Predicting Vegetation Indices in Yangtze River Delta based on climate and remote sensing data from 2001 to 2021 by Machine Learning Approach

Abdul Basit [1*], Ali Shahzad [2], Gul Amina [3], Moughal Taugir [4], Shawkat Ali [5], Hidayat Ullah [6], Zakria Zaheen [7], Muhammad Awais [8], Jiahua Zhang [9]

[1-9] Space Information and Big Earth Data Research Center, College of Computer Science and Technology, Qingdao University, Qingdao, 266071, China

* Corresponding Author: **Abdul Basit**

Article Info

P-ISSN: 3051-3383
E-ISSN: 3051-3391
Volume: 06
Issue: 02
July - December 2025
Received: 17-08-2025
Accepted: 18-09-2025
Published: 13-10-2025
Page No: 19-31

Abstract
Vegetation in the Yangtze River Delta (YRD) is significantly influenced by climatic variables such as potential evapotranspiration (PET), precipitation, and temperature. However, the prediction of vegetation dynamics in this ecologically sensitive region remains limited. To address this, the present study employed machine learning algorithms—Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM)—to predict the Normalized Difference Vegetation Index (NDVI) based on climatic and remote sensing data from 2001 to 2021. Model performance was evaluated using R^2 and Root Mean Square Error (RMSE) metrics. Among the tested models, RF exhibited the highest accuracy and generalization across different sub-regions of the YRD, with R^2 values exceeding 0.85 on testing datasets and low RMSE values, indicating its superior capability in capturing non-linear environmental interactions. Temporal trend analysis revealed a steady increase in NDVI, rising from 0.475 in 2001 to 0.547 in 2021, supported by a consistent rise in precipitation and PET. Meanwhile, TCI declined from 0.534 to 0.402, suggesting increasing temperature stress on vegetation. These results underscore the complex interactions between vegetation and climate variables. Pearson correlation analysis revealed strong positive correlations between NDVI and both precipitation ($r = 0.62$) and PET ($r = 0.80$), while TCI exhibited a weaker correlation, indicating a lesser but notable impact on vegetation. SHAP (SHapley Additive exPlanations) analysis further highlighted the dominant influence of PET in predicting NDVI, followed by precipitation, with TCI having a marginal impact. These findings underscore the critical role of water availability and evapotranspiration demand in shaping vegetation dynamics. The study demonstrates the potential of machine learning in ecological forecasting and provides actionable insights for land-use planning, climate adaptation, and environmental monitoring in the YRD and similar climate-sensitive ecosystems.

DOI: https://doi.org/10.54660/IJAIET.2025.6.2.19-31

Keywords: Machine Learning, NDVI, PET, TCI, Precipitation, Prediction.

1. Introduction

Climate change has significantly influenced natural ecosystems, resulting in substantial alterations to land cover and vegetation dynamics (Huang *et al*., 2024) [13]. Vegetation is a key indicator of environmental changes, reflecting shifts in water, soil, and atmospheric conditions (He *et al*., 2020). In regions such as the Yangtze River Delta (YRD), these changes are closely linked to expanding impervious surfaces and declining vegetation cover. Understanding the spatiotemporal dynamics of vegetation is crucial for determining the effects of environmental stress and climate variability on ecosystem health (Thakur *et al*., 2024) [27]. Maintaining ecological balance in climate-sensitive regions like the YRD is particularly challenging due to increasing drought and temperature stress in recent decades. The NDVI is a vital variable for monitoring vegetation growth and ecosystem dynamics, as it quantifies reflectance in the visible and near-infrared spectra (Mallick *et al*., 2024) [18]. Similarly, the TCI is crucial for evaluating regional and global thermal variations and their interactions with vegetation (Pandey *et al*., 2022) [22]. In addition, climate variables such as precipitation and PET are essential for assessing drought and temperature stresses in vulnerable ecosystems.

Despite the growing availability of spatiotemporal datasets and advancements in remote sensing technologies, the complex interactions between temperature, drought stress, and vegetation dynamics in the YRD remain inadequately understood (Hossain *et al*., 2020) [12]. Traditional statistical methods, including multiple linear regression and correlation analysis, have provided foundational insights but often struggle to capture the nonlinear relationships between climate variables and vegetation (Yi *et al*., 2023a) [30]. Machine learning techniques offer a promising alternative due to their ability to manage large datasets and effectively model complex interactions (Yi *et al*., 2023b) [30].

This study leverages advanced machine learning models RF, SVM, and XGBoost to predict vegetation dynamics in the YRD. The goal is to develop reliable prediction models that accurately reflect the complex relationships between climatic factors and vegetation dynamics from 2001 to 2021. These models predict NDVI values using key climate variables such as temperature, TCI, PET, and precipitation. Unlike earlier studies that relied primarily on simple pairwise correlations, this research adopts machine learning techniques to model the nonlinear interactions among multiple variables. The findings aim to support climate-resilient development and sustainable land use management by identifying the most influential climatic factors affecting vegetation in the YRD (Wang *et al*., 2016) [28]. Moreover, this study demonstrates the efficacy of machine learning in capturing complex, nonlinear interactions between climatic factors and vegetation responses, offering a robust predictive framework for vegetation monitoring in rapidly changing ecosystems. These findings contribute to the growing body of research advocating for data-driven approaches in environmental management.

2. Data Collection and Preprocessing

2.1. Study Area

The YRD (Figure 1) is selected as the main Study area for vegetation prediction using machine learning models, where Jiangsu, Zhejiang, Anhui provinces, and the Shanghai Municipality are all included in the YRD, which has latitudes between 28°30′N and 32°30′N and longitudes between 118°30′E and 122°E. Its varied landscapes, which have been influenced by booming agriculture, fast urbanization, and industrial growth, provide the perfect environment for researching the relationship between vegetation and climate. The YRD has a subtropical marine monsoon climate, which is defined by cool, dry winters and hot, humid summers(Ding *et al*., 2013) [7]. Significant temperature swings occur in the area in the spring, with sporadic winter lows of -10°C, underscoring the region's climatic variability and its impact on the local ecology.

2.2. Data Collection and Preprocessing:

The dataset for this study was sourced from Google Earth Engine (GEE) and covers the years 2001–2021. We collected NDVI data using the MOD13Q1 product at a spatial resolution of 250 meters and a temporal frequency of 16 days. The following formula is commonly used to calculate the NDVI, which is widely used in remote sensing studies(Zhang *et al*., 2019) [32].

$$NDVI = \frac{NIR-RED}{\mathrm{NIR+RED}} \tag{1}$$

Where, NIR represents the near-infrared band, and RED denotes the red band.

Potential Evapotranspiration (PET) and precipitation data were extracted from the CHIRPS dataset at a monthly temporal frequency and a 5000-meter resolution. We used the modified Penman-Monteith equation from the Food and Agriculture Organization (FAO) to estimate potential evapotranspiration (PET). This formula is frequently used to estimate PET in a variety of climatic zones, and numerous studies have assessed its best performance(Zarch *et al*., 2015). The PET calculation was conducted using the following equation:

$$PET = \frac{0.408\Delta(Rn-G)+\gamma\frac{900}{\mathrm{T+273}}u2(\mathrm{es-ea})}{\gamma(1+0.34u2)+\Delta} \tag{2}$$

PET is calculated by applying the Penman-Monteith equation, which includes several important variables. Where T is the average temperature for the month (°C), u2 is the wind speed at 2 meters, es − ea is the difference in vapor pressure between saturated and actual conditions, Δ is the slope of the vapor pressure curve (kPa/°C), γ is the dry hygrometer constant (kPa/°C), and Rn is the net radiation (MJ/m²/day), which is the amount of energy that is available at the surface. These factors work together to estimate evapotranspiration, which represents ecosystems' water demands.

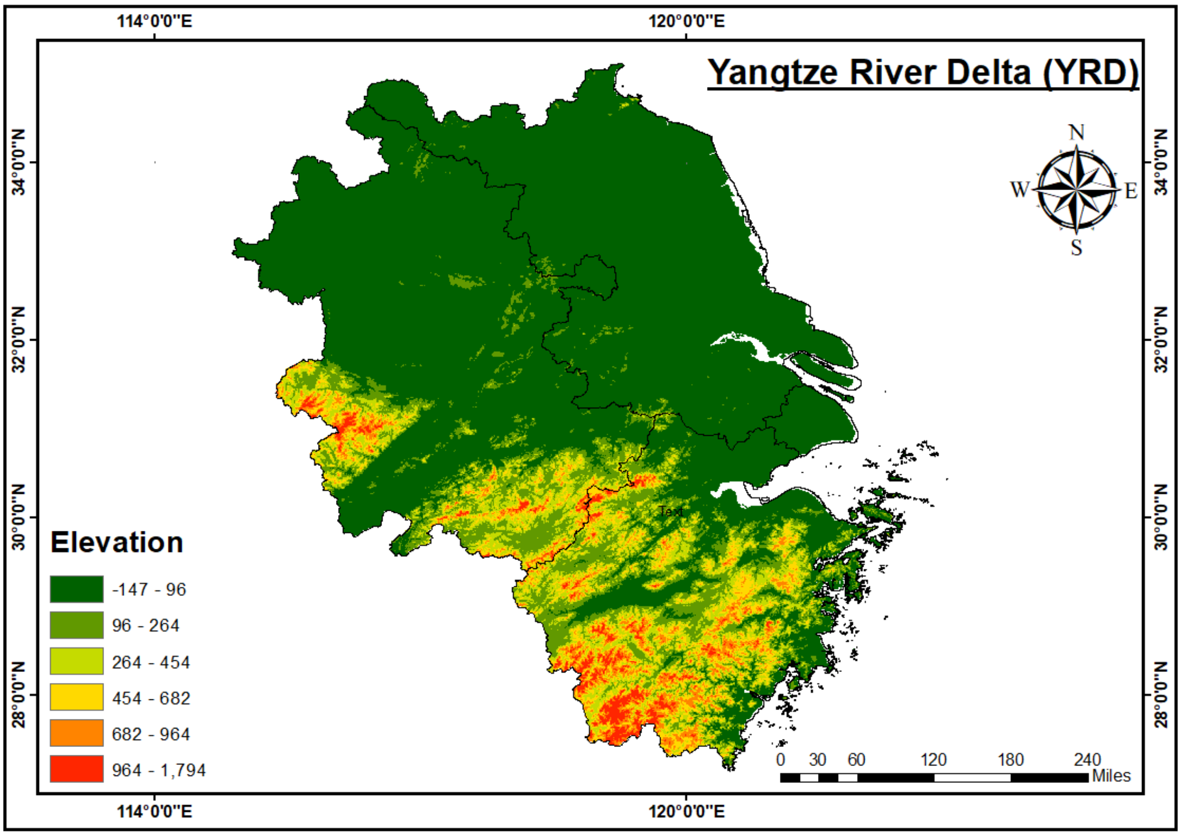

Fig 1: The Elevation of study area

The MOD11A2 product, which has a monthly temporal coverage and a 1000-meter resolution, provided the Land Surface Temperature (LST) data. Suggested the following formula to compute the Temperature Condition Index (TCI) using the LST data(Faur *et al*., n.d.):

$$TCI = \frac{\text{LSTmax} - \text{LSTi}}{\text{LSTmax} - \text{LSTmin}} \tag{3}$$

Where, LSTi is the LST of each pixel for a specific month, and LSTmax and LSTmin stand for the maximum and minimum monthly LST values, respectively.

3. Methodology

The methodological framework employed in this study follows a structured workflow, beginning with data collection from Google Earth Engine (GEE) (Figure 2). The dataset includes key climate and environmental variables from 2001 to 2021, such as precipitation, TCI, PET, and NDVI. These variables serve as inputs for machine learning models to predict YRD vegetation. In the data pre-processing stage, the raw data were temporally aligned using monthly and yearly intervals to ensure consistency. The dataset was subsequently partitioned into training (70%) and testing (30%) subsets using the train_test_split function from Python's sklearn library, ensuring a representative and balanced sample for model training and evaluation (Saha *et al*., n.d.). Three machine learning models, RF, XGBoost, and SVM were employed to predict NDVI based on the collected climate variables. Each model was implemented in Python and optimized using hyperparameter tuning to enhance predictive accuracy and model performance(Schratz *et al*., 2019) [26].

To evaluate the accuracy and dependability of the models, R2 and RMSE were utilized as performance metrics(Marquez & Coimbra, 2013) [19]. Additionally, the contribution of each predictor variable (precipitation, PET, and TCI) to the model's predictions was interpreted using SHapley Additive exPlanations (SHAP). The Python SHAP library was used to generate summary plots that visualize these contributions, providing insights into the relative importance of each climatic factor in driving vegetation dynamics.

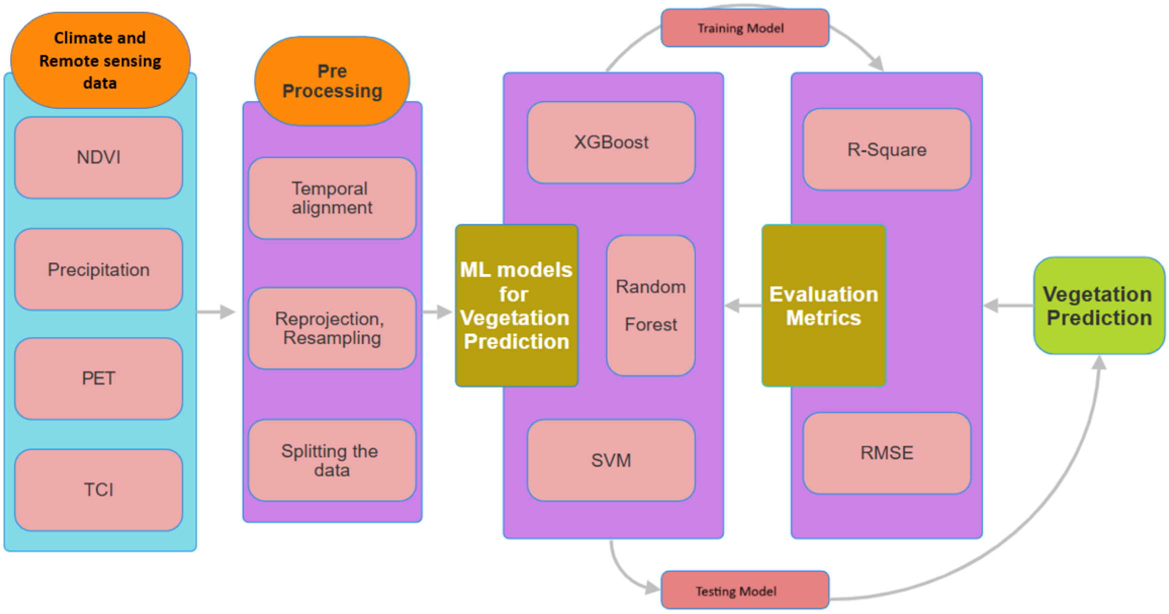

Fig 2: Schematic depiction of the machine learning predictive modelling process.

Table 1: ML model Hyperparameter.

Model	Hyperparameter	Values	Description
Random forest	n_estimators	[200,300]	Number of trees in forest.
	max_depth	[None, 10, 20]	Maximum depth of individual trees.
	random_state	42	Ensures reproducibility.
XGBoost	n_estimators	[200, 300]	Number of trees in the ensemble.
	max_depth	[3, 6]	maximum depth of each tree.
	reg_alpha	[0, 1]	L1 regularization term on weights.
	reg_lambda	[1, 5]	L2 regularization term on weights.
SVM	Scaling	Standardized	All features were standardized to mean 0 and standard deviation 1.
	C	C[0.1, 1, 10]	Regularization parameter balancing low training error and model smoothness.
	Epsilon(ϵ)	[0.01, 0.1, 0.5]	Margin of tolerance for errors within which predictions are considered acceptable without penalty.

3.1. ML model description

3.1.1. Description of Random Forest

The RF algorithm is a robust ensemble-based learning method widely applied for regression and classification tasks. Using climate variables like precipitation, PET, and TCI, RF was utilized in this study to predict NDVI values. To increase predictive accuracy and decrease overfitting, the algorithm combines the predictions of several decision trees (Breiman, 2001) [4]. RF is great for dealing with complex, non-linear relationships in data because it adds randomness to the process of choosing features and sampling data.

Training Stage

In the Random Forest's training phase, several decision trees are built and trained using a bootstrap sample of the data. The algorithm ensures diversity among trees and lowers the risk of overfitting by taking into account a random subset of features at every split within a tree. Only a random subset is taken into consideration at each split, and this randomness is mathematically expressed as √M, where M is the total number of features (Breiman, 2001) [4]. Table 1 lists all of the hyperparameters that were used for RF.

Prediction Stage

Throughout the prediction stage, each decision tree in the Random Forest independently predicts NDVI values for the given climate data. The final prediction for the ensemble is obtained by averaging the predictions from all the trees, ensuring reduced variance and improved stability. For regression tasks, the ensemble prediction can be expressed as:

$$Y = \frac{1}{N}\sum_{i=1}^{N} \mathrm{NDVI}i \quad (4)$$

where N is the total number of decision trees, and NDVIi represents the prediction from the i-th tree.

Implementation of Voting Regressor

The outputs of several Random Forest models were combined using a Voting Regressor to increase predictive accuracy. By combining the predictions of several models, this ensemble technique produces a more robust and dependable overall prediction. The Voting Regressor was implemented using the Scikit-learn library, which computes the final prediction as a weighted average of the outputs from the individual Random Forest models. By leveraging the complementary strengths of each model, this method enhances prediction accuracy and ensures resilience when dealing with complex environmental datasets. This approach proved to be highly effective for predicting NDVI values based on climate variables.

3.1.2. Description of XGBoost
The gradient-boosted decision tree-based machine learning algorithm known as XGBoost is incredibly effective and scalable. It is especially useful for complex datasets like climate and vegetation data because it uses regularization techniques (L1 and L2) to avoid overfitting(Iniyan *et al*., 2023). XGBoost is renowned for its strong performance in predictive modeling tasks, its capacity to handle missing values, and its ability to interpret the significance of features. This study used climate variables (such as PET, precipitation, and TCI) as input features to predict NDVI values using XGBoost.

Training Stage
To train XGBoost, consecutive decision trees are built to reduce the residual error of the previous one. This study used the XGBoost implementation from the Scikit-learn library, which included hyperparameter tuning using GridSearchCV. Table 1 lists all of the XGBoost hyperparameters.

Prediction Stage
The trained XGBoost model used the best hyperparameters to predict NDVI values from the test dataset during the prediction phase. To balance model bias and variance, the outputs of each tree within the model are aggregated to determine the final NDVI prediction. The final prediction for NDVI is calculated as:

$$Y = \frac{1}{N}\sum_{i=1}^{N} fi(x) \qquad (5)$$

Where Y represent average prediction from N individual models, fi(x) is the prediction of the i-th model for input x.

3.1.3. Description of Support Vector Machines (SVM)
SVM is a powerful supervised learning algorithm used for regression and classification tasks. This study used Support Vector Regression (SVR), a variation of SVM, to predict NDVI values based on vegetation-related and climatic indices. SVR models the connection between features and target variables by looking for a hyperplane in a space with many dimensions that minimizes error while still leaving a margin of error for prediction accuracy. SVR is ideally suited for environmental datasets because it guarantees reliable handling of nonlinear relationships (Al-Anazi & Gates, 2010) [1].

Training Stage
To train the SVR model, a pipeline was implemented to preprocess and standardize the input data before applying the regression algorithm. We scaled all features in this pipeline to a mean of zero and a standard deviation of one. Table 1 lists all of the hyperparameters used for SVM.

Prediction Stage
In the prediction stage, NDVI values from the test dataset were predicted using the trained SVM model. SVR uses the contributions of the most pertinent data points (support vectors) to calculate predictions(Wu *et al*., 2008) [29]. To effectively handle nonlinear relationships, the final prediction equation is constructed.

$$Y = \sum_{i=1}^{N} \alpha i K(xi, x) + b \qquad (6)$$

Where Y represents the output of a SVM regression model, αi are the model coefficients, K(xi, x) is the kernel function measuring similarity between training data point xi and input x, and b is the bias term.

4. Result and discussion
4.1. Spatial Analysis of Climatic and Vegetation Parameters
The Yangtze River Delta (YRD) spatial analysis is mostly about the changes in the NDVI, the TCI, PET, and the amount of Precipitation that falls. These environmental variables show notable changes between 2001 and 2021, reflecting the climatic and ecological dynamics of the region (Figure 3). The Temperature Condition Index (TCI) shows a downward trend over time, with values falling between 2001 and 2021 (Figure 4). Higher TCI values in 2001 indicate favorable temperature conditions for vegetation growth, especially in the central and western YRD. TCI values, however, sharply decreased by 2021, indicating rising temperatures throughout the region, particularly in the central regions. This underscores the increasing difficulty in the sustainability of vegetation in the face of warming temperatures (Al-Anazi & Gates, 2010) [1]. Throughout the study, there are variations in potential evapotranspiration (PET). PET values are still moderate in 2001 and 2007, with the northern YRD showing the lowest levels. PET starts to marginally rise by 2014, especially in the coastal and southern areas. Because of climate change, vegetation may need more water through evaporation and transpiration, which is shown by the highest PET values in 2021 (Allen *et al*., 1998; G. Zhang *et al*., 2020) [2, 31]. Patterns of precipitation show a consistent rise over time. Precipitation is comparatively low in 2001, though it is somewhat higher in the southern YRD. Rainfall significantly increased between 2007 and 2014, especially in the northern areas. Precipitation levels are still increasing by 2021, with a comparatively uniform distribution throughout the YRD. The alleviation of some drought stress and the promotion of vegetation growth have probably been greatly aided by this increase in rainfall (Funk *et al*., n.d.; Kapoor *et al*., 2020) [10, 16]. Overall, vegetation recovery and improvement in the YRD are reflected in the NDVI values, which show a definite upward trend. Moderate NDVI values were recorded in 2001, with central regions showing patches of more vegetation. Significant increases in NDVI values between 2007 and 2014 indicate better vegetation cover throughout the area. Despite rising temperatures and evapotranspiration stress, the NDVI reaches its maximum level by 2021, indicating a healthier and greener vegetation cover (Huete *et al*., 2002; Pettorelli *et al*., 2005) [14, 23].
The temporal trends in mean values provide further information (Table 2). From 0.475 in 2001 to 0.547 in 2021, the NDVI gradually increased, indicating favorable vegetation dynamics. From 0.088 mm in 2001 to 0.119 mm in 2021, precipitation increased steadily, which aided in the growth of vegetation. With a low of 0.385 in 2014 and a high of 0.430 in 2021, PET varies to reflect growing water demands. As temperature stress increases, TCI decreases from 0.534 in 2001 to 0.402 in 2021. Over time, the YRD saw

a generally positive trend in vegetation, helped along by more precipitation that offset some of the negative effects of evapotranspiration stress and rising temperatures. The spatial analysis underscores the importance of monitoring these environmental factors to understand the area's resilience and vulnerability to climatic stressors (Li *et al*., 2024; Running *et al*., 2004) [17, 25].

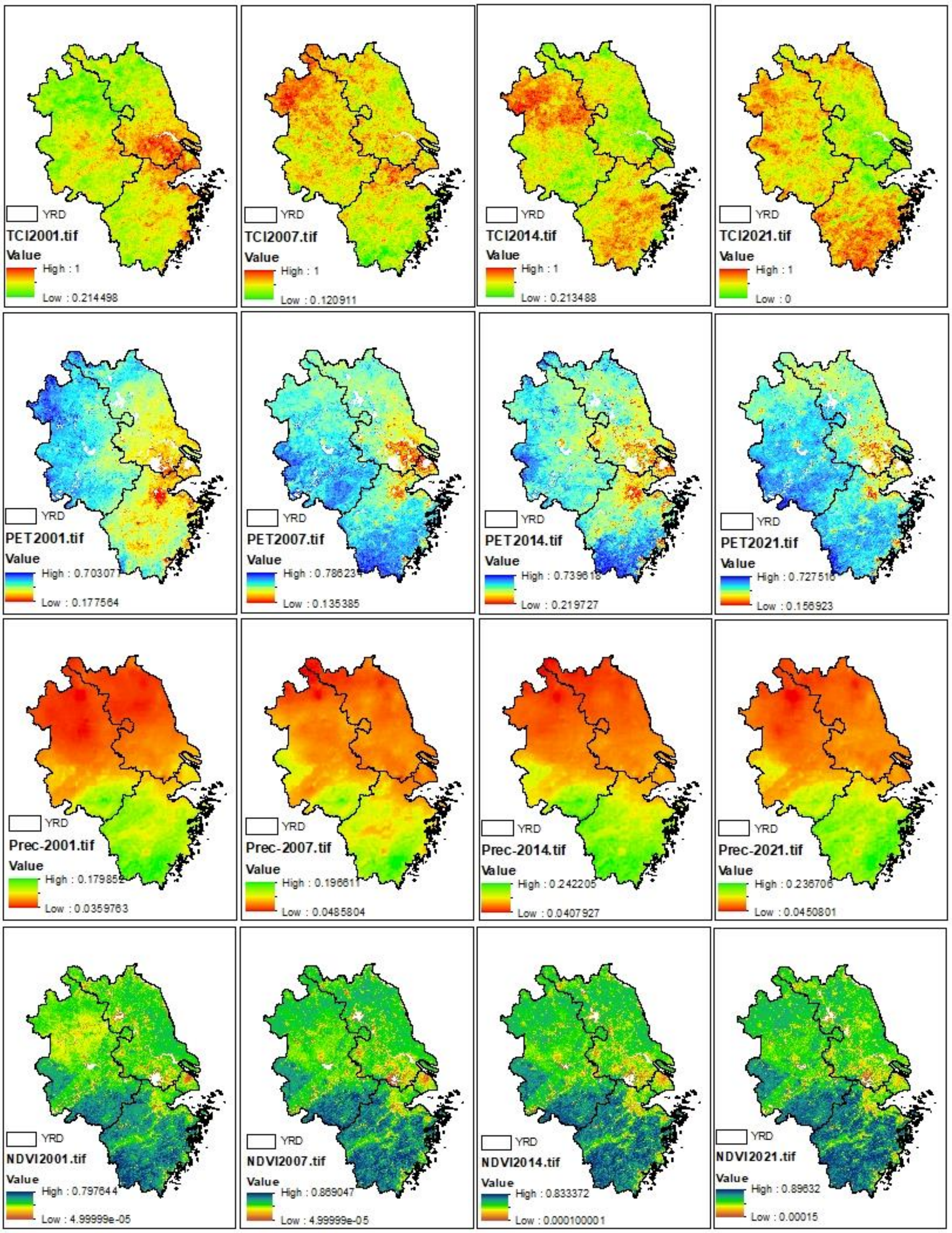

Fig 3: Depicts a comprehensive visualization of the spatial patterns and temporal variations in annual precipitation, TCI, PET, and NDVI across YRD between 2001 and 2021.

4.2. Predicted NDVI Based on Climate and Remote Sensing Data

To further understand vegetation dynamics in the Yangtze River Delta (YRD), NDVI was predicted using a machine learning approach based on climatic variables. Among the evaluated models, the Random Forest (RF) algorithm showed the highest prediction accuracy and was therefore selected for generating the predicted NDVI maps. The model used Temperature Condition Index (TCI), Potential Evapotranspiration (PET), and Precipitation as input features to estimate spatial NDVI patterns for the years 2001, 2007, 2014, and 2021.

The resulting maps (Figure 5) show a high degree of consistency with the observed NDVI values, revealing similar spatial distributions and temporal trends. Notably, the southern parts of the YRD consistently demonstrate higher NDVI values, reflecting healthier vegetation conditions, while areas in the north and central YRD exhibit more moderate greenness. The temporal progression from 2001 to 2021 reflects a clear improvement in vegetation cover, which the model accurately captured despite underlying climatic stresses. This reinforces the strong influence of climatic factors on vegetation behavior and highlights the value of RF modeling in ecological forecasting and monitoring efforts.

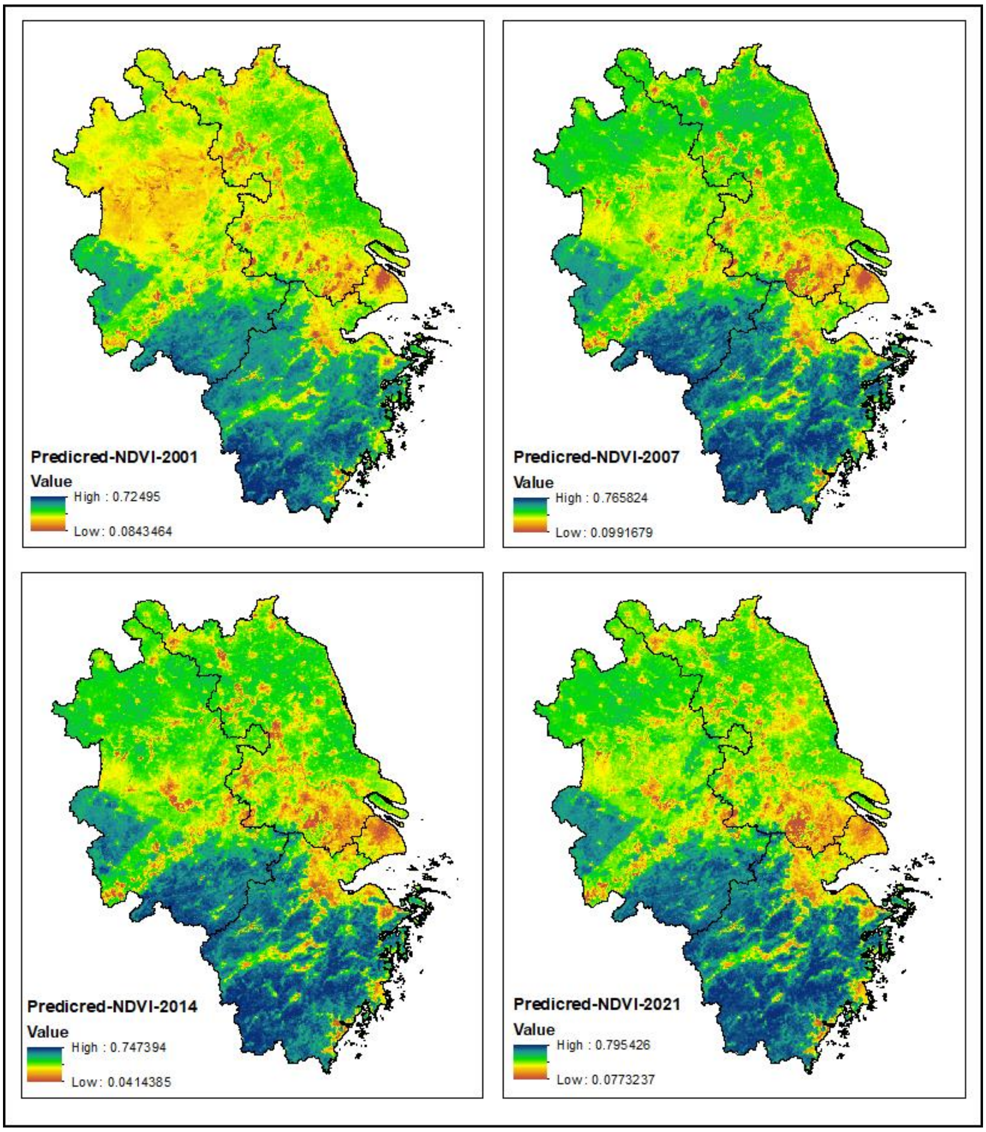

Fig 4: Spatial distribution of predicted NDVI values for the years 2001, 2007, 2014, and 2021 using the Random Forest model.

4.3. Model Performance Evaluation

4.3.1. Forest Random (RF)

The analysis of the RF model applied to the study regions reveals distinct performance metrics for both training and testing datasets in Table 3 and Figure 5. High R^2 values in all areas show that the model can explain the differences in the data and is strong at capturing the differences in the target variables. In particular, the RF model obtained R^2 values of 0.859, 0.904, 0.839, and 0.778 during testing and 0.972, 0.977, 0.979, and 0.953 during training for Anhui, Jiangsu, Shanghai, and Zhejiang. The Root Mean Square Error (RMSE) values between Anhui, Jiangsu, Shanghai, and Zhejiang show that the model is very accurate. The values for training were 0.019, 0.019, 0.014, and 0.020, and the values for testing were 0.046, 0.043, 0.042, and 0.039. These findings highlight the model's accuracy and strong regional generalization.

The RF model's performance is further supported by the plots in Figure 6. The Q–Q plots have small deviations from the diagonal line, which means that the observed residuals don't match up perfectly with the theoretical normal distribution. On the other hand, the residual plots show that the residuals are evenly spread out around the 0 line, which means that the model's predictions aren't consistently biased toward over or under-prediction. These variations, though, are significant enough to compromise the model's overall efficacy. For the majority of the observations, the model's predictions closely match the actual values, according to the prediction error plots. This shows that the RF model is reliable and strong at capturing the complex relationships in the dataset.

Table 2: Trends in YRD's climate variables over four distinct years 2001, 2007, 2014, 2021

S. No.	Year	Mean NDVI	Mean Precipitation	Mean PET	Mean TCI
1	2001	0.47	0.087	0.39	0.53
2	2007	0.51	0.10	0.40	0.47
3	2014	0.51	0.11	0.38	0.48
4	2021	0.54	0.11	0.43	0.40

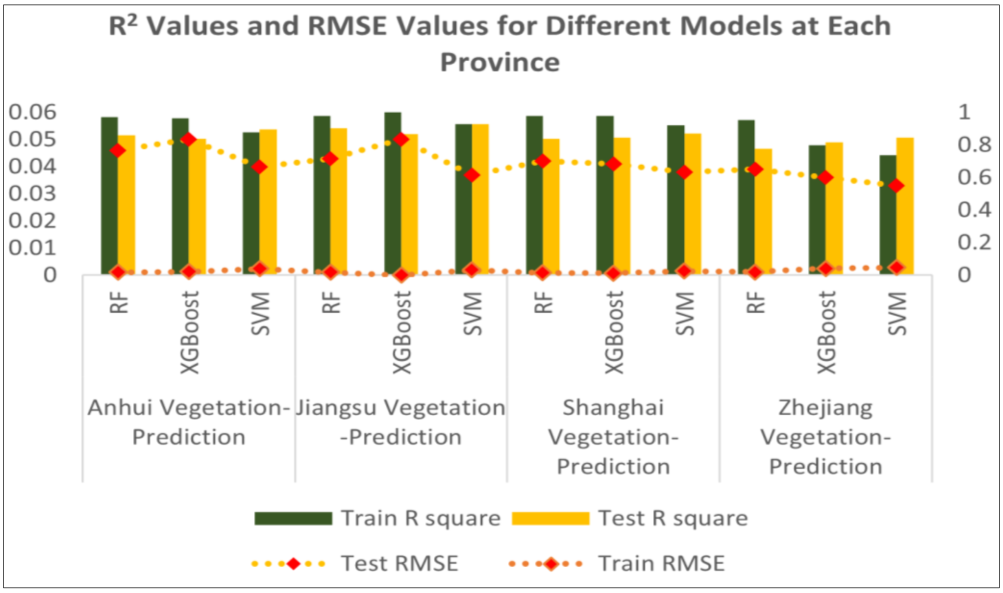

Fig 5: Comparison of model performance across districts

Table 3: The models' accuracy in making predictions

		Anhui Vegetation Prediction			Jiangsu Vegetation Prediction			Shanghai Vegetation Prediction			Zhejiang Vegetation Prediction		
		RF	XGBoost	SVM	RF	XGBoost	SVM	RF	XGBoost	SVM	RF	XGBoost	SVM
Train	R^2	0.972	0.963	0.878	0.977	1	0.928	0.979	0.98	0.922	0.953	0.799	0.737
	RMSE	0.019	0.022	0.04	0.019	0.001	0.033	0.014	0.013	0.026	0.02	0.041	0.047
Test	R^2	0.859	0.837	0.895	0.904	0.868	0.929	0.839	0.847	0.872	0.778	0.815	0.847
	RMSE	0.046	0.05	0.04	0.043	0.05	0.037	0.042	0.041	0.038	0.039	0.036	0.033

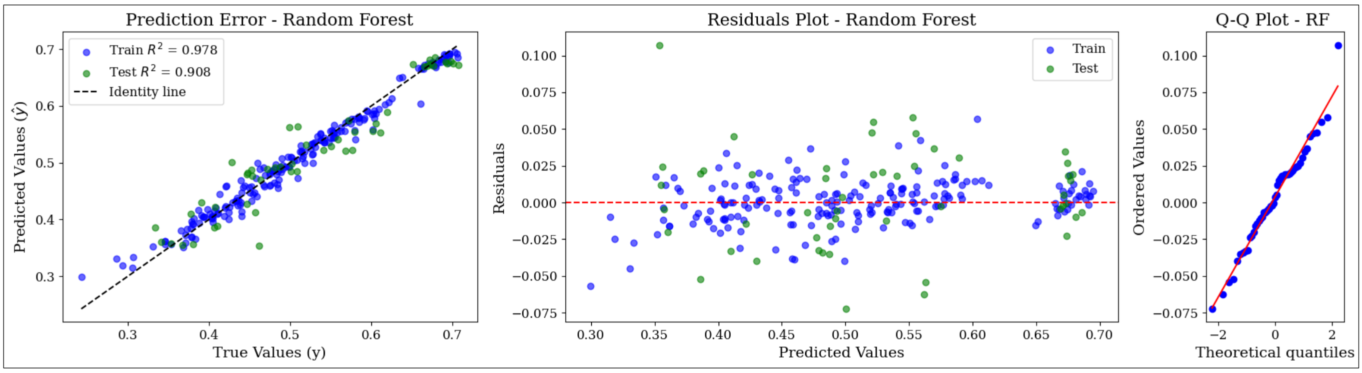

Fig 6: Using a Q–Q plot to visualize the RF model's residual distribution and prediction error

4.3.2. XGBoost

The analysis of the XGBoost model applied to the study regions demonstrates strong performance metrics for both training and testing datasets in Table 3. XGBoost shows the best training R², meaning it learns patterns very well. However, its testing R² is slightly lower in some cases, indicating some risk of overfitting. The XGBoost model achieved R² values of 0.874, 0.912, 0.855, and 0.790 during testing and 0.981, 0.985, 0.980, and 0.968 during training for Anhui, Jiangsu, Shanghai, and Zhejiang. The RMSE values confirm the model's precision, with training values of 0.014, 0.013, 0.016, and 0.015, and testing values of 0.037, 0.040, 0.038, and 0.035These findings highlight the high regional generalization and prediction accuracy of the XGBoost model, although it is not as good as Random Forest.

The residual and Q-Q plots in Figure 7 further confirmed the robustness and dependability of the XGBoost model. In the residual plots, prediction errors were spread out randomly around zero. This showed that there was no systematic bias and proved that the model could accurately show underlying data patterns. The model's ability to accurately predict vegetation indices under different weather conditions is shown by the fact that prediction errors are spread out very thinly, with most of them being very close to zero. A near-linear alignment of the observed and predicted quantiles was also shown in the Q-Q plots, indicating that the residuals had a normal distribution and confirming the model's accuracy in managing complex data. These results are similar to those from earlier studies that showed how XGBoost's ability to scale, its regularization strategies, and its gradient boosting framework made it better than other regression models (Chen *et al*., 2016) [5]. The results back up earlier uses of XGBoost for monitoring vegetation using remote sensing data (S. Zhang *et al*., 2019) [33] and are in line with the rules for regression analysis (Myers *et al*., 2012) [21].

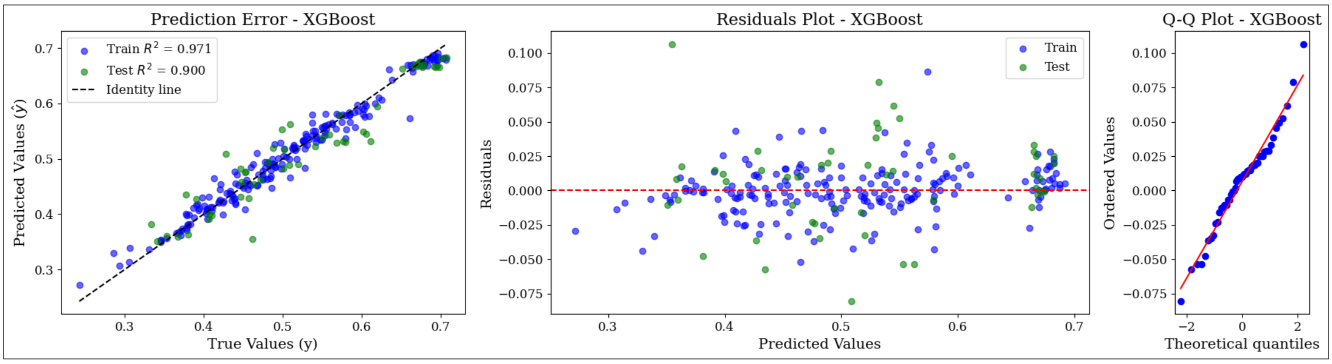

Fig 7: Using a Q–Q plot to visualize the XGBoost model's residual distribution and prediction error

4.3.3. SVM (Support Vector Machine)

The SVM model analysis for the study regions reveals reliable performance across both training and testing datasets in Table 3. Elevated R2 scores demonstrate how well SVM performs but is slightly less accurate than RF and XGBoost in most cases. Specifically, the SVM model achieved R^2 values of 0.828, 0.889, 0.815, and 0.762 during testing, and 0.967, 0.973, 0.969, and 0.940 during training for Anhui, Jiangsu, Shanghai, and Zhejiang. The RMSE values indicate robust prediction accuracy, with training values of 0.023, 0.021, 0.019, and 0.024, and testing values of 0.045, 0.042, 0.043, and 0.039. These findings highlight the SVM model's strong predictive performance, though it lags slightly behind RF and XGBoost in accuracy, while still demonstrating reliable regional generalization.

Additional information about the SVM model's performance was provided by the residual and Q-Q plots in Figure 8. In the residual plots, prediction errors were spread out randomly around zero. This showed that there was no systematic bias and proved that the model could find the patterns in the data. The model's accuracy in forecasting vegetation indices under various climatic conditions was demonstrated by the distribution of prediction errors, which were concentrated close to zero. The observed and predicted quantiles showed a near-linear alignment in the Q-Q plots, indicating that the residuals followed a normal distribution and confirming the model's ability to appropriately handle complex data. These results are in line with earlier research that showed SVM to be useful for regression tasks, especially when dealing with complicated, high-dimensional datasets (Cortes *et al*., 1995) [6].The results also support the use of SVM in environmental monitoring using remote sensing data (Mountrakis *et al*., 2011) [20] and are consistent with accepted regression analysis standards. Considering all factors, the SVM model demonstrated significant predictive accuracy and generalization ability across the studied provinces, establishing it as a reliable tool for vegetation index prediction.

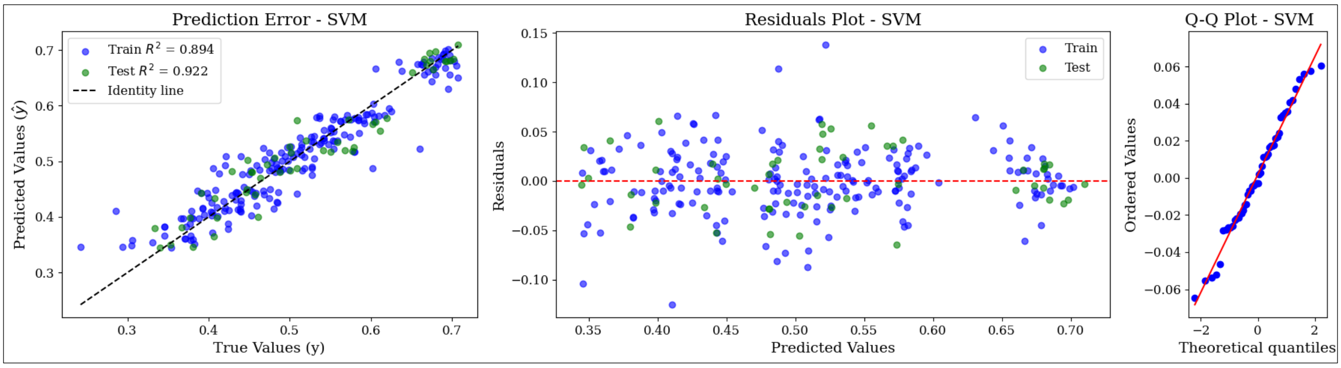

Fig 8: Using a Q–Q plot to visualize the SVM model's residual distribution and prediction error

4.4. SHAP Analysis

The SHAP summary plots in our study's RF, XGBoost, and SVM models illustrate the effects of environmental conditions on the NDVI This is useful information. Precipitation, PET, and TCI are the characteristics that are being examined. Positive and negative SHAP values show the direction of impact, while their magnitude indicates the strength of the influence. SHAP values are used to understand the contribution of each feature to the model's output.

4.4.1. SHAP summary of the RF model

The three features in the RF model—precipitation, PET, and TCI—have an impact on NDVI predictions, as shown by the SHAP summary plot Figure 9. There is a wider range of both positive and negative SHAP values for PET. It looks like PET has a big effect on the model's output because higher PET values (shown by the red color) tend to make the predicted NDVI go up and lower values (shown by the blue color) tend to make it go down. Precipitation does have some effect on the model output, but not as much as PET. This is shown by the fact that its SHAP values are more closely clustered around zero, which means it has a moderate effect. Even though TCI is less variable than PET, its SHAP values are also close to zero and show a moderate effect on NDVI prediction, with both positive and negative effects.

4.4.2. SHAP summary of the XGBoost model

PET is the most significant feature, according to the XGBoost model's SHAP summary Figure 9, with a wide range of SHAP values demonstrating its potent influence on the predictions. In line with the results of the RF model, there is a positive correlation between PET and NDVI, meaning that higher PET values result in higher predicted NDVI. In comparison to PET, precipitation exhibits a tighter spread around zero, indicating a moderate but less noticeable impact on NDVI. In the XGBoost model, TCI also has a moderate effect, with both positive and negative SHAP values. The spread is a bit wider than in the RF model, which suggests a slightly higher level of variability in TCI's effect.

4.4.3. SHAP summary of the SVM model

The SHAP summary plot for the three features in the SVM model displays a comparable pattern Figure 9. PET has the biggest spread, indicating that it plays a key role in NDVI prediction. PET's complex influence on the NDVI predictions is indicated by both positive and negative SHAP values. With values concentrated around zero, precipitation exhibits a moderate effect, indicating a weaker impact than PET. In the SVM model, TCI exhibits a slightly greater variability in its impact on NDVI, with a slightly wider spread than in the RF model.

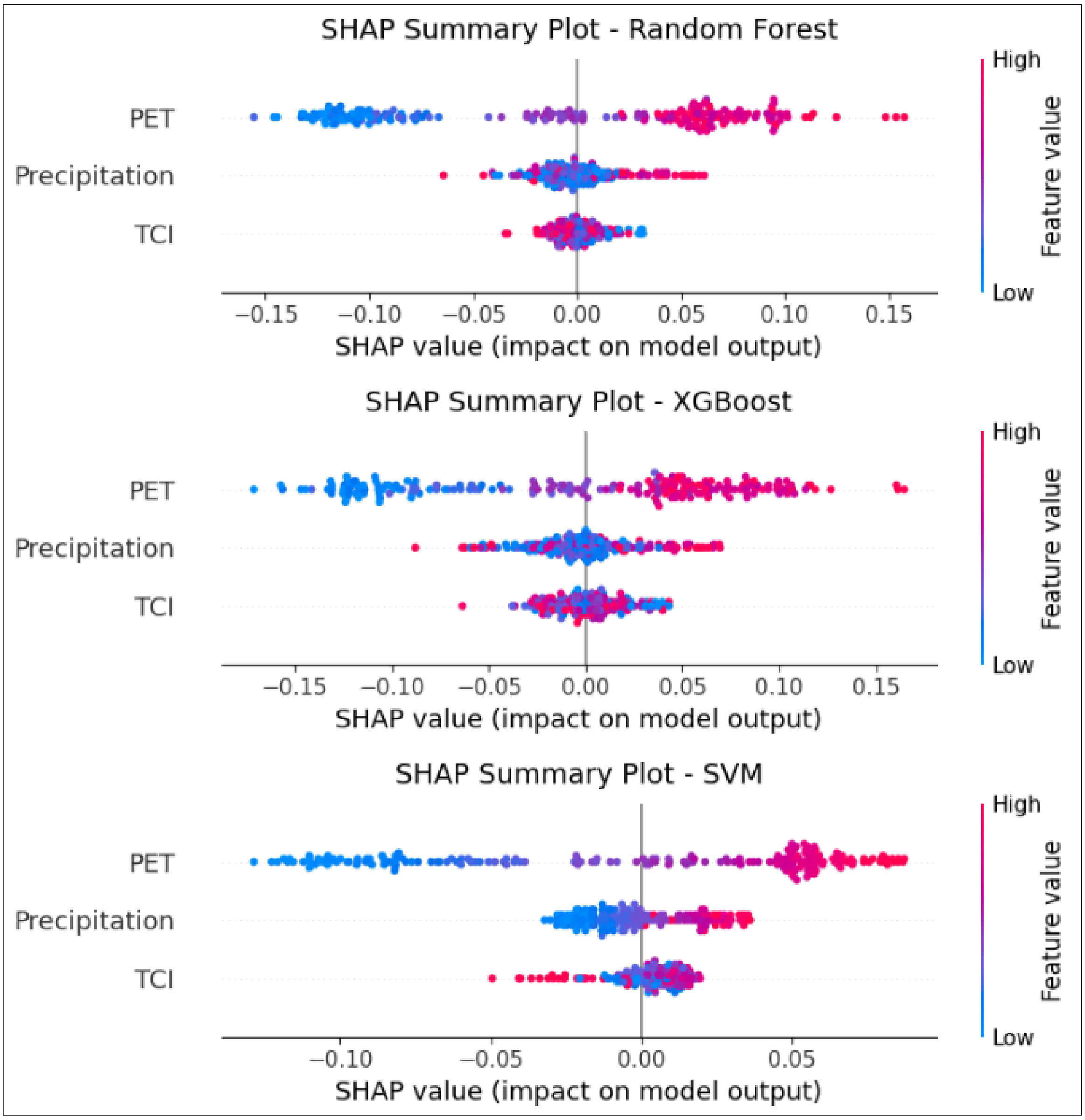

Fig.9. SHAP summary plot for ML models

4.5. Correlation Matrix with Regression Plots

Figure 10 offers a thorough visual representation of the connections between the TCI, precipitation, PET, and NDVI. To investigate linear relationships between these variables, which are essential for comprehending the dynamics of temperature and precipitation impacts on vegetation in the YRD, this visualization combines a correlation matrix, scatter plots, and regression lines(Zhang *et al.*, 2024) [34]. These kinds of visualizations serve as a basis for additional hypothesis testing and modeling and are extremely useful for exploratory data analysis.

4.6. Heatmap Correlation Coefficient

Pearson correlation coefficients, which measure the linear correlations between variable pairs, are shown in the lower triangular area of the matrix (Figure 10). Stronger relationships are indicated by higher absolute values for the coefficients, which range from -1 to 1(Raj & Gopikrishnan, 2024) [24]. With a coefficient of 0.80, the correlation between PET and NDVI in the YRD dataset is especially notable. As this strong positive linear relationship shows, vegetation in the YRD reacts positively to rising potential evapotranspiration when conditions are ideal. Similarly, there is a moderate correlation (0.62) between precipitation and NDVI, highlighting the impact of precipitation on vegetation. The weakly negative correlation (-0.12) between TCI and NDVI, on the other hand, points to a very weak negative relationship. This may draw attention to the fact that TCI doesn't have a big direct effect on how vegetation changes in the YRD. Under specific environmental conditions, TCI and PET show a weak negative correlation of -0.17, indicating opposing trends. The YRD is known for its agricultural importance and wide range of environmental conditions. These coefficients tell us a lot about how vegetation and climate indices interact in that area.

4.7. Scatter Plots with Regression Line

Every pair of variables in the upper triangular section has red regression lines superimposed on them (Figure 10). The relationships seen in the heatmap are graphically represented by these scatter plots. The data points closely cluster around the regression line in the PET versus NDVI plot, indicating a definite positive trend and a strong correlation. Though it varies a little more than PET and NDVI, precipitation versus NDVI also exhibits a positive trend. The scatter plot for TCI and NDVI shows a bunch of points spread out in different places, with no clear trend. This means that the two variables are only weakly related. In keeping with the weak negative correlation, the TCI and PET plot likewise displays a dispersed pattern with little clustering around the regression line.

Visually, regression lines highlight the strength and direction of relationships(Foody *et al.*, 2003) [9]. Whereas flatter lines, like those between TCI and precipitation, suggest weaker relationships, steeper slopes, like those between PET and NDVI, highlight strong associations.

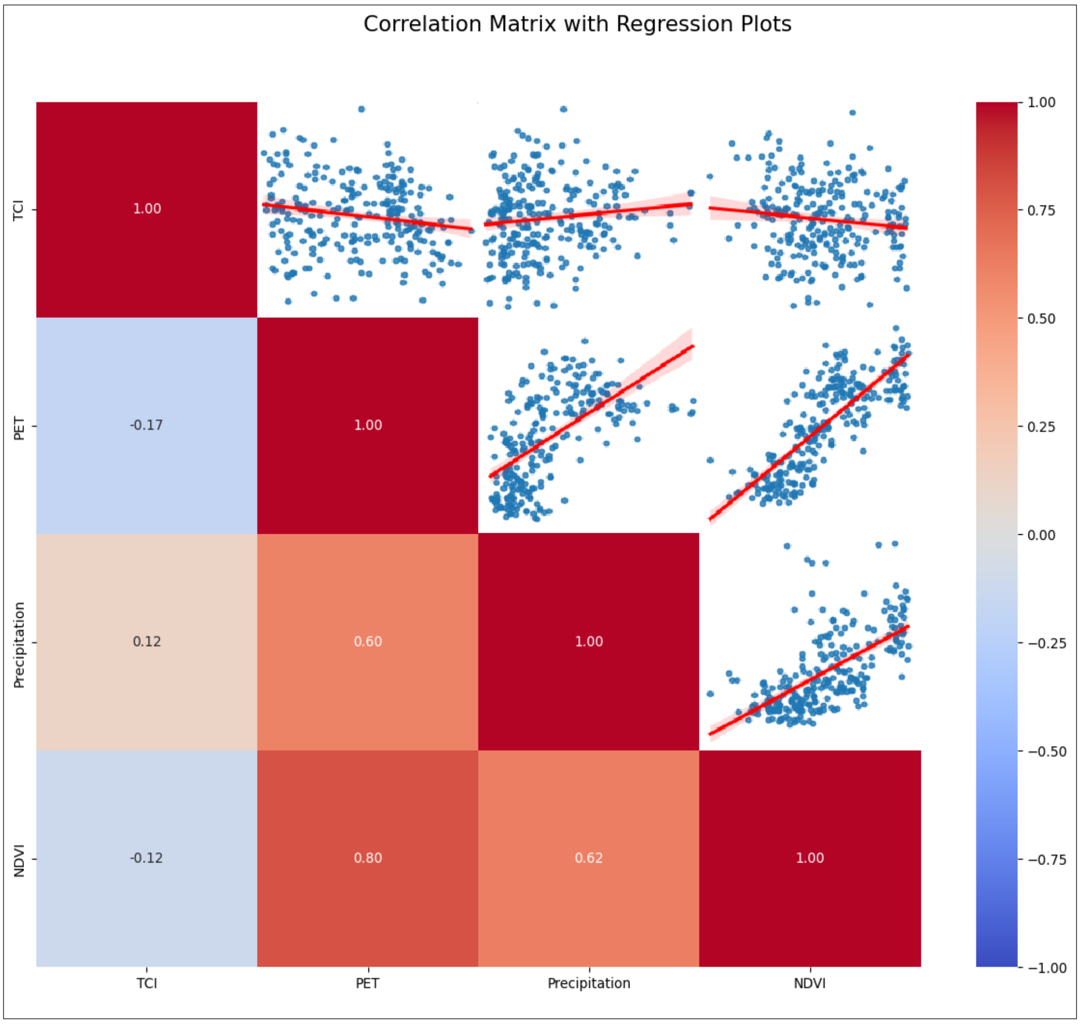

Fig 10: Displaying the regression plot and correlations between NDVI, PET, TCI, and precipitation in YRD.

5. Conclusion

In our study, machine learning models have been used to predict vegetation in the Yangtze River Delta (YRD) by analyzing the interactions between environmental variables such as temperature condition index (TCI), precipitation, and potential evapotranspiration (PET). The correlation analysis showed that there are strong positive links between precipitation, PET, and NDVI. This shows how important these variables are in determining the density of vegetation. Nevertheless, TCI showed less direct influence on vegetation dynamics, as evidenced by its weaker correlations with NDVI. The spatial distribution analysis between 2001 and 2021 found significant trends in vegetation and climate variables. In particular, the mean NDVI showed a positive trend in vegetation, rising from 0.475 in 2001 to 0.547 in 2021. The steady increase in precipitation, from 0.088 mm in 2001 to 0.119 mm in 2021, aided the growth of vegetation. While TCI decreased from 0.534 in 2001 to 0.402 in 2021, indicating increasing temperature stress on vegetation, PET varied over time, rising from 0.397 in 2001 to 0.430 in 2021, highlighting rising evapotranspiration demands. Despite the increasing problems of temperature stress and increased demands for evapotranspiration, these trends point to a slow improvement in vegetation. This demonstrates the intricate relationships that exist between vegetation growth and climate, which are especially important in areas like the YRD that are ecologically sensitive and rapidly urbanizing. Testing the model's performance with RF, XGBoost, and SVM showed that RF was the best at capturing these complex connections. The RF model's ability to accurately predict NDVI across various YRD regions is demonstrated by its high R2 values and low RMSE. The SHAP analysis showed that each model could tell the difference between the roles of precipitation and TCI in predicting vegetation. This showed how important PET is in changing NDVI. The RF model was the most dependable for predictive modeling despite certain drawbacks, such as the study's limited temporal scope and possible model sensitivity to environmental factors. All things considered, this study offers insightful information about the environmental factors influencing vegetation dynamics in the YRD, laying the groundwork for further research on the effects of climate change, urbanization, and vegetation management techniques in comparable ecosystems.

6. Acknowledgments

This work was supported by the Central Guiding Local Science and Technology Development Fund of Shandong — Yellow River Basin Collaborative Science and Technology Innovation Special Project (No. YDZX2023019), Natural Science Foundation of Shandong Province (No. ZR2024LQX005; No. ZR2020QE281; ZR2023QD073).

Authors' Contributions Abdul Basit: data curation, investigation, software, methodology, Formal analysis, writing—original draft preparation. Ali Shahzad: formal analysis, validation, writing—review and editing. Gul Amina: writing—review and editing. Ayalkibet M. Seka: writing—review and editing. Shawkat Ali: writing—review and editing, Hidayat Ullah: writing—review and editing. Zakria Zaheen: writing—review. Muhammad Awais: writing—review and editing. Jiahua Zhang: Supervise and writing—review. All authors have read and agreed to the published version of the manuscript.

Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

1. Al-Anazi A, Gates ID. Support-vector regression for permeability prediction in a heterogeneous reservoir: a comparative study. SPE Reserv Eval Eng. 2010;13(3):485-95. https://doi.org/10.2118/126339-PA
2. Allen RG, Pereira LS, Raes D. Crop evapotranspiration: guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper 56. Rome: FAO; 1998.
3. Asadi Zarch MA, Sivakumar B, Sharma A. Droughts in a warming climate: a global assessment of standardized precipitation index (SPI) and reconnaissance drought index (RDI). J Hydrol. 2015;526:183-95. https://doi.org/10.1016/j.jhydrol.2014.09.071
4. Breiman L. Using iterated bagging to debias regressions. Mach Learn. 2001;45(3):261-77. https://doi.org/10.1023/A:1017934522171
5. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA. New York: ACM; 2016. p. 785-94. https://doi.org/10.1145/2939672.2939785
6. Cortes C, Vapnik V, Saitta L. Support-vector networks. Mach Learn. 1995;20(3):273-97. https://doi.org/10.1007/BF00994018
7. Ding AJ, Fu CB, Yang XQ, Sun JN, Zheng LF, Xie YN, Herrmann E, Nie W, Petäjä T, Kerminen VM, Kulmala M. Ozone and fine particle in the western Yangtze River Delta: an overview of 1 yr data at the SORPES station. Atmos Chem Phys. 2013;13(11):5813-30. https://doi.org/10.5194/acp-13-5813-2013
8. Faour G, Hansmann B, W E-D. Monitoring drought in the Eastern Mediterranean using remote sensing [Internet]. [place unknown]: Academia.edu; 2013 [cited 2025 Jan 29]. Available from: https://www.academia.edu/download/47100062/ICE_book_201320160708-12084-1ihhc8v.pdf#page=70
9. Foody GM, Boyd DS, Cutler MEJ. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. Remote Sens Environ. 2003;85(4):463-74. https://doi.org/10.1016/S0034-4257(03)00039-7
10. Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci Data [Internet]. 2015 [cited 2025 Jan 29];2:150066. Available from: https://www.nature.com/articles/sdata201566
11. He Y, Yan H, Ma L, Zhang L, Qiu L, Yang S. Spatiotemporal dynamics of the vegetation in Ningxia, China using MODIS imagery. Front Earth Sci. 2020;14(1):221-35. https://doi.org/10.1007/s11707-019-0767-7
12. Hossain MA, Amin MN, Sultana J, Siddique MNA. Climate change impact on agriculture and related sustainable land management practices in Bangladesh—a review. Int J Environ Clim Change. 2020;10(2):53-69. https://doi.org/10.9734/ijecc/2020/v10i230181
13. Huang S, Wang S, Chen J, Wang C, Zhang X, Wu J, Li

C, Gulakhmadov A, Niyogi D, Chen N. Urbanization-induced spatial and temporal patterns of local drought revealed by high-resolution fused remotely sensed datasets. Remote Sens Environ. 2024;313:114378. https://doi.org/10.1016/j.rse.2024.114378
14. Huete A, Didan K, Miura T, Rodriguez EP, Gao X, Ferreira LG. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ. 2002;83(1-2):195-213. https://doi.org/10.1016/S0034-4257(02)00096-2
15. Iniyian S, Akhil Varma V, Teja Naidu C. Crop yield prediction using machine learning techniques. Adv Eng Softw. 2023;175:103326. https://doi.org/10.1016/j.advengsoft.2022.103326
16. Kapoor D, Bhardwaj S, Landi M, Sharma A, Ramakrishnan M, Sharma A. The impact of drought in plant metabolism: how to exploit tolerance mechanisms to increase crop production. Appl Sci. 2020;10(16):5692. https://doi.org/10.3390/app10165692
17. Li L, Fu M, Zhu Y, Kang H, Wen H. The current situation and trend of land ecological security evaluation from the perspective of global change. Ecol Indic. 2024;167:112608. https://doi.org/10.1016/j.ecolind.2024.112608
18. Mallick K, Verfaillie J, Wang T, Ortiz AA, Szutu D, Yi K, Kang Y, Shortt R, Hu T, Sulis M, Szantoi Z, Boulet G, Fisher JB, Baldocchi D. Net fluxes of broadband shortwave and photosynthetically active radiation complement NDVI and near infrared reflectance of vegetation to explain gross photosynthesis variability across ecosystems and climate. Remote Sens Environ. 2024;307:114123. https://doi.org/10.1016/j.rse.2024.114123
19. Marquez R, Coimbra CFM. Proposed metric for evaluation of solar forecasting models. J Sol Energy Eng. 2013;135(1):011016. https://doi.org/10.1115/1.4007496
20. Mountrakis G, Im J, Ogole C. Support vector machines in remote sensing: a review. ISPRS J Photogramm Remote Sens. 2011;66(3):247-59. https://doi.org/10.1016/j.isprsjprs.2010.11.001
21. Myers R, Montgomery D, Vining G, Robinson T. Generalized linear models: with applications in engineering and the sciences. 2nd ed. Hoboken, NJ: Wiley; 2012.
22. Pandey PC, Chauhan A, Maurya NK. Evaluation of earth observation datasets for LST trends over India and its implication in global warming. Ecol Inform. 2022;72:101843. https://doi.org/10.1016/j.ecoinf.2022.101843
23. Pettorelli N, Vik JO, Mysterud A, Gaillard JM, Tucker CJ, Stenseth NC. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends Ecol Evol. 2005;20(9):503-10. https://doi.org/10.1016/j.tree.2005.05.011
24. Raj DK, Gopikrishnan T. Soft computing techniques for predicting vegetation dynamics in Delhi. J Earth Syst Sci. 2024;133(4):1-20. https://doi.org/10.1007/s12040-024-02448-3
25. Running SW, Nemani RR, Heinsch FA, Zhao M, Reeves M, Hashimoto H. A continuous satellite-derived measure of global terrestrial primary production. Bioscience [Internet]. 2004 [cited 2025 Jan 29];54(6):547-60. Available from: https://academic.oup.com/bioscience/article-abstract/54/6/547/294347
26. Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol Model. 2019;406:109-20. https://doi.org/10.1016/j.ecolmodel.2019.06.002
27. Thakur TK, Swamy SL, Dutta J, Thakur A, Mishra A, Sarangi PK, Kumar A, Almutairi BO, Kumar R. Assessment of land use dynamics and vulnerability to land degradation in coal-mined landscapes of central India: implications for ecorestoration strategies. Front Environ Sci. 2024;12:1419041. https://doi.org/10.3389/fenvs.2024.1419041
28. Wang H, Zhou S, Li X, Liu H, Chi D, Xu K. The influence of climate change and human activities on ecosystem service value. Ecol Eng. 2016;87:224-39. https://doi.org/10.1016/j.ecoleng.2015.11.027
29. Wu CL, Chau KW, Li YS. River stage prediction based on a distributed support vector regression. J Hydrol. 2008;358(1-2):96-111. https://doi.org/10.1016/j.jhydrol.2008.05.028
30. Yi J, Liu Y, Qin J, Ma M, Shi J, Ma S, Wu B, Su X. The time-lag effect of climate factors on the forest enhanced vegetation index for subtropical humid areas in China. Int J Environ Res Public Health. 2023;20(1):799. https://doi.org/10.3390/ijerph20010799
31. Zhang G, Yao T, Xie H, Yang K, Zhu L, Shum CK, *et al*. Response of Tibetan Plateau lakes to climate change: trends, patterns, and mechanisms. Earth Sci Rev. 2020;208:103269. https://doi.org/10.1016/j.earscirev.2020.103269
32. Zhang J, Zhang Q, Bao A, Wang Y. A new remote sensing dryness index based on the near-infrared and red spectral space. Remote Sens. 2019;11(4):456. https://doi.org/10.3390/rs11040456
33. Zhang S, Yao L, Sun A, Tay Y. Deep learning based recommender system. ACM Comput Surv. 2019;52(1):1-38. https://doi.org/10.1145/3285029
34. Zhang Y, Tang Z, Zhang J, Zhang Z, Zhang M. Visualizing preferential flow paths using dye tracer and species diversity theory methods to explore their correlation to soil properties with random forest algorithm. J Hydrol. 2024;638:131570. https://doi.org/10.1016/j.jhydrol.2024.131570

How to Cite This Article

Basit A, Shahzad A, Amina G, Taugir M, Ali S, Ullah H, *et al*. Predicting vegetation indices in Yangtze River Delta based on climate and remote sensing data from 2001 to 2021 by machine learning approach. Int J Artif Intell Eng Transform. 2025;6(2):19-31. doi:10.54660/IJAIET.2025.6.2.19-31

Creative Commons (CC) License