-making tree in the region I, the precipitation in the whole region I was higher than the normal level by \(-\sim 500\) mm (Figure 9b), especially in the mid-east parts of the region I (higher than 300 mm).

**Table 3.** Rule sets of summer rainfall forecast model of “Whether less (lower than 0.5 times standard deviation)” in cluster2 of Poyang Lake basin.

| Rules          | Attributes (Factors and Levels)                                                                 | Training Accuracy Rate |
|----------------|---------------------------------------------------------------------------------------------|------------------------|
| Rule A         | If (Atlantic-European Polar Vortex Area Index \(\leq 13.968\)) Then less than normal precipitation | 5/5 = 100%             |
| Rule B         | If (Atlantic-European Polar Vortex Area Index \(> 13.968\) and Total Sunspot Number Index \(> 573.00\)) Then more than normal precipitation | 6/7 = 85.7%            |
| Rule C         | If (Atlantic-European Polar Vortex Area Index \(> 13.968\) and Total Sunspot Number Index \(\leq 573.00\)) and Pacific Polar Vortex Intensity Index \(> 3902.41\) Then less than normal precipitation | 13/14 = 92.9%          |
| Rule D         | If (Atlantic-European Polar Vortex Area Index \(> 13.968\) And Total Sunspot Number Index \(\leq 573.00\)) and Pacific Polar Vortex Intensity Index \(\leq 3902.41\) and Cold-tongue ENSO Index \(\leq -0.149\) Then more than normal precipitation | 4/6 = 66.7%            |
| Rule E         | If (Atlantic-European Polar Vortex Area Index \(> 13.968\) and Total Sunspot Number Index \(\leq 573.00\)) and Pacific Polar Vortex Intensity Index \(\leq 3902.41\) and Cold-tongue ENSO Index \(> -0.149\) Then less precipitation | 6/8 = 75%              |

**Figure 9.** Summer precipitation anomaly in the Poyang Lake basin (a). Summer precipitation anomaly under rule B in region I (b). Summer precipitation anomaly under rule D in region I (c). Summer precipitation anomaly under rule B in region II (d). Summer precipitation anomaly under rule D in the region II.
When the climatic indices during the early spring were in accord with the rule B of the decision-making tree in region II, most regions of region II were experiencing above-normal precipitation ~50 mm, especially in the eastern parts with the increased precipitation up to 100 mm (Figure 9c). When the climatic indices in the early spring were in accord with the rule D of the decision-making tree in region II, the precipitation in the whole region II was increased by 100 mm than the normal level (Figure 9d). In some western regions of region II, the precipitation was increased by ~250 mm in comparison with the normal level.

5. Discussion

The summer precipitation of the Poyang Lake basin, if influenced by several factors and patterns. The climate within the basin is also subjected to differences in mean precipitation magnitude, and deviations in seasonality under the same weather system, which are large. Thus, an accurate prediction of the summer precipitation in this lake basin has important and practical significant implications at a local scale. In this study, we used the K-Means clustering algorithm of the machine learning technique to reasonably and objectively separate the summer precipitation in the Poyang Lake basin into two clusters. Then, we built the C4.5 based algorithm decision tree prediction model for investigating whether the summer precipitation in the Poyang Lake was more than normal or not, with an insight into the possible large-scale climate drivers.

The K-Means clustering technique is among the commonly used partition clustering algorithm with simplicity and efficiency. It has become the most widely used among all clustering algorithms. The C4.5 can quickly and effectively discover potential drivers and key information from a large number of complex climate indexes and combine these drivers and information to select reasonable factors to construct a prediction model for summer precipitation in the Poyang Lake Basin. The results generally show a significant difference in summer precipitation between the southern and northern parts of the PLB. The K-Means clustering and C4.5 attribution generally agree with previous studies, which studied the regional precipitation variation due to large-scale atmospheric drivers embedded within complex earth system climate. Such forcing can range from the atmospheric response towards the oceanic forcing, solar radiation, and much more. The accuracy of model forecasting can reach a considerable level [63,64]. Min et al. [65] found that the spring SST in the South China Sea, the Bay of Bengal, and the Arabian Sea was positively correlated with the summer precipitation in the Yangtze River Basin. Lu et al. [66] proposed that the western Pacific subtropical high and the subtropical monsoon had strong (weak) linkage, whereas the West Indian Ocean circulation was negatively (positive) related to the summer rainfall in the Yangtze-Huai River. Gong et al. [67] pointed out that the Arctic Oscillation (AO) index was negatively correlated with Meiyyu. Wang et al. [68] found that the North Atlantic Oscillation (NAO) in the previous winter had little effect on summer precipitation in my country, while the changes in the North Atlantic Oscillation (NAO) in the previous spring had a significant correlation with summer precipitation. All these studies individually reported the linkage between the large-scale circulations individually or forced by changes in sea surface temperature and their association with regional precipitation changes. All these studies individually reported the linkage between the large-scale circulations individually or forced by changes in sea surface temperature and their association of with regional precipitation changes. The current study attribution was indeed an initial appraisal using the machine learning algorithms and thus a more detailed study was scheduled to look into the dynamics of the indices as a forecast indicator of the above/below normal precipitation in the region.

With the continuous advances of the big data era, computing hardware and computational intelligence have been continuously strengthened. The data mining technique has been widely used to predict the short-term changes in weather and climate of summer precipitation regimes. In this study, we used the machine learning technique to separate the Poyang Lake basin precipitation into clusters and find the associated set of rules responsible for such changes in precipitation magnitude. We built an effective prediction
model for investigating whether the summer precipitation is more than the normal or not in different regions, which offers a significant reference for the short-term climate forecast of the summer precipitation in the Poyang Lake basin. However, machine learning demands a larger number of data samples and a higher demand for the computation speed of the computing devices in comparison with traditional mathematical statistics, which need more complex training strategies. We suggest that there still exists a potential in the prediction accuracy, with the continually accumulated data sample and the continually optimized training strategies and parameters. In the next step, we plan to focus on calculating the effects of the dominated climate factors that have been identified in this study on the summer precipitation in regions I (C1) and II (C2) of the Poyang Lake based on the sensitivity experiments.

6. Conclusions

The study used long-term precipitation observations from a dense network of rain-gauges in a diverse climate during 1961–2016 to classify and attribute large-scale atmospheric drivers to regional precipitation variability. Using machine learning techniques, the study found significant variation in regional precipitation variation across the Poyang Lake basin (PLB), which, however, is termed to be a homogenous precipitation region. The choice of the techniques used to verify the sub-regional clusters based on precipitation magnitude variability on an interannual scale, monthly scale, were justified by attribution from larger-scale indices as a predictor of the changes in the mean precipitation. The specific conclusions are as follows:

1. Based on the K-Means algorithm, we investigated the level of similarity in the summer precipitation within the Poyang Lake basin. Based on the principle of “similarity within classes, dissimilarity between classes,” the Poyang Lake basin precipitation is separated into clusters, comprised of north and south regions, namely region I and region II. These two regions are integrated, continuous, and mutually independent, meeting the objective and reasonable criterion for the separation with a distinct boundary obvious.

2. Comparing region I and region II as Cluster 1 (C1) and Cluster 2 (C2), the changes in the summer precipitation of these two regions have their respective individual characteristics on an interannual and monthly scale. On the interannual scale, the summer precipitation always exhibited significant differences between the two clusters. On the monthly scale, the amount of precipitation in June and July in region I is higher than region II, while that in August is smaller than region II (include the fraction values).

3. The C4.5 based decision tree prediction model is built to investigate whether the summer precipitation in the C1 and C2 is more than the normal or not under specific years with 0.5 standard deviations from the mean. The learning accuracy of the model in the C1 and the C2 are up to 90.0% and 85%, respectively. After checking the model by test set, the accuracy rate of the test is up to 87.5% and 93.8%, respectively. From region I to region II, the root node and the leaf node from the decision tree abstract out four and five decision rules, respectively, forming into the concise and scientific rule sets. Each rule has its own learning accuracy, which would be convenient for application use.

4. From region I to region II, based on the decision tree models, the main climate factors of the summer precipitation in region I are the modifications of the subtropical ridgeline over North Africa and the 850 hPa western pacific trade wind. The dominant factors about whether the summer precipitation is more than the normal or not in region II are the Atlantic-European Polar Vortex Area Index in the Atlantic and Europe and the changes in the number of sunspots.

The study thus concludes the potential of the machine learning techniques application valuable in atmospheric sciences for better decision making and feature attribution. Further studies should include a large-scale dynamics-based verification of the techniques.
with detailed methods and datasets for a better depiction of the ability of the machine learning techniques.

**Author Contributions:** Conceptualization, D.L. and D.S.; methodology, D.L. and D.S.; software, M.Y. and Y.C. (Yutian Chen); validation, M.Y. and Y.C. (Yuanfang Chai) and D.S.; formal analysis, D.L. and D.S.; investigation, D.L., M.Y., and D.S.; resources, D.L., M.Y.; data curation, M.Y. and Y.C. (Yutian Chen); writing—original draft preparation, D.L. and D.S.; writing—review and editing, W.U. and Y.C. (Yuanfang Chai); visualization, W.U. and Y.C. (Yuanfang Chai); supervision, G.W.; project administration, D.S. and G.W.; funding acquisition, D.S. and G.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Key R&D Program of China, grant number 2019YFC1510203, Huaihe Basin Meteorological Open Founding, grant number HRM201602 and China Scholarship Council.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Zhu, H.; Xu, L.; Jiang, J.; Fan, H. Spatiotemporal Variations of Summer Precipitation and Their Correlations with the East Asian Summer Monsoon in the Poyang Lake Basin, China. *Water* 2019, 11, 1705. [CrossRef]
2. Zhang, Q.; Xiao, M.; Li, J.; Singh, V.P.; Wang, Z. Topography-based spatial patterns of precipitation extremes in the Poyang Lake basin, China: Changing properties and causes. *J. Hydrol.* 2014, 512, 229–239. [CrossRef]
3. Zhang, Y.; You, Q.; Ye, L.; Chen, C. Spatio-temporal characteristics and possible mechanisms of rainy season precipitation in Poyang Lake Basin, China. *Clim. Res.* 2017, 72, 129–140. [CrossRef]
4. Li, X.; Zhang, Q.; Xu, C.Y. Assessing the performance of satellite-based precipitation products and its dependence on topography over Poyang Lake basin. *Theor. Appl. Climatol.* 2014, 115, 713–729. [CrossRef]
5. Han, T.; Li, S.; Hao, X.; Xinyi, G. A statistical prediction model for summer extreme precipitation days over the northern Central China. *Int. J. Climatol.* 2019, 40, 4189–4202. [CrossRef]
6. Lee, E.; Hong, S.-Y. Impact of the Sea Surface Salinity on Simulated Precipitation in a Global Numerical Weather Prediction Model. *J. Geophys. Res. Atmos.* 2019, 124, 719–730. [CrossRef]
7. Johny, K.; Pai, M.L. Empirical forecasting and Indian Ocean dipole teleconnections of south—West monsoon rainfall in Kerala. *Meteorol. Atmos. Phys.* 2019, 131, 1055–1065. [CrossRef]
8. Sachindra, D.A.; Ahmed, K.; Rashid, M.M.; Shahid, S.; Perera, B.J.C. Statistical downscaling of precipitation using machine learning techniques. *Atmos. Res.* 2018, 212, 240–258. [CrossRef]
9. Whan, K.; Schmeits, M. Comparing Area Probability Forecasts of (Extreme) Local Precipitation Using Parametric and Machine Learning Statistical Postprocessing Methods. *Mon. Weather Rev.* 2018, 146, 3651–3673. [CrossRef]
10. Rahnama, A.; Clark, S.; Sridhar, S. Machine learning for predicting occurrence of interphase precipitation in HSLA steels. *Comput. Mater. Sci.* 2018, 154, 169–177. [CrossRef]
11. Alexander, L. V Global observed long-term changes in temperature and precipitation extremes: A review of progress and limitations in IPCC assessments and beyond. *Weather Clim. Extrem.* 2016, 11, 4–16. [CrossRef]
12. Nabeel, A.; Athar, H. Stochastic projection of precipitation and wet and dry spells over Pakistan using IPCC AR5 based AOGCMs. *Atmos. Res.* 2020, 234, 104742. [CrossRef]
13. Tapiador, F.J.; Navarro, A.; Levizzani, V.; Garcia-Ortega, E.; Huffman, G.J.; Kidd, C.; Kucera, P.A.; Kummerow, C.D.; Masunaga, H.; Petersen, W.A.; et al. Global precipitation measurements for validating climate models. *Atmos. Res.* 2017, 197, 1–20. [CrossRef]
14. Eden, J.; Van Oldenborgh, G.J.; Hawkins, E. A global empirical system for probabilistic seasonal climate prediction. *Geosci. Model Dev. Discuss.* 2015, 8, 3941–3970. [CrossRef]
15. Totz, S.; Tziperman, E.; Coumou, D.; Pfeiffer, K.; Cohen, J. Winter Precipitation Forecast in the European and Mediterranean Regions Using Cluster Analysis. *Geophys. Res. Lett.* 2017, 44, 12–412. [CrossRef]
16. Gupta, S.; Agrawal, A.; Gopalakrishnan, K.; Narayanan, P. Deep Learning with Limited Numerical Precision. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 37, pp. 1737–1746.
17. Singh, S.; Kaushik, M.; Gupta, A.; Malviya, A. Weather Forecasting using Machine Learning Techniques. In Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019, Kamla Nehru Institute of Technology Sultanpur, UP, India, 2019. [CrossRef]
18. O’Gorman, P.A.; Dwyer, J.G. Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. J. Adv. Model. Earth Syst. 2018, 10, 2548–2563. [CrossRef]

19. Ardabili, S.; Mosavi, A.; Dehghani, M.; Varkonyi-Koczy, A. Deep Learning and Machine Learning in Hydrological Processes, Climate Change and Earth Systems: A Systematic Review. In International Conference on Global Research and Education; Springer: Cham, Switzerland, 2019.

20. Hartigan, J.; MacNamara, S.; Leslie, L. Application of Machine Learning to Attribution and Prediction of Seasonal Precipitation and Temperature Trends in Canberra, Australia. Climate 2020, 8, 76. [CrossRef]

21. Chivers, B.D.; Wallbank, J.; Cole, S.J.; Sebek, O.; Stanley, S.; Fry, M.; Leontidis, G. Imputation of missing sub-hourly precipitation data in a large sensor network. A machine learning approach. J. Hydrol. 2020, 588, 125126. [CrossRef]

22. Teegavarapu, R.S.V.; Aly, A.; Patahk, C.S.; Ahlquist, J.; Fuelberg, H.; Hood, J. Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: Use of optimal weighting parameters and nearest neighbour-based corrections. Int. J. Clim. 2018, 10, 776–793. [CrossRef]

23. Miao, Q.; Pan, B.; Wang, H.; Hsu, K.; Sorooshian, S. Improving Monsoon Precipitation Prediction Using Combined Convolutional and Long Short Term Memory Neural Network. Water 2019, 11, 977. [CrossRef]

24. Huang, T.; Xu, L.; Fan, H. Drought Characteristics and Its Response to the Global Climate Variability in the Yangtze River Basin, China. Water 2019, 11, 13. [CrossRef]

25. Xiao, L.; Tian, W.; Lv, L. Temporal and spatial change characteristics of precipitation concentration index in Poyang Lake Basin. J. Nanchang Inst. Technol. 2020, 39, 25–31.

26. Shi, N. Meteorological Statistical Forecast; China Meteorological Press: Beijing, China, 2009; pp. 128–142. (In Chinese)

27. Wei, F. Regional consensus forecast method with dynamic weighting for summer precipitation over China. Q. J. Appl. Meteorol. 1999, 10, 402–409. (In Chinese)

28. Ding, Y.; Li, W.; Li, Q. Advance in seasonal dynamical prediction operation in China. Acta Meteorol. Sin. 2004, 62, 598–612. (In Chinese).

29. Haiyang, H.; Yijia, H.; Zhong, Z.; Yimin, Z. Double nested dynamical downscaling research on summer precipitation over China with WRF model. J. Meteorol. Sci. 2015, 35, 413–421. (In Chinese)

30. Zhang, W.; Leung, Y.; Chan, J.C.L. The Analysis of Tropical Cyclone Tracks in the Western North Pacific through Data Mining. Double nested dynamical downscaling research on summer precipitation over China with WRF model. J. Meteorol. Sci. 2015, 35, 413–421. (In Chinese)

31. Bhatia, N.; Sojan, J.M.; Simonovic, S.; Srivastav, R. Role of cluster validity indices in delineation of precipitation regions. Water 2020, 12, 1372. [CrossRef]

32. Pike, M.; Lintner, B.R. Application of clustering algorithms to TRMM precipitation over the tropical and South Pacific Ocean. J. Clim. 2020, 33, 5767–5785. [CrossRef]

33. Salzberg, S. C4.5: Programs for Machine Learning. Mach. Learn. 1994, 16, 235–240. [CrossRef]

34. Hamerly, G.; Elkan, C. Alternatives to the k-means algorithm that find better clusterings. In Proceedings of the International Conference on Information and Knowledge Management, McLean, VA, USA, 4–9 November 2002; pp. 600–607. [CrossRef]

35. Silhouettes, R. A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 1987, 20, 53–65.

36. Liao, K.; Zhou, Z.; Lai, X.; Zhu, Q.; Feng, H. Evaluation of different approaches for identifying optimal sites to predict mean hillslope soil moisture content. J. Hydrol. 2017, 547, 10–20. [CrossRef]

37. Friedl, M.A.; Sulla-Menashe, D.; Tan, B.; Schneider, A.; Ramankutty, N.; Sibley, A.; Huang, X. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sens. Environ. 2010, 114, 168–182. [CrossRef]

38. Yang, W.; You, Q.; Fang, N.; Xu, L.; Zhou, Y.; Wu, N.; Ni, C.; Liu, Y.; Liu, G.; Yang, T.; et al. Assessment of wetland health status of the climate change of temperatures and precipitation. Phys. Chem. Earth 2016, 94, 22–28. [CrossRef]
48. Zhang, W.; Fu, B.; Peng, M.S.; Li, T. Discriminating developing versus nondeveloping tropical disturbances in the Western North Pacific through decision tree analysis. *Weather Forecast.* 2015, 30, 446–454. [CrossRef]

49. Kim, J.M.; Ahn, H.K.; Lee, D.H. A study on the occurrence of crimes due to climate changes using decision tree. In *Lecture Notes in Electrical Engineering*; Springer: Dordrecht, The Netherlands, 2013; Volume 215 LNEE, pp. 1027–1036. [CrossRef]

50. Hasan, N.; Uddin, T.; Chowdhury, N.K. Automated weather event analysis with machine learning. In Proceedings of the 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET), Dhaka, Bangladesh, 28–29 October 2016. [CrossRef]

51. Veenadhari, S.; Misra, B.; Singh, C.D. Machine learning approach for forecasting crop yield based on climatic parameters. In Proceedings of the 2014 International Conference on Computer Communication and Informatics, Combatore, India, 3–5 January 2014. [CrossRef]

52. Coria, S.R.; Gay-García, C.; Villers-Ruíz, L.; Guzmán-Arenas, A.; Ávila-Barrón, O.R.; Pérez-Meza, M.; Cruz-Núñez, X.; Martínez-Luna, G.L. Climate patterns of political division units obtained using automatic classification trees. *Atmosfera* 2016, 29, 359–377. [CrossRef]

53. Zhang, W.; Gao, S.; Chen, B.; Cao, K. The application of decision tree to intensity change classification of tropical cyclones in western North Pacific. *Geophys. Res. Lett.* 2013, 40, 1883–1887. [CrossRef]

54. Guo, H.; Jiang, T.; Wang, G.; Su, B.; Wang, Y. Observed trends and jumps of climate change over Lake Poyang Basin, China: 1961-2003. *J. Lake Sci.* 2006, 18, 443–451. (In Chinese)

55. Miao, C.; Dongpo, H.E.; Wang, J.; Shi, D. Research and application of summer rainfall prediction model in the middle and lower reaches of the Yangtze River based on C4.5 algorithm. *J. Meteorol. Sci.* 2017, 37, 256–264. [CrossRef]

56. Zhang, J.; Yao, Y.; Cao, N. Prediction of whether precipitation based on decision tree. *J. Geomat.* 2017, 42, 107–109. [CrossRef]

57. Treshansky, A. Overview of clustering algorithms. *Proc. SPIE* 2001, 4367, 41–51. [CrossRef]

58. Clausi, D.A. K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation. *Pattern Recognit.* 2002, 35, 1959–1972. [CrossRef]

59. Anitha, P.; Patil, M.M. RFM model for customer purchase behavior using K-Means algorithm. *J. King Saud. Univ. Comput. Inf. Sci.* 2019. [CrossRef]

60. Jujjuri, R.D.; Venkateswara Rao, M. Evaluation of enhanced subspace clustering validity using silhouette coefficient internal measure. *J. Adv. Res. Dyn. Control Syst.* 2019, 11, 321–328.

61. Li, X.; Hu, Q. Spatiotemporal Changes in Extreme Precipitation and Its Dependence on Topography over the Poyang Lake Basin, China. *Adv. Meteorol.* 2019, 1, 1–15. [CrossRef]

62. Prein, A.F.; Langhans, W.; Fosser, G.; Ferrone, A.; Ban, N.; Goergen, K.; Keller, M.; Tölle, M.; Gutjahr, O.; Feser, F.; et al. A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.* 2015, 53, 323–361. [CrossRef] [PubMed]

63. Liu, Y.; Ke, Z.; Ding, Y. Predictability of East Asian summer monsoon in seasonal climate forecast models. *Int. J. Climatol.* 2019, 39, 5688–5701. [CrossRef]

64. Mulholland, D.P.; Haines, K.; Sparrow, S.N.; Wallom, D. Climate model forecast biases assessed with a perturbed physics ensemble. *Clim. Dyn.* 2017, 49, 1729–1746. [CrossRef]

65. Min, J.; Guo, Y.; Wang, G. Impacts of Soil Moisture on Typical Frontal Rainstorm in Yangtze River Basin. *Atmosphere* 2016, 7, 42. [CrossRef]

66. Ming, L.U.; Tan, G.; Chen, H.; Hang, Y.; Chen, Q. The relationship between summer rainfall anomalies in Yangtze-Huaihe valley and atmospheric circulation anomalies over western Indian Ocean. *J. Meteorol. Sci.* 2013, 27, 992–1006.

67. Gong, D. Arctic Oscillation’s Significance for Prediction of East Asian Summer Monsoon Rainfall. *Meteorol. Mon.* 2003, 29, 3–6.

68. Wang, Y.; Shi, N. The North Atlantic Oscillation In Relation To Summer Weather-Climate Anomaly In China And East Asian Summer Monsoon. *Sci. Meteorol. Sin.* 2001, 21, 271–278.