https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

Hydrological drought prediction and its influencing factors analysis based on a machine learning model

Min Li[1], Yuhang Yao[1], Zilong Feng[1], Ming Ou[1]

[1]College of Hydraulic Science and Engineering, Yangzhou University, Yangzhou, China, 225000

Correspondence to: Min Li (limintju@126.com)

Abstract: Predicting future drought conditions is crucial for effective disaster management. In this study, a machine learning framework is proposed to predict hydrological drought in the Huaihe River Basin, China. The interpretable Extreme Gradient Boosting (XGBoost) model is applied to forecast four drought categories in 28 grid regions, using 26 factors for monthly and 18 for seasonal predictions. The framework also integrates the Shapley Additive Explanation (SHAP) variable importance index to infer drought prediction factors. The model achieves 79.9% accuracy in classifying droughts, with the Standard Precipitation Index (SPI) being the most influential factor. The SHAP values of SPI are 0.360, 0.261, 0.169, and 0.247 for spring, summer, autumn, and winter, respectively. Soil moisture content and evapotranspiration are particularly affected in spring and autumn, while large-scale climatic factors are more significant in summer and winter. Overall, this study offers valuable decision support for regional drought management and water resource allocation.

Keywords: XGBoost; SHAP; Drought prediction; SRI; Huaihe River Basin

1 Introduction

Drought is a global disaster characterized by its long duration and extensive impacts, resulting in severe implications for the economy, agriculture, and environment (Fu et al., 2018; Shi et al., 2018; Zhou et al., 2020; 2021). Over the past 20 years, the frequency and severity of global drought events have increased (Dai 2011; 2012; 2013; Zhang et al., 2019), affecting water security, economic growth, and food supply in some areas. Therefore, drought prediction is of great significance for managing water resources and reducing losses caused by drought.

Consequently, according to the different effects of drought, previous studies have divided it into several different types. Among them, four types of droughts are widely used: meteorology, hydrology, agriculture, and social economy (Wilhite and Glantz, 1985; American Meteorological Society, 2013). In

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

the past few decades, more than one hundred drought indices based on single or multiple hydroclimatic variables have been proposed to represent different drought characteristics. For example, the Palmer Drought Severity Index (PDSI) (Palmer 1965), the Standardized Precipitation Index (SPI) (McKee et al., 1993), and the Standardized Runoff Index (SRI) (Shukla and Wood, 2008). SPI index and SRI index are robust, statistically straightforward to compute, and well-suited to long-term time series data. Therefore, this study chooses the SPI index and SRI index to characterize meteorological drought and hydrological drought.

In recent years, there has been an increasing trend toward utilizing machine learning to predict droughts (Ardabili et al., 2020; Sun and Scanlon, 2019). Compared to conventional regression models, machine learning-based models better capture non-linear characteristics inherent in drought problems and exhibit more robustness, especially when dealing with high-dimensional datasets (Mishra and Singh, 2010; Kikon and Deka, 2022; Prodhan et al., 2022; Wu et al., 2022). Multiple machine learning models such as artificial neural networks (Orimoloye et al., 2021; Orimoloye et al., 2022), support vector machines (Li et al., 2021), random forests(Park et al., 2019), and extreme gradient boosting (XGBoost) (Choi et al. 2018; Han et al. 2019; Zhang et al., 2023) have been extensively employed in the research field of drought. Machine Learning models can learn the input-output relationships in training data and can effectively leverage big data to improve prediction accuracy (Mardian et al., 2023). By training tree-based machine learning models, Bachmair et al. (2016) discovered that tree-based machine learning models outperform baseline models. Jungho and Kim (2023) employed a tree-structured XGBoost model to predict the likelihood of impact occurrence (LIO) of drought on public water supply. Their findings demonstrated that the XGBoost model exhibited high accuracy and low uncertainty. Furthermore, the XGBoost model necessitates only minor hyperparameter tuning, and its performance is relatively insensitive to the selection of hyperparameters (Gao and Ding, 2020; Barnwal et al., 2022).

Previous research indicates that numerous factors significantly impact hydrological drought. Zou et al. (2018) demonstrated that climate change is the primary factor affecting hydrological drought on long-term scales. Wang et al. (2021) found that climatic variables such as precipitation and evapotranspiration significantly influence the duration of hydrological drought. Additionally, Gan et al. (2023) revealed that large-scale climatic factors and sunspot activity have a substantial impact on hydrological drought events in the Huaihe River Basin. Despite many studies showing that machine learning models outperform physical models in terms of prediction accuracy, these models lack transparency and interpretability.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

Most research on machine learning models for drought prediction focuses on model performance, often neglecting the role of different factors influencing drought occurrence in model predictions. For example, Xu et al. (2022) established a hybrid model combining autoregressive integrated moving averages (ARIMA) and long short-term memory (LSTM) to predict the standardized precipitation evapotranspiration index at multiple time scales. Yu et al. (2023) combined the Hydrologiska Byrans Vattenbalansavdelning (HBV) model with an LSTM neural network to improve the prediction ability for semi-arid basins. Yalcin et al. (2023) proposed a hybrid model of convolutional neural networks (CNN) and LSTM to enhance the prediction accuracy of the standardized precipitation evapotranspiration index. However, these studies do not consider the influence of different factors on the model output.

Recent advancements in Explainable AI (XAI) techniques have provided opportunities for understanding why models make certain predictions (Gunning et al., 2019; Islam et al., 2022). Recently, local interpretability methods have been developed and can be implemented for neural network and random forest model architectures (Ribeiro et al., 2016a). The Local Interpretable Model-Agnostic Explanation (LIME) method has been widely used, but it exhibits a high degree of instability due to considerable variation in its explanations upon repeated use (Ribeiro et al., 2016b). Therefore, the Shapley Additive Explanations (SHAP) approach was proposed as a solution. Grounded in the strong theoretical basis of game theory, it provides more robust mathematical accuracy and consistent extension on top of the LIME framework (Lundberg and Lee, 2017; Molnar, 2022). At present, there are few studies on interpretable machine learning using the SHAP algorithm. For example, Dikshit and Pradhan (2021) employed an LSTM model combined with the SHAP algorithm to predict droughts, demonstrating that the inclusion of climate variables as predictors can enhance prediction accuracy. Similarly, Mardian et al. (2023) utilized an XGBoost model and SHAP to forecast droughts in the Canadian prairies, and clarified the importance of spatial and temporal predictors, drought indicators, GRACE groundwater distribution and teleconnection in drought prediction. However, the range of drought-influencing factors considered in their research is still not comprehensive enough. For example, soil temperature and water content, surface thermal radiation and other factors are also important factors affecting drought (Raposo et al., 2023).

In light of the above, the novelty of this study is to employ interpretable machine learning models for hydrological drought prediction and to identify the contribution of different influencing factors to the model prediction results. While SPI is a precursor to SRI, this study disentangles the hierarchy of

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

contributing factors, including SPI, large-scale climate indices, and soil moisture. Soil moisture directly affects hydrological drought, and it can analyze the contribution of different factors to drought when it is predicted together with drought factors such as large-scale climate factors. For example, Mardian et al. (2023) employed a method combining the XGBoost model with SHAP (Shapley Additive Explanations) values, utilizing a variety of drought influencing factors such as large-scale climatic factors and soil moisture, to predict drought conditions in the context of the Canadian Drought Monitor (CDM) and to understand the underlying driving factors. Therefore, the objectives of the study are: ⅰ) Utilizing the XGBoost model, combined with 26 factors predicted monthly and 18 factors predicted seasonally, the hydrological drought in the Huaihe River Basin is predicted, and the performance evaluation is carried out by using precision and recall indicators; ii) Various SHAP plots were employed to gain insights into the model outputs and analyze the influence of different drought variables on the predictive results of the model.

2 Study area and data

2.1 Study area

In this paper, as shown in Figure 1, the Huaihe River Basin is selected as the research area, and the grid is divided at a resolution of 1°lat×1°lon, with a total of 28 grid regions, which takes into account the computational feasibility and spatial heterogeneity. Although large-cale climatic factors have spatial consistency, their effects on regional precipitation can be different through local terrain-atmosphere feedback (Lu et al., 2006). Gridded analysis identifies sensitive subregions, supporting targeted mitigation. The Huaihe River Basin is located at 111°55'–121°25'E, 30°55'–36°36'N, covering an area of approximately 270,000 square kilometers. It experiences significant spatiotemporal variations in precipitation, with an average annual precipitation of around 883 millimeters. Situated in the transitional climatic zone from south to north, the southern part of the basin falls under a subtropical climate, while the northern part experiences a warm temperate climate. The average annual temperature ranges from 11 to 16°C. The winter and spring seasons in the basin are relatively dry, while the autumn and summer seasons are hot and rainy, resulting in pronounced seasonal fluctuations between droughts and floods. The average annual runoff depth in the basin is 230 millimeters. Due to its unique geographical location, the area is prone to frequent flooding, leading to high water levels and prolonged flood conditions. In

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

addition, the annual average water surface evaporation in the Huaihe River Basin ranges from 900 to 1500 millimeters. As one of the important agricultural production bases in China, the basin is densely populated with substantial water demands. However, the region frequently suffers from drought disasters. Since the beginning of the 21st century, an average of 2.698 million hectares of crops, accounting for 21% of the total cultivated land area in the basin, have been affected annually.

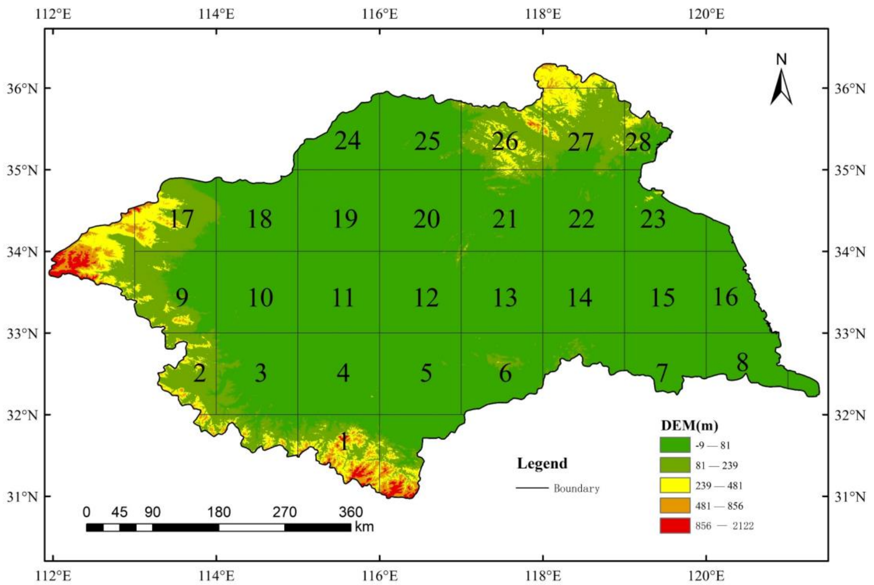

Figure 1: Huaihe River Basin and 28 grid area location.

2.2 Data

We obtained monthly average precipitation, wind speed, temperature, evapotranspiration, monthly average runoff, 0-10cm soil moisture, and 100-200cm soil moisture data sets for the Huaihe River Basin from the website https://disc.gsfc.nasa.gov/datasets/GLDAS_NOAH10_M_2.0/ for the period 1960 to 2014. The monthly average 2 m dewpoint temperature, surface net solar radiation, surface net thermal radiation, surface pressure, and leaf area index data sets were obtained from the ERA5-Land reanalysis dataset (https://cds.climate.copernicus.eu/). According to whether the grid center point falls within the basin, 28 grid regions are defined. If the center point of the grid is not within the basin boundary, the region is not divided into grids. The grid analysis is carried out with these grid points as the center and 1°lat×1°lon as the resolution, covering a total of 28 grid regions. Using the interpolation method in array, the data of Huaihe River Basin are interpolated to 28 grid regions.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

Numerous studies have demonstrated the significant influence of large-scale climate indices, including the Atlantic Multidecadal Oscillation (AMO), Arctic Oscillation (AO), North Pacific pattern (NP), Pacific Decadal Oscillation (PDO), and Nino3.4, on drought dynamics(Gan et al., 2023; Phan-Van et al., 2022; Wu and Xu, 2020; Xiao et al., 2019). For example, the positive phase of AMO leads to a decrease in summer precipitation in the Huaihe River Basin by enhancing the western Pacific subtropical high (Lu et al., 2006); the Pacific Decadal Oscillation (PDO) has the most significant impact on the monthly runoff in the Huaihe River Basin (Sun et al., 2018). These selected climate factors (Nino3.4, AMO, TPI, PDO, AO, TNI, and NP) for the Huaihe River basin analysis were acquired from the National Oceanic and Atmospheric Administration (NOAA) climate database (http://www.esrl.noaa.gov/psd/data/climateindices) , covering the period from 1960 to 2014.

3 Methods

3.1 Drought index

In this study, the standardized precipitation index (SPI) (McKee et al., 1993) is used to characterize meteorological drought. SPI is widely used for drought risk assessment and monitoring due to its ease of calculation and ability to work on multiple time scales. The calculation method of SPI is as follows:

$$f(x) = \frac{x^{\alpha-1} e^{\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \tag{1}$$

$$F(x) = \int_0^x f(x)dx \tag{2}$$

Assuming that the precipitation series x at a certain time scale follows a stationary gamma distribution, where α and β are the scale and shape parameters ($\alpha > 0$, $\beta > 0$). The cumulative probability $F(x)$ of each item is normalized to obtain the corresponding SPI.

The standardized runoff index (SRI) was first proposed by Shukla and Wood (2008) as an effective and accurate index for describing hydrological drought characteristics. It has been widely used in hydrological drought identification. SRI is also calculated by transforming the cumulative flow distribution of a given time scale into a standard normal distribution using equiprobability transformation, similar to the calculation method of SPI. The SPI/SRI classes are classified as shown in Table 1 (Li et al. 2024). In this study, drought is classified into four classes, namely, Normal (ND), Mild drought (D1),

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

Moderate drought (D2), and Severe drought and Extreme drought (D3), according to Table 1. However, due to the limited number of extreme drought events, it posed an issue in training the model. Therefore, the classes of Severe drought and Extreme drought were merged into one.

Table 1: Drought class classification and corresponding SPI values and SRI value.

SPI/SRI value	Class
> 0	Normal
0 to -1.0	Mild
-1.0 to -1.5	Moderate
-1.5 to -2.0	Severe
≤ -2.0	Extreme

3.2 Machine learning models

In this paper, the XGBoost model is used for multi-input single-output regression prediction problems to predict the hydrological drought in the Huaihe River Basin. The XGBoost model is an ensemble learning algorithm belonging to the Boosting algorithm category. It utilizes decision trees as its basic elements and implements a gradient-boosting algorithm to minimize loss when adding new models. XGBoost aims to improve the training speed and predictive performance of gradient-boosting decision trees. The foundational knowledge about the mechanism and implementation behind XGBoost can be found in the paper by Chen and Guestrin (2016). Assuming we have K base models denoted as $f_t(x) \in F \; t = 1,2,......,K$, where F the model space contains all the base models, the XGBoost model can be represented using the following function:

$$\hat{y} = F(x) = \sum_{t=1}^{k} f_t(x) \tag{3}$$

Where the parameters of the XGBoost model primarily consist of the structure of each tree and the scores in the leaf nodes, that is, the learning of each function $f_t(x)$.

As each base model is generated in a certain sequential order, the creation of the subsequent tree takes into account the predictions made by the preceding tree. Therefore, the objective function of the t base model can be expressed as follows:

$$y^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

Here, $l\left(y_i, \hat{y}_i^{(t-1)}\right)$ represents the loss function related to $y_i, \hat{y}_i^{(t-1)}$, $y_i^{(t-1)}$ denotes the predictions of the first $t-1$ decision trees for sample i (i.e., the sum of predictions made by the first $t-1$ trees), y_i represents the actual value of sample i, $f_t(x_i)$ represents the prediction of the t decision tree for sample i, and $\Omega(f_t)$ represents the model complexity of the t tree. Therefore, the predictions of the first k trees for the sample i are equal to the predictions of the first $k-1$ trees plus the prediction of the k tree.

3.3 Model input data

The XGBoost model for 28 grid areas is established, and the data types used in each region are the same. As shown in Table 1, for the monthly data analysis, 26 different drought-influencing factors were considered. These include a month-scale SPI (SPI-1) and SPI indices at different time scales of 1 month and 2 months in advance. Large-scale climate indices (AMO, TPI, PDO, AO, TNI, NP), evapotranspiration, wind speed, 2 m dewpoint temperature, soil moisture content, surface net thermal radiation, surface net solar radiation, surface pressure and leaf area index were considered.

As shown in Table 2, for seasonal data analysis, the basin data are classified by season, and 18 different drought influencing factors are used. It includes SPI-3 value, soil moisture content, evapotranspiration, surface net thermal radiation, air temperature, NINO3.4, NP, wind speed, TNI, PDO, TPI, surface pressure, AO, AMO, leaf area index, 2 m dewpoint temperature and surface net solar radiation in four seasons.

For monthly and seasonal data sets, SHAP (Shapley Additive Explanation) values were used to analyze the contribution of 28 grid regions to determine the impact of each factor.

Monthly-scale predictions capture the rapid onset of drought, which is critical for early warning systems, whereas seasonal analysis aligns with agricultural planning cycles. Thus, our study employs both monthly and seasonal analyses to comprehensively assess short-term variability and long-term trends in hydrological drought.

Table 2: The drought impact factors of the monthly scale prediction input of the model (T is the lead time, SPI-1, SPI-3, SPI-6, and SPI-9 are SPI values at different monthly scales.).

Drought influencing factors（monthly）

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

1	SPI-1
2	T=1 SPI-1
3	T=1 SPI-3
4	T=1 SPI-6
5	T=1 SPI-9
6	T=2 SPI-1
7	T=2 SPI-3
8	T=2 SPI-6
9	T=2 SPI-9
10	d2m temperature
11	surface pressure
12	evapotranspiration
13	Air temperature
14	wind speed
15	surface net solar radiation
16	surface net thermal radiation
17	0-10cm soil moisture
18	100-200cm soil moisture
19	Nino3.4

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

20	AMO
21	PDO
22	AO
23	TNI
24	NP
25	TPI
26	leaf area index

Table 3: The drought impact factors of seasonal prediction input of model.

	Drought influencing factors（seasonal）
1	SPI-3 (different seasons)
2	d2m temperature
3	surface pressure
4	evapotranspiration
5	Air temperature
6	wind speed
7	surface net solar radiation
8	surface net thermal radiation
9	0-10cm soil moisture
10	100-200cm soil moisture

11	Nino3.4
12	AMO
13	PDO
14	AO
15	TNI
16	NP
17	TPI
18	leaf area index

3.4 Model evaluation

Based on the optimal parameters obtained during the training phase, the XGBoost model is utilized to predict the hydrological drought situation in the Huaihe River Basin from 2004 to 2014. These predictions will be assessed using precision and recall as measurement metrics. Precision is defined as the ratio of correctly classified instances of a specific class to the total number of predicted instances, quantifying the model's precision in predicting drought conditions and evaluating its reliability. Conversely, recall represents the ratio of correctly classified instances of a specific class to the total number of observed instances in that class, capturing the probability of the model predicting observed drought conditions and reflecting its sensitivity (Mardian et al., 2023; Zhang et al., 2023).

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Where the classification evaluation metrics employed are True Positives (TP), False Positives (FP), and False Negatives (FN). TP denotes the number of actual positive samples correctly predicted as positive, FP represents the number of actual negative samples incorrectly predicted as positive, and FN signifies the number of actual positive samples incorrectly predicted as negative.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

3.5 Shapley Additive Explanations (SHAP)

SHAP, a machine learning interpretability method, provides a unified approach by combining elements from additional variable attribution methods with Shapley values as a measure of variable importance. Shapley values were originally introduced in game theory to determine the contributions made by each player in cooperative games. The fundamental idea is that each player receives a corresponding payout based on their contribution (Shapley, 1953). The interpretation of SHAP values is straightforward: larger absolute SHAP values indicate greater weight of the variable in predicting the model, while negative (positive) SHAP values exert a negative (positive) influence on the prediction process. Lundberg and Lee (2017) developed the SHAP method based on the theoretical foundation of Shapley values to explain the influence of each variable on model predictions, thereby providing increased transparency to the model. The Shapley value is calculated as the average marginal contribution based on all possible variable permutations. The mathematical expression for the classic SHAP value is as follows:

$$\varphi_i = \sum_{S \subseteq N} \frac{|S|!\left(n-|S|-1\right)!}{n!}\left[v\left(S \cup \{i\}\right) - v\left(S\right)\right] \quad (7)$$

Where φ_i represents the contribution of variable i, N represent the set of all variables, n denote the number of variables N, S indicate the subset of N that includes variable i, and $v(N)$ represent the baseline, which signifies the predicted outcome of each variable in N when their values are unknown.

The model results for each observed value are estimated by summing the SHAP values of each variable corresponding to that observed value. Hence, formulating the explanation model as follows:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \quad (8)$$

Where, $z' \in \{0,1\}^M$, the variable quantity is denoted as M, and the value ϕ_i can be obtained from equation (7). SHAP offers a variety of AI model explainers.

In this study, we utilized a tree explainer to compute SHAP values based on the best XGBoost model for assessing drought impacts, aiming to estimate the contributions of each variable.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

4 Results

4.1 Model performance

The study period for this research spans from 1960 to 2014, with the model training period from 1960 to 2003 and the prediction period from 2004 to 2014. The input and output data types for 28 grid areas are the same. Take the 7th grid area as an example, when using monthly data, the input was 26 different drought influencing factors, and the output was SRI-1. The number of input samples during model training was 13767, and the number of output samples was 526. There are 3432 input samples and 132 output samples during the model prediction period. When using seasonal data, the input is 18 factors without drought, and the output is SRI-3 in different seasons. The number of input samples during model training is 792, and the number of output samples is 44. The number of input samples in the model prediction period is 198, and the number of output samples is 11. According to the data in Table 4 and Figure 2, the overall precision of the XGBoost model is 79.9%, which means that it has a 79.9% ability to correctly identify drought classes. In the identification of the ND drought class, the performance of the model is particularly excellent. Figure 2 shows that the median precision and recall rate of the ND class are both more than 0.8. It can be seen from the data in Table 4 that the recall rate of the ND drought class is 91% and the precision rate is 88%, which proves that the model has high sensitivity and reliability in predicting the ND drought class. At the same time, the precision rates of ND and D3 drought classes are 88% and 86%, respectively, indicating that the model had good prediction accuracy for these two types of droughts. However, the precision rates of the D1 and D2 drought classes are 74% and 61%, respectively, reflecting the lack of prediction accuracy of the model in these classes.

In addition, the boxplot in Figure 2 further reveals the precision and recall performance of the model for each drought class in 28 grid regions. Although the median precision and recall of the D1 drought class are both close to 0.8, indicating that the model has a high predictive ability in this class, the performance of the D2 and D3 drought classes is relatively poor. Especially for the D3 drought class, the median recall rate does not exceed 0.5, indicating that the model is not sensitive to the identification of such drought events, and there are some limitations in the prediction. owever, although the recall rate of the D3 drought class is low, its precision is almost as high as the ND drought class, which is mainly due to the low frequency of D3 drought class events. The model can successfully capture all D3 drought class events in some grid areas, thereby improving the precision of this class.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

Table 4: The average accuracy and recall rate indicators for each drought level predicted by the 28 regional models.

Class	Precision (%)	Recall (%)
ND	88	91
D1	74	78
D2	61	47
D3	86	50

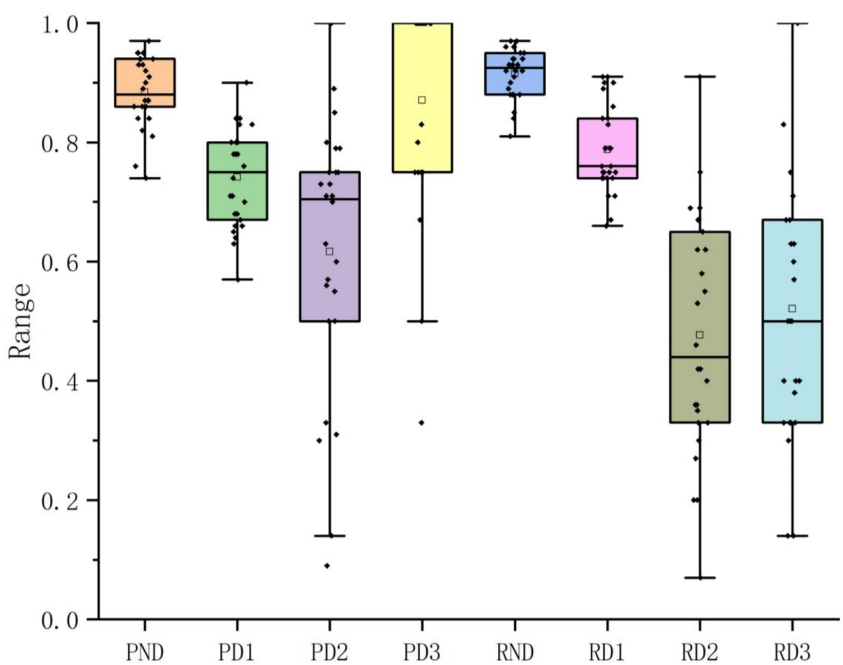

Figure 2: Box plots of the accuracy and recall rates of the four drought categories predicted by the 28 regional models ('P' represents the accuracy rate, and 'R' represents the recall rate. The small square represents the average.).

4.2 Prediction maps

According to the predicted drought data, 2011 was identified as a year with relatively severe drought conditions. To visually assess the predictive capability of the model, drought predicted, observed, and difference maps were created for each month of 2011 (Figure 3 to Figure 5). In 2011, the model accurately captured drought situations across most regions. In January, the drought situation was severe, and the drought class was mainly in the D2 and D3 classes. However, the prediction map of the model shows

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

that the drought degree in most regions is lighter than the actual drought situation, and the drought class is mainly classified as D1, which relatively underestimates the actual situation of drought. In February, the drought situation was rapidly reduced, and the prediction map of the model was basically consistent with the observation map. In March and April, the drought conditions in the entire basin rapidly escalated and became severe, and most of the areas in the observation map reached the drought classes of D2 and D3, and only a few areas in the north were classified as D1 drought class. Consequently, this period poses a considerable challenge to the predictive ability of the model, making it an appropriate period to evaluate the predictive performance of the model. In general, the model effectively predicts the occurrence and deterioration of drought and captures the spatial distribution pattern. However, in some parts of the central and western regions, the model still underestimates the drought situation.

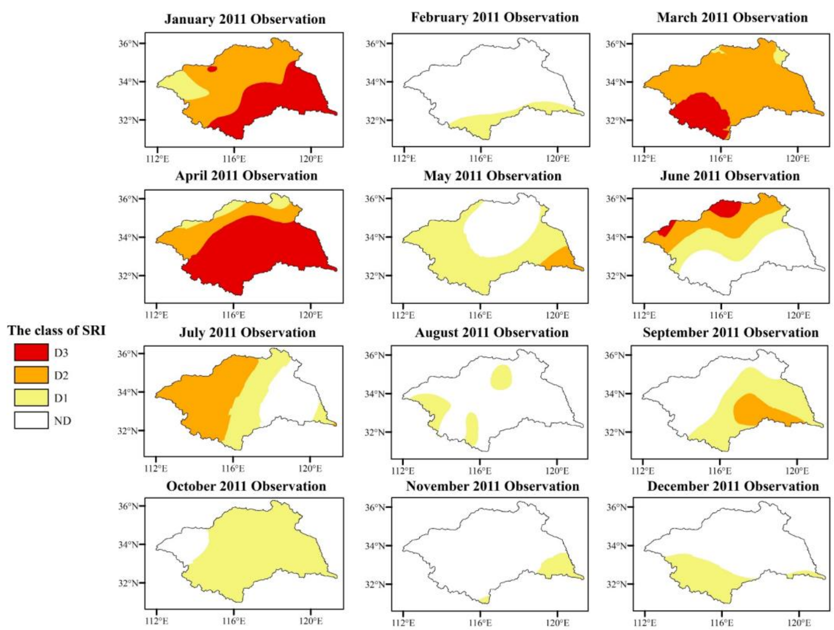

Figure 3: The observed drought types of each month in 2011.

In May, the severity of the drought situation decreased relative to the previous two months, and the actual observed map and the model-predicted map were largely consistent. According to the observed map, in June, a drought occurrence was observed in the northern region where no drought had been previously recorded. Furthermore, in July, the drought area shifted from the northern to the western region. It was not until August that drought gradually diminished in most areas. Basically, the model

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

captures the change of drought, but for some areas of D3 drought class, the model predicts them as D2 drought class.

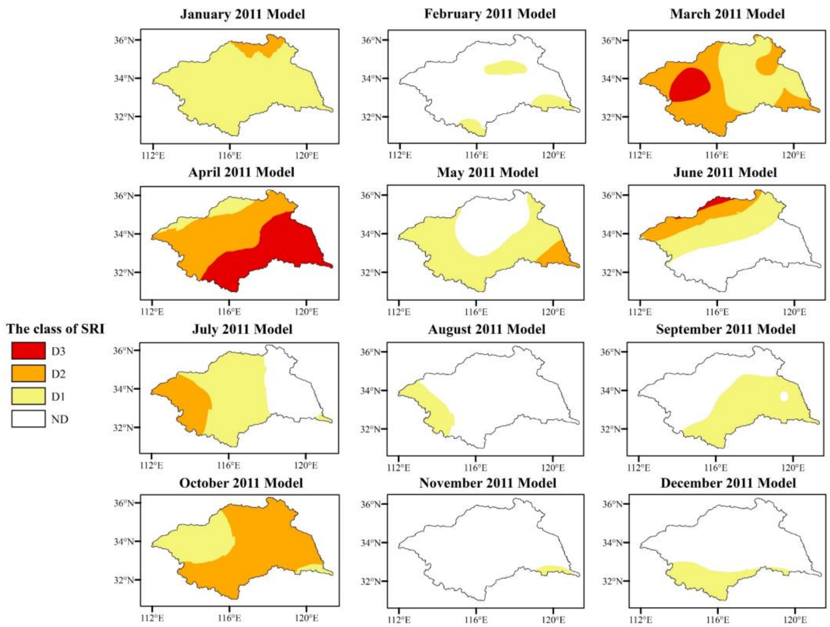

Figure 4: The drought types of each month in 2011 predicted by the Model.

In September, drought conditions were found in the eastern and southern regions on the observed map. However, the drought situation in some areas is underestimated on the map predicted by the model. In October, the model significantly overestimated the severity of the drought situation. According to the observed map, all regions except a small part of the western region experienced the D1 drought class. In contrast, the model-predicted map shows widespread drought across the region, with most of the regions classified in the D2 drought class. In November and December, the drought in the observation map dissipated rapidly, and the drought situation was basically the same as that in the model prediction map.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

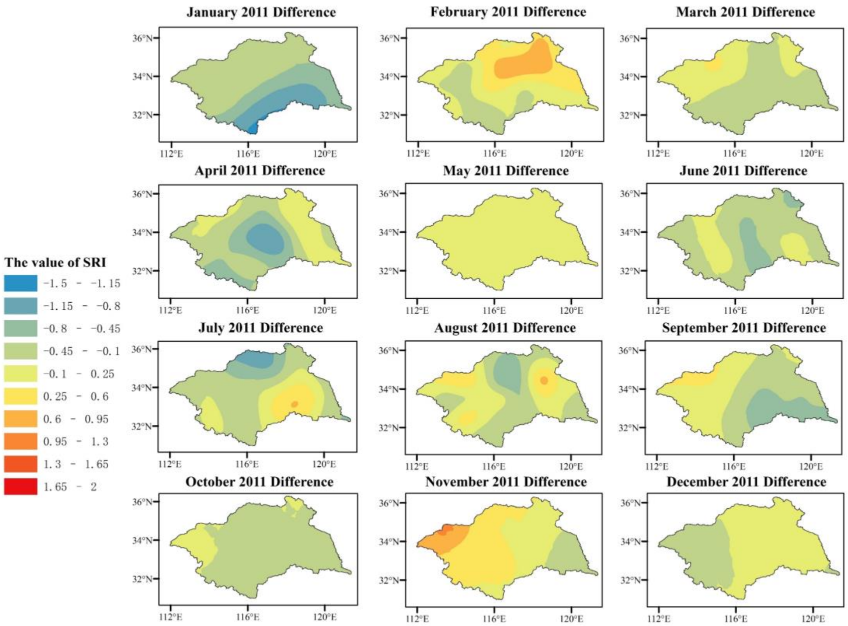

Figure 5: The difference between the predicted results of the model and the observed data values (Difference = SRIprediction - SRIactual).

In general, the XGBoost model has a great performance in capturing the spatial structure and temporal dynamics of drought events during the 12-month period of 2011. However, the model indicates that while the model can distinguish between drought and non-drought conditions, it lacks clarity in defining the boundaries between different drought classes. In most cases, the model underestimates drought conditions compared to the observed results.

4.3 Variable importance analysis

4.3.1 Monthly prediction analysis

To study the effects of different factors on drought, 26 different drought influencing factors were considered, and the corresponding influencing factors are analyzed for 28 grid regions, and the contribution analysis is made with SHAP values. Due to the limited space, only the analysis of the 7th grid region is shown in Figure 6. Figure 6 reveals the contribution of each input feature based on the SHAP value of each instance in 28 grid regions. In the vertical direction, the variables in the bee colony graph are sorted according to their absolute SHAP values, which also reflects the importance of ranking variables. The density of points represents the eigenvalues of each instance in each row. The X-axis

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

shows the SHAP value corresponding to a single instance. The left side of the Y-axis of the bee colony graph represents the negative total contribution of the features in the XGBoost model, while the right side represents the positive total contribution. The analysis reveals that SPI plays a dominant role, followed by AMO and evapotranspiration.

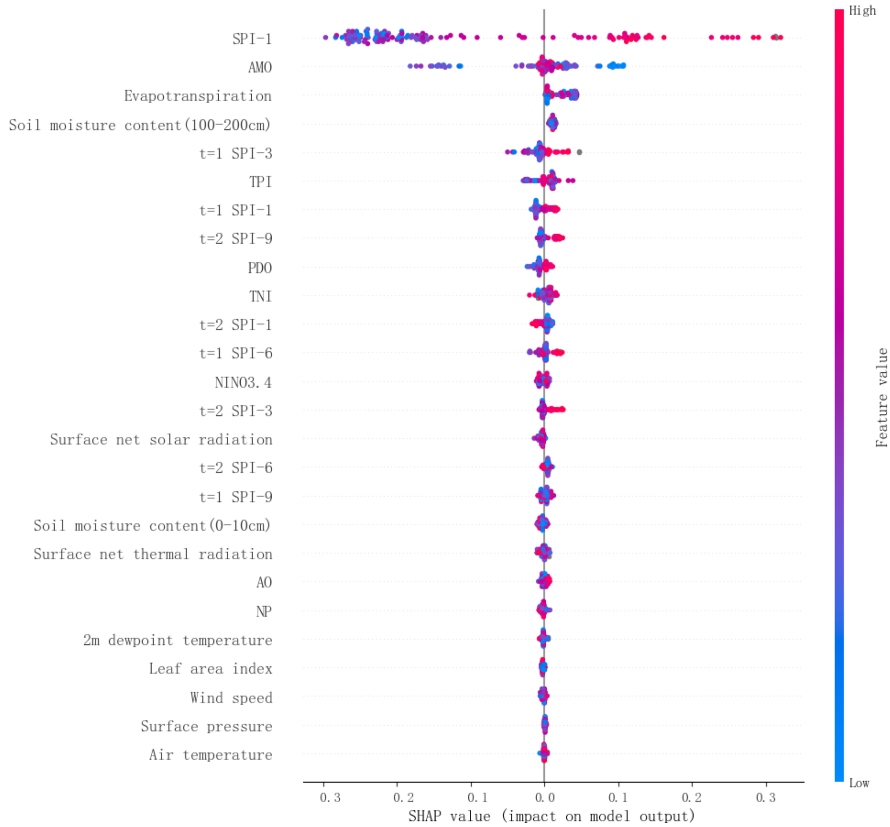

Figure 6: The SHAP values of 26 different influencing factors in each month of the 7th grid region from 2004 to 2014.

Figure 7 illustrates the interpretability of the XGBoost model focusing on the 7th grid region, providing insights into the average impact of the 26 influencing factors on model output. These findings corroborate the insights from Figure 6, highlighting that SPI, AMO, and evapotranspiration are the predominant factors influencing the predictions of the model. Table 3 indicates that the absolute average SHAP value of SPI, incorporating monthly precipitation data for the entire basin, is 0.190, marking it as the most substantial influence on hydrological drought within the 7th grid region.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

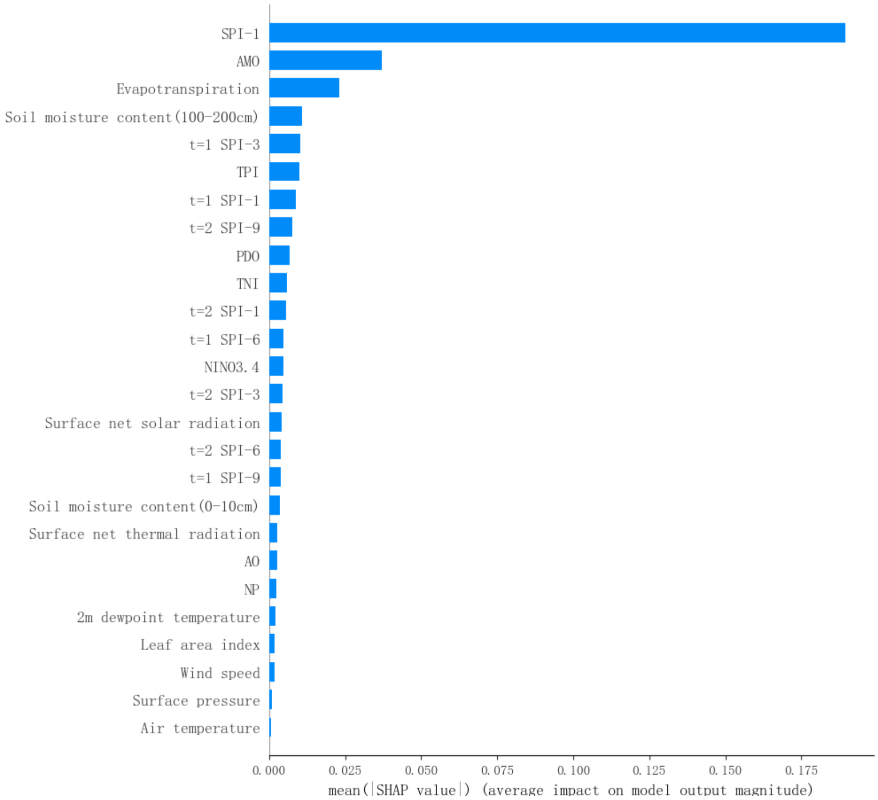

Figure 7: The absolute average SHAP values of 26 different influencing factors at the 7th grid region from 2004 to 2014.

To gain a deeper understanding of the factors contributing to drought events in the study area, As shown in Figure 8, this study shows the spatial distribution of the first three main drought-influencing factors and discusses the changes of drought-influencing factors in the basin. The results show that the main influencing factor of hydrological drought in the Huaihe River Basin is meteorological drought. As shown in Table 5, the absolute average SHAP value of the first influencing factor is significantly higher than that of the second and third influencing factors. Large-scale climate factors (particularly AMO) emerge as the secondary major influence, and about half of the North Central Basin is significantly dependent on these factors. For the third influencing factor, a diverse range of large-scale climate variables, such as TPI, PDO, NP, TNI, and AMO, affect almost half of the study area. In summary, the foremost determinant of hydrological drought is meteorological drought. Large-scale climate factors (notably AMO) rank second in importance, followed by factors like soil moisture content, and so on.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

The findings demonstrate that the Standardized Precipitation Index (SPI) serves as the dominant driver of hydrological drought in the Huaihe River Basin, consistent with the conclusions of Gan et al. (2023), who identified meteorological drought as a critical precursor to hydrological extremes in this region. Further support arises from Wang et al. (2021), whose analysis of drought propagation mechanisms in the Huaihe Basin revealed indirect hydrological drought impacts mediated through soil moisture and evapotranspiration—a pattern corroborated by the secondary influence of soil moisture and evapotranspiration in this study. However, compared with the study of Zou et al. (2018) in the Weihe River Basin, the influence of large-scale climate factors in this study is more prominent, which may be related to the fact that the Huaihe River Basin is located in the climate transition zone and is more sensitive to the air-sea coupling phenomenon.

Table 5: The first three drought influencing factors and the SHAP value of the absolute average influence of 28 grid areas in Huaihe River Basin.

SHAP value / grid area	The first influencing factor	Average SHAP value	The second influencing factor	Average SHAP value	The third influencing factor	Average SHAP value
1	SPI-1	0.160	Evapotranspiration	0.040	TPI	0.038
2	SPI-1	0.190	AO	0.018	Soil moisture content(100-200cm)	0.014
3	SPI-1	0.189	TPI	0.030	Soil moisture content(100-200cm)	0.023
4	SPI-1	0.178	NP	0.020	PDO	0.016
5	SPI-1	0.147	Evapotranspiration	0.044	NP	0.017
6	SPI-1	0.180	TPI	0.025	Evapotranspiration	0.021
7	SPI-1	0.190	AMO	0.037	Evapotranspiration	0.023
8	SPI-1	0.212	TPI	0.030	TNI	0.020
9	SPI-1	0.161	AMO	0.034	T=2 SPI-6	0.028
10	SPI-1	0.195	AMO	0.037	Surface net thermal radiation	0.031
11	SPI-1	0.226	AMO	0.037	TNI	0.012
12	SPI-1	0.221	AMO	0.033	T=2 SPI-3	0.017
13	SPI-1	0.228	AMO	0.028	NP	0.026
14	SPI-1	0.204	Soil moisture content(100-200cm)	0.057	T=1 SPI-1	0.029

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

15	SPI-1	0.160	Soil moisture content(100-200cm)	0.033	NP	0.032
16	SPI-1	0.157	Wind speed	0.033	AMO	0.030
17	SPI-1	0.186	AMO	0.064	Evapotranspiration	0.025
18	SPI-1	0.235	AMO	0.040	Soil moisture content(0-10cm)	0.032
19	SPI-1	0.168	TPI	0.055	AMO	0.035
20	SPI-1	0.172	AMO	0.038	T=2 SPI-3	0.026
21	SPI-1	0.165	AMO	0.039	PDO	0.039
22	SPI-1	0.179	AMO	0.042	Evapotranspiration	0.025
23	SPI-1	0.176	AMO	0.029	T=1 SPI-9	0.022
24	SPI-1	0.189	PDO	0.053	AMO	0.021
25	SPI-1	0.149	AMO	0.055	TPI	0.024
26	SPI-1	0.160	AMO	0.043	PDO	0.030
27	SPI-1	0.169	AMO	0.047	T=2 SPI-3	0.018
28	SPI-1	0.287	NP	0.025	T=1 SPI-1	0.016

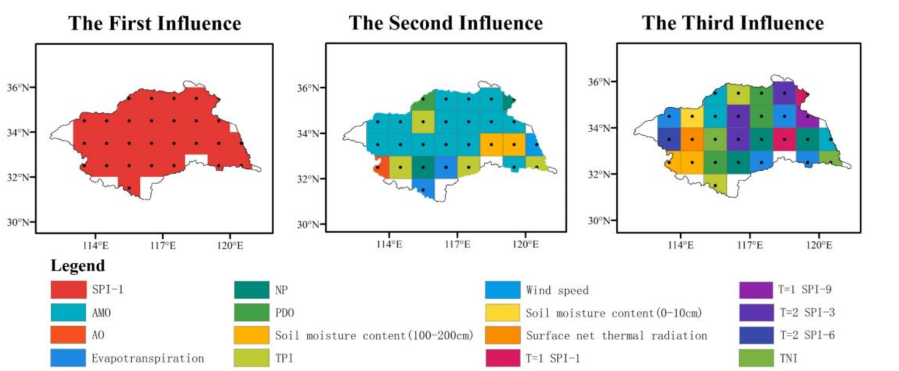

Figure 8: The first three drought-influencing factors of 28 grid areas in the Huaihe River Basin.

4.3.2 Seasonal prediction analysis

To accurately reflect the differences in drought-influencing factors across different seasons, this study utilized 18 different drought-influencing factors to predict the hydrological drought in the Huaihe River Basin. Histograms of the absolute average SHAP values for different influencing factors in four seasons in the 7th grid region are presented in Figure 9. The absolute average SHAP values of SPI-3 in spring, summer, autumn, and winter were 0.360, 0.261, 0.169, and 0.247 respectively, which had the greatest impact on hydrological drought in the same season. In addition, the absolute average SHAP values of evapotranspiration, soil moisture content, air temperature, and surface net thermal radiation were close to or exceeded 0.05, which also had a significant impact on hydrological drought in the Huaihe River Basin.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

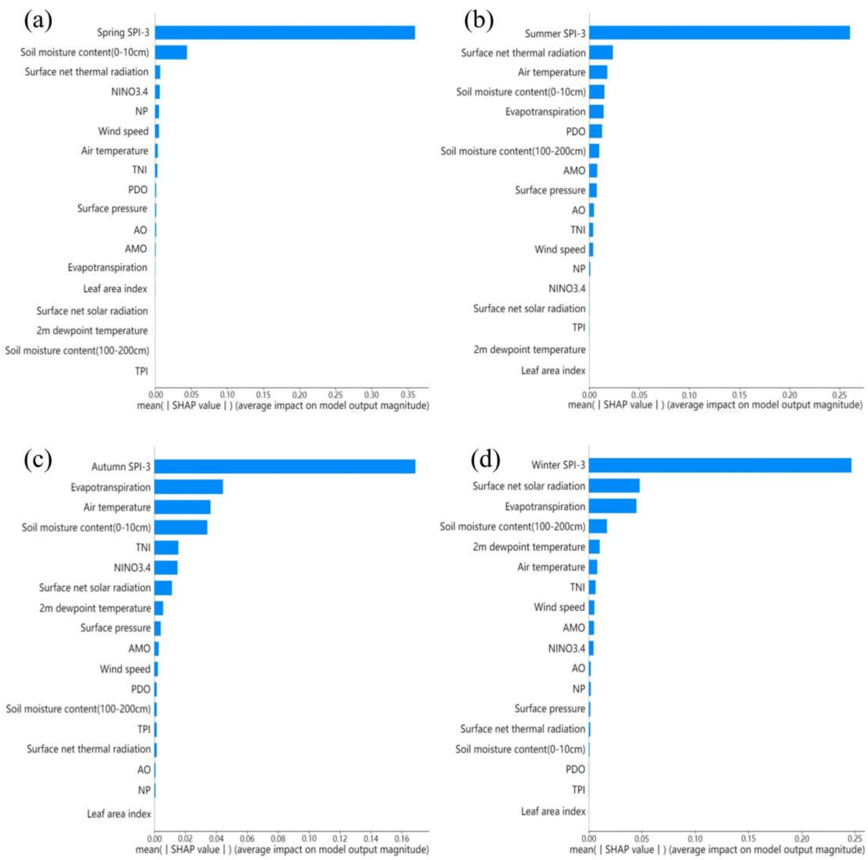

Figure 9: The absolute average SHAP values of 18 different influencing factors in the 7th grid region of four seasons ((a) Spring; (b) Summer; (c) Autumn; (d) Winter).

To understand the spatial and temporal distribution characteristics of drought and the potential impact mechanism, Figure 10 displays the spatial distribution of the top three influencing factors in each season. The leading influencing factors across the four seasons include SPI-3, soil moisture content, and surface net thermal radiation, with SPI-3 being predominant across all seasons and regions. As shown in Figure 11, the absolute average SHAP value of the primary factor exceeded the sum SHAP values of the second and third factors. Aside from SPI-3, soil moisture content also exerts a significant influence on hydrological drought in summer and autumn, particularly in the southern and southeastern parts of the river basin. In winter, certain areas in the central part of the river basin are mainly affected by surface net thermal radiation and surface net solar radiation.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

From the perspective of the second influencing factor, hydrological drought in most areas of the basin in spring is mainly affected by soil water content and evapotranspiration. In the rest of the region, surface pressure, temperature, radiation, and other factors also play an important role. It is worth noting that in the 15th grid region, the surface pressure becomes a key secondary influencing factor, and its absolute average SHAP value reaches 0.175. This value is significantly higher than the second impact factor in other regions, and even close to the primary impact factor in the same grid area. This indicates that it is extremely sensitive to surface pressure in this particular place. During summer, the influence of large-scale climatic factors such as the AMO, PDO, and TPI becomes more pronounced compared to spring. Additionally, soil moisture content and surface radiation continue to account for a substantial proportion of the influence on hydrological drought. Regions with absolute average SHAP values surpassing 0.1 in summer constitute approximately one-seventh of the study area, indicating elevated sensitivity to these factors during this season. Similar to spring, soil moisture content and evapotranspiration remain predominant influencing factors for hydrological drought in half of the grid areas during autumn and winter. The remaining regions are mainly influenced by surface net thermal radiation and surface net solar radiation. Specifically, during winter, the second influencing factors for three grid regions (the 12th, 13th, and 21st grid regions) in the central part of the basin are soil moisture content and evapotranspiration, with absolute average SHAP values exceeding 0.1. This indicates a relatively higher influence of these secondary factors in these regions compared to others.

Compared with the second impact factor, the large-scale climatic factors in the third impact factor have an increased influence on hydrological drought in the four seasons. In spring and autumn, soil moisture content exhibits a more substantial influence on hydrological drought, while in summer, air temperature is considered to be a more important factor. However, in winter, half of the study areas continue to be dominated by soil moisture content and evapotranspiration, whereas most of the remaining study areas are primarily influenced by large-scale climate factors such as TNI, PDO, NP, and AO.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

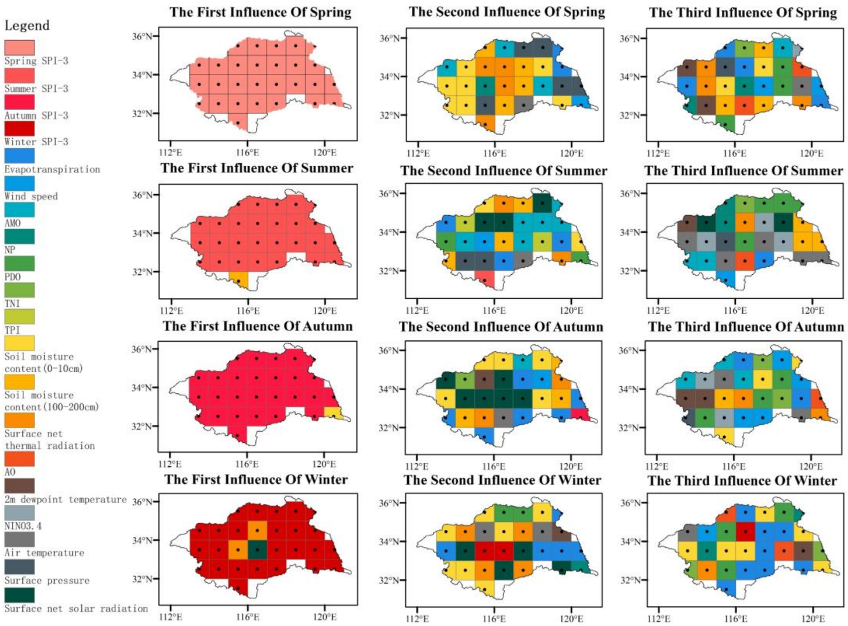

Figure 10: The first three drought-influencing factors of 28 grid points in Huaihe River Basin in each season.

According to the above results, there were significant differences in the influencing factors of drought among the four seasons. This diversity highlights the need for us to pay more attention to the weights and dynamic changes of various influencing factors when predicting and understanding the spatial-temporal distribution characteristics of drought. Although the SPI factor continues to dominate, at some grid points, factors such as soil moisture content in summer and autumn, as well as thermal radiation in winter, cannot be ignored. This suggests that even for the same influencing factor, its influence can vary greatly in different seasons and regions. Furthermore, in addition to the influence of meteorological drought, the influencing factors of spring hydrological drought are mainly biased toward soil moisture content and evapotranspiration, in addition to surface pressure, temperature, radiation, and other related factors. The absolute average SHAP value of these influencing factors is basically no more than 0.1, which is very different from SPI-3, but its impact on hydrological drought cannot be ignored. In autumn and winter, the above factors still dominate, but at the same time, the proportion of large-scale climate factors gradually increases, indicating that climate change between different seasons may play an important regulatory role in the composition of drought-influencing factors.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

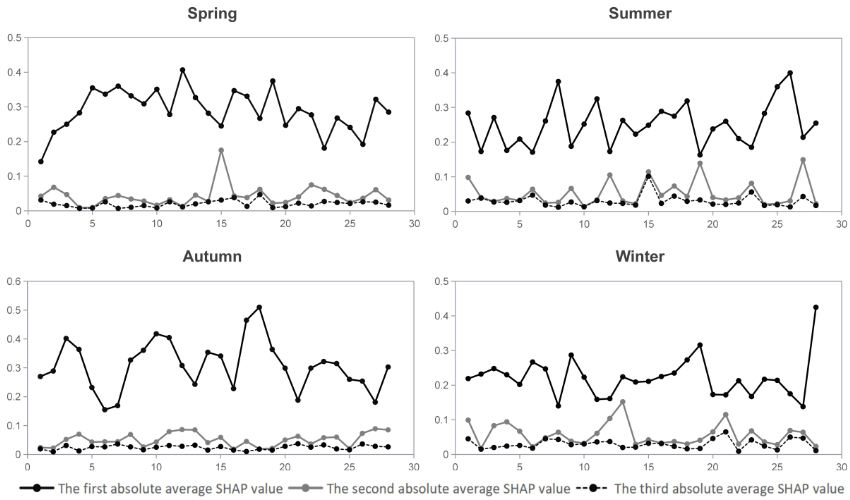

Figure 11: The absolute average SHAP values of the first three drought-influencing factors in each season (The X-axis represents 28 grid regions in the Huaihe River Basin).

5 Discussion

SHAP analysis based on the XGBoost model unequivocally identifies the SPI as the most influential predictor of hydrological drought across the Huaihe River Basin. Beyond SPI, the key secondary drivers exhibit a distinct spatial and seasonal differences. In terms of space, the hydrological drought in the northern part of the basin shows higher sensitivity to large-scale climate oscillations such as AMO, indicating that large-scale climate factors regulate regional precipitation patterns (Yu et al., 2024). On the contrary, the secondary factors affecting the hydrological drought in the southern part of the basin are mainly surface processes, especially soil moisture and evapotranspiration.(Mtupili et al., 2025; Zhu et al., 2025). The difference in the second influencing factors of hydrological drought in the southern and northern parts of the basin may be due to the fact that the basin belongs to the temperate-subtropical transition position. For the seasonal scale, in spring, soil moisture and evapotranspiration account for a large proportion of the explanatory power of the model. In summer, the relative weight of large-scale climatic factors increases, which is consistent with the enhancement of water vapor transport (Yu et al., 2024). In autumn and winter, radiative fluxes (net solar and thermal radiation) assume greater importance (Jin et al., 2025). Collectively, these findings underscore SPI as the primary driver while revealing the nuanced spatio-temporal controls exerted by secondary factors, thereby providing a scientific foundation

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

for developing more targeted drought mitigation and water resource management strategies across the diverse Huaihe River Basin.

When studying the influence of large-scale climate indices on drought, the correlation between climate indices and drought for the same period and a certain lead time is often considered, and the results show that climate indices for the same period and different lead times have a certain influence on drought in the basin, and the degree of influence varies with the changes in the study area. For example, Ren et al. (2017) studied the correlation between SPI and large-scale climate indices with advance periods of 0, 1, 2, and 3 months, and the correlation results show that Nino3.4 has significant correlation in August-October, and PDO has significant correlation in January-May and June-December of the same period. Lv et al. (2022) analyzed the correlation between large-scale climatic factors and drought in different lag periods. The results show that large-scale climatic factors in the same period also have an impact on drought. Due to the many influencing factors considered in this paper, only the effect of climate indices on drought in the basin during the same period was considered when selecting the large-scale climate indices. Subsequent studies can consider selecting the most relevant large-scale climate factors in different months or seasons as the influencing factors for basin drought prediction to further improve the accuracy of drought prediction. Before inputting the influencing factors into the machine learning model for training, methods such as random forest and principal component analysis (PCA) can be used to select the influencing factors. The application of these methods can optimize the influencing factors and provide strong support for more accurate drought trend prediction and management strategies.

6 Conclusions

Drought is one of the most significant environmental and climate problems in the world, and drought prediction is a crucial means of drought prevention. In this study, the integration of SHAP and XGBoost provides a novel framework that can not only improve the prediction accuracy, but also show the impact of different drought influencing factors on drought. The framework can provide two types of support for decision makers: (1) giving priority to high weight factors in real-time drought warning; (2) Identifying early risk signals in long-term water resources planning. The main conclusions are as follows:

1) The XGBoost model achieved an accuracy of 79.9% for identifying drought classes. The model performs particularly well in predicting ND and D1 drought classes, with a precision rate of

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

88 % and 74 %, respectively. It also has a recall rate of 91 % and 78 %. However, the prediction performance of the model for the D2 and D3 drought classes is relatively poor, especially for the D3 drought, the recall rate should not exceed 0.5, indicating that the recognition sensitivity of the model for the D3 class is limited. In general, the model has high prediction reliability for ND and D1 classes, but limits in the prediction performance of D2 and D3 classes.

2) This study determined that SPI is the most critical factor affecting hydrological drought in the Huaihe River Basin. In 28 grid regions, the absolute average SHAP value of SPI is not less than 0.147, which is much higher than other influencing factors. In addition, large-scale climate factors, soil moisture content, and evapotranspiration play a significant role in hydrological drought in the basin.

3) The SPI remains a major influence in all seasons with absolute average SHAP values of 0.360, 0.261, 0.169, and 0.247 in spring, summer, autumn, and winter respectively. Additional factors such as soil moisture content, net heat radiation, and solar radiation also play seasonal roles. Soil moisture content and evapotranspiration are significant factors in spring and autumn, while temperature and large-scale climate factors are critical in summer and winter.

Reference

[1] Ardabili, S., Mosavi, A., Dehghani, M., Várkonyi-Kóczy, A.R., 2020. Deep Learning and Machine Learning in Hydrological Processes Climate Change and Earth Systems a Systematic Review, in: Várkonyi-Kóczy, A.R. (Tran.), Engineering for Sustainable Future. Springer International Publishing, Cham, pp. 52–62.

[2] Bachmair, S., Svensson, C., Hannaford, J., Barker, L.J., Stahl, K., 2016. A quantitative analysis to objectively appraise drought indicators and model drought impacts. Hydrol. EARTH Syst. Sci. 20, 2589–2609. https://doi.org/10.5194/hess-20-2589-2016

[3] Barnwal, A., Cho, H., Hocking, T., 2022. Survival Regression with Accelerated Failure Time Model in XGBoost. J. Comput. Graph. Stat. 31, 1292–1302. https://doi.org/10.1080/10618600.2022.2067548

[4] Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

KDD ’16. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[5] Choi, H.-S., Kim, S., Oh, J.E., Yoon, J.E., Park, J.A., Yun, C.-H., Yoon, S., 2018. XGBoost-Based Instantaneous Drowsiness Detection Framework Using Multitaper Spectral Information of Electroencephalography, ACM-BCB’18: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Presented at the 9th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB), Washington, DC, USA. https://doi.org/10.1145/3233547.3233567

[6] Dai, A., 2013. Increasing drought under global warming in observations and models (vol 3, pg 52, 2013). Nat. Clim. CHANGE 3, 171–171. https://doi.org/10.1038/NCLIMATE1811

[7] Dai, A., 2012. Erratum: Drought under global warming: a review. WILEY Interdiscip. Rev.-Clim. CHANGE 3, 617–617. https://doi.org/10.1002/wcc.190

[8] Dai, A., 2011. Characteristics and trends in various forms of the Palmer Drought Severity Index during 1900-2008. J. Geophys. Res.-ATMOSPHERES 116, D12115. https://doi.org/10.1029/2010JD015541

[9] Dikshit, A., Pradhan, B., 2021. Interpretable and explainable AI (XAI) model for spatial drought prediction. Sci. TOTAL Environ. 801, 149797. https://doi.org/10.1016/j.scitotenv.2021.149797

[10] Fu, Q., Zhou, Z., Li, T., Liu, D., Hou, R., Cui, S., Yan, P., 2018. Spatiotemporal characteristics of droughts and floods in northeastern China and their impacts on agriculture. Stoch. Environ. Res. RISK Assess. 32, 2913–2931. https://doi.org/10.1007/s00477-018-1543-z

[11] Gan, R., Li, D., Chen, C., Yang, F., Zhang, X., Guo, X., 2023. Spatiotemporal characteristics of extreme hydrometeorological events and its potential influencing factors in the Huaihe River Basin, China. Stoch. Environ. Res. RISK Assess. 37, 2693–2712. https://doi.org/10.1007/s00477-023-02413-4

[12] Gao, L., Ding, Y., 2020. Disease prediction via Bayesian hyperparameter optimization and ensemble learning. BMC Res. NOTES 13, 205. https://doi.org/10.1186/s13104-020-05050-0

[13] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z., 2019. XAI-Explainable artificial intelligence. Sci. Robot. 4, eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[14] Han, Y., Wu, J., Zhai, B., Pan, Y., Huang, G., Wu, L., Zeng, W., 2019. Coupling a Bat Algorithm

with XGBoost to Estimate Reference Evapotranspiration in the Arid and Semiarid Regions of China. Adv. Meteorol. 2019, 9575782. https://doi.org/10.1155/2019/9575782

[15] Islam, M.R., Ahmed, M.U., Barua, S., Begum, S., 2022. A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. Appl. Sci.-BASEL 12, 1353. https://doi.org/10.3390/app12031353

[16] Jin, H., Zhang, K., Zhang, P., Liu, G., Liu, M., Chen, X., Willems, P., 2025. Spatiotemporal evolution of drought status and its driving factors attribution in China. Sci. Total Environ. 958, 178131–178131. https://doi.org/10.1016/j.scitotenv.2024.178131

[17] Jungho, S., Kim, Y., 2023. Assessing the likelihood of drought impact occurrence with extreme gradient boosting: a case study on the public water supply in South Korea. J. HYDROINFORMATICS 25, 191–207. https://doi.org/10.2166/hydro.2023.064

[18] Kikon, A., Deka, P.C., 2022. Artificial intelligence application in drought assessment, monitoring and forecasting: a review. Stoch. Environ. Res. RISK Assess. 36, 1197–1214. https://doi.org/10.1007/s00477-021-02129-3

[19] Li, J., Wang, Z., Wu, X., Xu, C.-Y., Guo, S., Chen, X., Zhang, Z., 2021. Robust Meteorological Drought Prediction Using Antecedent SST Fluctuations and Machine Learning. WATER Resour. Res. 57, e2020WR029413. https://doi.org/10.1029/2020WR029413

[20] Li, M., Feng, Z., Zhang, M., Yao, Y., 2024. Influence of large-scale climate indices and regional meteorological elements on drought characteristics in the Luanhe River Basin. ATMOSPHERIC Res. 300, 107219. https://doi.org/10.1016/j.atmosres.2024.107219

[21] Lu, R., Dong, B., Ding, H., 2006. Impact of the Atlantic Multidecadal Oscillation on the Asian summer monsoon. Geophys. Res. Lett. 33. https://doi.org/10.1029/2006GL027655

[22] Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.

[23] Lv, A., Fan, L., Zhang, W., 2022. Impact of ENSO Events on Droughts in China. Atmosphere 13. https://doi.org/10.3390/atmos13111764

[24] Mardian, J., Champagne, C., Bonsal, B., Berg, A., 2023. A Machine Learning Framework for Predicting and Understanding the Canadian Drought Monitor. WATER Resour. Res. 59, e2022WR033847. https://doi.org/10.1029/2022WR033847

[25] McKee, T.B., Doesken, N.J., Kleist, J., 1993. THE RELATIONSHIP OF DROUGHT FREQUENCY AND DURATION TO TIME SCALES.

[26] Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. J. Hydrol. 391, 204–216. https://doi.org/10.1016/j.jhydrol.2010.07.012

[27] Molnar, C., 2022. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed.

[28] Mtupili, M., Wang, R., Gu, L., Yin, J., 2025. Delayed Response of Soil Moisture and Hydrological Droughts to Meteorological Drought Over East Asia. Int. J. Climatol. https://doi.org/10.1002/joc.8883

[29] Orimoloye, I.R., Ololade, O.O., Belle, J.A., 2021. Satellite-based application in drought disaster assessment using terra MOD13Q1 data across free state province, South Africa. J. Environ. Manage. 285, 112112. https://doi.org/10.1016/j.jenvman.2021.112112

[30] Orimoloye, I.R., Olusola, A.O., Belle, J.A., Pande, C.B., Ololade, O.O., 2022. Drought disaster monitoring and land use dynamics: identification of drought drivers using regression-based algorithms. Nat. HAZARDS 112, 1085–1106. https://doi.org/10.1007/s11069-022-05219-9

[31] Park, H., Kim, K., Lee, D.K., 2019. Prediction of Severe Drought Area Based on Random Forest: Using Satellite Image and Topography Data. WATER 11, 705. https://doi.org/10.3390/w11040705

[32] Phan-Van, T., Nguyen-Ngoc-Bich, P., Ngo-Duc, T., Vu-Minh, T., Le, P.V.V., Trinh-Tuan, L., Nguyen-Thi, T., Pham-Thanh, H., Tran-Quang, D., 2022. Drought over Southeast Asia and Its Association with Large-Scale Drivers. J. Clim. 35, 4959–4978. https://doi.org/10.1175/JCLI-D-21-0770.1

[33] Prodhan, F.A., Zhang, J., Hasan, S.S., Sharma, T.P.P., Mohana, H.P., 2022. A review of machine learning methods for drought hazard monitoring and forecasting: Current research trends, challenges, and future research directions. Environ. Model. Softw. 149, 105327. https://doi.org/10.1016/j.envsoft.2022.105327

[34] Raposo, V. de M.B., Costa, V.A.F., Rodrigues, A.F., 2023. A review of recent developments on drought characterization, propagation, and influential factors. Sci. Total Environ. 898, 165550. https://doi.org/10.1016/j.scitotenv.2023.165550

[35] Ren, W., Wang, Y., Li, J., Feng, P., Smith, R.J., 2017. Drought forecasting in Luanhe River

basin involving climatic indices. Theor. Appl. Climatol. 130, 1133–1148. https://doi.org/10.1007/s00704-016-1952-1

[36] Ribeiro, M.T., Singh, S., Guestrin, C., 2016a. Model-Agnostic Interpretability of Machine Learning.

[37] Ribeiro, M.T., Singh, S., Guestrin, C., 2016b. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

[38] Shapley, L.S., 1953. 17. A Value for n-Person Games, in: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press, Princeton, pp. 307–318. https://doi.org/doi:10.1515/9781400881970-018

[39] Shi, H., Chen, J., Wang, K., Niu, J., 2018. A new method and a new index for identifying socioeconomic drought events under climate change: A case study of the East River basin in China. Sci. TOTAL Environ. 616, 363–375. https://doi.org/10.1016/j.scitotenv.2017.10.321

[40] Shukla, S., Wood, A.W., 2008. Use of a standardized runoff index for characterizing hydrologic drought. Geophys. Res. Lett. 35, L02405. https://doi.org/10.1029/2007GL032487

[41] Sun, A.Y., Scanlon, B.R., 2019. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ. Res. Lett. 14, 073001. https://doi.org/10.1088/1748-9326/ab1b7d

[42] Sun, P., Sun, Y., Zhang, Q., Yao, R., 2018. Hydrological Processes in the Huaihe River Basin, China: Seasonal Variations, Causes and Implications. Chin. Geogr. Sci. 28, 636–653. https://doi.org/10.1007/s11769-018-0969-z

[43] Wang, J., Wang, W., Cheng, H., Wang, H., Zhu, Y., 2021. Propagation from Meteorological to Hydrological Drought and Its Influencing Factors in the Huaihe River Basin. WATER 13, 1985. https://doi.org/10.3390/w13141985

[44] Wu, Y., Xu, Y., 2020. Assessing the Climate Tendency over the Yangtze River Delta, China: Properties, Dry/Wet Event Frequencies, and Causes. Water 12. https://doi.org/10.3390/w12102734

[45] Wu, Z., Yin, H., He, H., Li, Y., 2022. Dynamic-LSTM hybrid models to improve seasonal drought predictions over China. J. Hydrol. 615, 128706.

https://doi.org/10.1016/j.jhydrol.2022.128706

[46] Xiao, L., Chen, X., Zhang, R., Zhang, Z., 2019. Spatiotemporal Evolution of Droughts and Their Teleconnections with Large-Scale Climate Indices over Guizhou Province in Southwest China. Water 11. https://doi.org/10.3390/w11102104

[47] Xu, D., Zhang, Q., Ding, Y., Zhang, D., 2022. Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. Environ. Sci. Pollut. Res. 29, 4128–4144. https://doi.org/10.1007/s11356-021-15325-z

[48] Yalcin, S., Esit, M., Coban, O., 2023. A new deep learning method for meteorological drought estimation based-on standard precipitation evapotranspiration index. Eng. Appl. Artif. Intell. 124, 106550. https://doi.org/10.1016/j.engappai.2023.106550

[49] Yu, J., Li, Q., Ding, Y., Wen, Z., Gong, Z., Sun, X., Shen, X., Dong, L., 2024. AMO modulation of interdecadal background of persistent heavy rainfall in summer over the Huaihe River Basin. Clim. Dyn. 62, 3621–3640. https://doi.org/10.1007/s00382-023-07088-9

[50] Yu, Q., Jiang, L., Wang, Y., Liu, J., 2023. Enhancing streamflow simulation using hybridized machine learning models in a semi-arid basin of the Chinese loess Plateau. J. Hydrol. 617, 129115. https://doi.org/10.1016/j.jhydrol.2023.129115

[51] Zhang, B., Abu Salem, F.K., Hayes, M.J., Smith, K.H., Tadesse, T., Wardlow, B.D., 2023. Explainable machine learning for the prediction and assessment of complex drought impacts. Sci. TOTAL Environ. 898, 165509. https://doi.org/10.1016/j.scitotenv.2023.165509

[52] Zhang, R., Chen, Z.-Y., Xu, L.-J., Ou, C.-Q., 2019. Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China. Sci. TOTAL Environ. 665, 338–346. https://doi.org/10.1016/j.scitotenv.2019.01.431

[53] Zhou, Z., Shi, H., Fu, Q., Ding, Y., Li, T., Liu, S., 2021. Investigating the Propagation From Meteorological to Hydrological Drought by Introducing the Nonlinear Dependence With Directed Information Transfer Index. WATER Resour. Res. 57, e2021WR030028. https://doi.org/10.1029/2021WR030028

[54] Zhou, Z., Shi, H., Fu, Q., Li, T., Gan, T.Y., Liu, S., 2020. Assessing spatiotemporal characteristics of drought and its effects on climate-induced yield of maize in Northeast China. J. Hydrol. 588, 125097. https://doi.org/10.1016/j.jhydrol.2020.125097

[55] Zhu, S., Huang, W., Wei, Y., Guo, J., Qin, H., 2025. Impact of driving factors on drought

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

propagation: perspectives on rainfall deficit and excessive evaporation. Clim. Dyn. 63, 204. https://doi.org/10.1007/s00382-025-07618-7

[56] Zou, L., Xia, J., She, D., 2018. Analysis of Impacts of Climate Change and Human Activities on Hydrological Drought: a Case Study in the Wei River Basin, China. WATER Resour. Manag. 32, 1421–1438. https://doi.org/10.1007/s11269-017-1877-1

Statements & Declarations

Software and data availability

The code and data set for prediction using python language (version 3.9.13) can be found in Mendeley Data: doi: 10.17632 / jnr2z36g77.1. The warehouse was created by Yuhang Yao (e-mail: 151746151 @ qq.com). The author 's experimental environment is as follows:

CPU: AMD Ryzen 9 7845HX 3.00 GHz;

GPU: NVIDIA GeForce RTX4060 8 GB；

RAM:16G.

Funding

This work was supported by the State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation (No. HESS-2206), the Open Fund of Key Laboratory of Flood & Drought Disaster Defense, the Ministry of Water Resources (KYFB202307260034), and Yangzhou University Graduate Student Research and Practice Innovation Program Funding projects (SJCX24_2250).

Competing Interests

The authors declare no conflict of interest.

Authors and Affiliations

College of Hydraulic Science and Engineering, Yangzhou University, Yangzhou, China

Min Li, Yuhang Yao, Zilong Feng, Ming Ou

Corresponding author

Correspondence to Min Li, Mail: limintju@126.com

Author Contributions

All authors contributed this paper: Conceptualization, M L; Data curation, M L, YH Y, ZL F, Ming Ou; Visualization, YH Y, M L; Validation, M L, YH Y; Methodology, M L, YH Y, ZL F; Formal Analysis, M L, YH Y; Funding acquisition, M L; Writing – original draft, YH Y; Writing – review & editing, M L.

https://doi.org/10.5194/egusphere-2025-1891
Preprint. Discussion started: 25 June 2025

Data availability

We are grateful to the National Oceanic and Atmospheric Administration (http://www.esrl.noaa.gov/psd/data/climateindices) for providing large-scale climate index data, and grateful to the GLDAS for providing monthly average precipitation, temperature, wind speed, soil water content, evapotranspiration data sets and runoff data sets, and to the European Centre for Medium-Range Weather Forecasts (https://cds.climate.copernicus.eu/) for providing monthly average 2 m dewpoint temperature, surface net solar radiation, surface net thermal radiation, surface pressure, and leaf area index data sets. The data and materials of this study are available.