Iran Agricultural Research (2026) 45(Special Issue) 45-54

Shiraz University

Iran Agricultural Research

Journal homepage: https://iar.shirazu.ac.ir/?lang=en

Research Article

An explainable machine learning framework to analyze the influence of environmental drivers on vegetation condition: Applications in agricultural drought assessment

Amirhossein Haddadi*, Behnam Rahimi

Department of Civil Engineering, Sharif University of Technology, Tehran, I. R. Iran

ARTICLE INFO

Keywords:
ANOVA
Remote sensing
SHAP analysis
Vegetation condition
XGBoost

Received: 09 October 2025
Revised: 11 February 2026
Accepted: 14 February 2026

ABSTRACT- Effective crop management depends on a clear understanding of the interactions among environmental variables and their combined effects on plant health. This study aims to evaluate the influence of environmental drivers—such as precipitation, air temperature, solar radiation, and aerosol optical depth (AOD)—on vegetation condition by integrating analysis of variance (ANOVA) within a multiple linear regression (MLR) framework and machine learning approach, SHAP (Shapley Additive exPlanations) on XGBoost model, using satellite and reanalysis data. Monthly mean values of multi-source remote sensing data were analyzed for the Shushtar sub-basin from 2001 to 2024, including ERA5-Land meteorological variables, AOD from Terra and Aqua satellites, and Landsat-derived NDVI used to compute the Vegetation Condition Index (VCI). The final results show that XGBoost outperformed MLR, achieving a higher accuracy ($R^2 = 0.86$, RMSE = 9.32) compared to MLR ($R^2 = 0.29$, RMSE = 19.06). The results of the MLR based on fitted coefficients indicate that volumetric soil water and precipitation exert the strongest linear influence on VCI, while wind speed shows non-significant effect. In addition, the ANOVA results identify AOD ($F = 20.11$, $P < 0.01$) and surface pressure ($F = 18.19$, $P < 0.01$) as the most statistically significant environmental drivers of VCI. Furthermore, SHAP-based result reveals that solar radiation (mean $|SHAP| \approx 3.0$) and air temperature (mean $|SHAP| \approx 1.6$) contribute most significantly to the XGBoost model's predictions. Overall, this framework enables interpretable quantification of complex and nonlinear relationships on vegetation condition, supporting data-driven agricultural monitoring and management.

INTRODUCTION

The process of crop health monitoring extends past basic stress detection and yield prediction because it enables detailed large-scale crop management systems (Omia et al., 2023). The system requires data that shows both current accuracy and complete coverage of the entire area throughout time. The monitoring of large agricultural areas requires more than ground-based methods because they fail to provide sufficient coverage. The current monitoring system faces two major problems because it costs too much and does not provide enough area coverage. Remote sensing technology revolutionizes this situation. The MODIS and Sentinel-2 and Landsat satellites operate continuously to monitor vegetation and air and water conditions. The system delivers precise measurements of temperature and rainfall data. The technology provides exact real-time information about crops and their environment which proves essential for areas experiencing droughts and dust storms and pollution. The combination of remote sensing technology provides better results than traditional fieldwork methods because it delivers consistent data across extensive and large areas (Bojanowski et al., 2022; Segarra, 2020).

The rapid development of satellite remote sensing has advanced vegetation health monitoring by providing spatially continuous, repetitive, and cost-effective observations over large regions (Anyamba et al., 2001; Ji and Peters, 2003). Remote sensing-based vegetation indices have been widely adopted to assess vegetation health, and stress responses to environmental variability (Wang et al., 2001). Among these, the Normalized Difference Vegetation Index (NDVI) has become one of the most commonly used indicators for monitoring vegetation dynamics due to its simplicity and sensitivity to photosynthetically active biomass (Seiler et al., 2000; Aboelghar et al., 2010). NDVI has been successfully applied in a wide range of studies, including crop condition assessment, yield estimation, land degradation analysis, and dryland monitoring (Mondal et al., 2014).

Building upon NDVI, Kogan (1995) introduced the Vegetation Condition Index (VCI) to isolate the weather-related component of vegetation variability by scaling NDVI values relative to their long-term minimums and

Corresponding Author: M. Sc. Student, Department of Civil Engineering, Sharif University of Technology, Tehran, I. R. Iran
E-mail address: amir.haddadi81@sharif.edu
DOI: 10.22099/iar.2026.54487.1718

maximums. In contrast to absolute NDVI values, the VCI reflects vegetation stress relative to local climatological conditions, making it particularly suitable for drought monitoring across regions with differing ecological characteristics (Kogan, 1995; Bajgiran et al., 2008). The VCI has been widely applied using NOAA-AVHRR and MODIS datasets to detect the onset, duration, intensity, and spatial extent of agricultural droughts at regional to global scales (Seiler et al., 1998; Domenikiotis et al., 2004; Quiring and Ganesh, 2010).

Despite its advantages, the NDVI-derived VCI has limitations. Several studies have reported inconsistent relationships between VCI and station-based drought indices, particularly in arid and semiarid regions (Dutta et al., 2013; Patel et al., 2012). NDVI is affected by soil background effects in sparsely vegetated areas and exhibits saturation in regions with dense vegetation, limiting its sensitivity to capture changes in canopy structure and biomass (Cheng et al., 2023). Additionally, NDVI-based products are vulnerable to atmospheric noise, cloud contamination, and variations in sun–sensor geometry, which can introduce uncertainties into drought assessments if not properly addressed (Klisch and Atzberger, 2016).

Traditional statistical methods, including multiple linear regression, geographically weighted regression, and structural equation modeling, have been widely employed to analyze the drivers of drought and vegetation variability (Quiring and Ganesh, 2010). While these approaches provide valuable insights, they often struggle to capture complex nonlinear relationships and interactions among environmental variables, particularly when multiple drivers operate simultaneously across different spatial and temporal scales (Hao and Singh, 2015). Recent advances in machine learning (ML) offer new opportunities to address these challenges. Algorithms such as extreme gradient boosting (XGBoost) have demonstrated strong predictive performance in environmental applications due to their ability to model nonlinear relationships and high-order feature interactions (Chen and Guestrin, 2016). However, the “black-box” nature of many ML models has raised concerns regarding interpretability and physical understanding. To address this issue, interpretable ML techniques such as Shapley Additive Explanations (SHAP) have been introduced, providing a consistent framework for quantifying the contribution of individual predictors to model outputs (Lundberg and Lee, 2017).

Despite substantial progress, significant knowledge gaps remain in understanding how multiple environmental drivers jointly influence vegetation condition and agricultural drought. In particular, the relative importance of climatic variables (e.g., precipitation, temperature, and radiation), land surface states (e.g., soil moisture and LST), and atmospheric factors (e.g., AOD) remains insufficiently quantified across different regions and temporal scales. Moreover, few studies have jointly applied classical statistical analysis with interpretable ML to clarify the relative contributions using long-term, multi-source remote sensing and reanalysis datasets. Against this background, the present study aims to provide a comprehensive and interpretable assessment of the environmental drivers controlling vegetation condition and agricultural drought by integrating satellite-based vegetation indices, climate reanalysis data, and advanced ML techniques. By coupling traditional statistical approaches with XGBoost and SHAP analysis, this study seeks to quantify the relative contributions of key environmental variables to vegetation condition variability, thereby enhancing the understanding of drought–vegetation interactions and supporting data-driven decision-making for agricultural monitoring and sustainable ecosystem management.

MATERIALS AND METHODS

The methodological framework of this study follows a systematic process to determine the influence of environmental drivers on vegetation condition (Fig. 1). Monthly averaged meteorological, atmospheric, and vegetation-related data were first collected over a long-term period, and vegetation indices were derived to represent vegetation condition. Multicollinearity among environmental variables was evaluated prior to modeling to ensure the robustness of the analyses. Also, mutual information test was applied before entering the data to the ML model to select suitable features without any prior assumptions about their relationships with the target variable. Correlation analysis was used to explore the relationships between drivers and vegetation condition. The dataset was then partitioned into training and testing subsets for model development and validation. A Multiple Linear Regression (MLR) model was applied to identify linear relationships and assess statistical significance, whereas a nonlinear machine learning model was employed to capture complex interactions among variables. Model interpretability and driver attribution were achieved through variance analysis and explainable machine learning techniques. Finally, results from the statistical and ML approaches were compared to determine the most influential environmental drivers shaping vegetation condition.

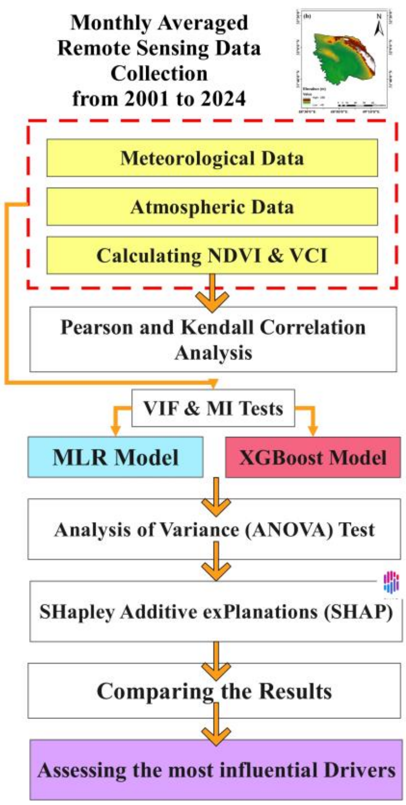

Fig. 1. Schematic overview of the methodological framework.

Study area

Khuzestan Province is located in southwestern Iran, between 30–33° N latitude and 48–50° E longitude, and covers an area of approximately 63,000 km² (Fig. 2). The province of Khuzestan is Iran's most environmentally vulnerable area because it experiences frequent dust storms, air contamination, and extended dry periods (Daniali and Karimi, 2019). The Shushtar sub-basin, in the middle section of Khuzestan Province, consists mainly of agricultural land, which makes up more than 80% of its total area through its extensive croplands, irrigated fields, and orchards (Fig. 2a). According to the ESRI Annual Land Cover, which is derived from ESA Sentinel-2 imagery at a 10 m resolution, Shushtar is characterized by fertile plains irrigated by the Karun River, with terrain rising northeastward and eastward toward the Zagros Mountains. (Fig. 2b). The local environment experiences changes in vegetation health because of the combined effects of temperature, humidity, and rainfall patterns throughout different seasons. The land use/land cover classification includes eight major categories representing the surface characteristics of the study area. Tree cover refers to areas dominated by tall woody vegetation that forms forests or dense tree stands. Shrubland is characterized by low-to medium-height woody plants with limited tree presence, whereas grassland consists mainly of grasses and herbaceous vegetation with minimal woody cover. Croplands represent areas actively used for agricultural production, including both irrigated and rain-fed systems. Built-up areas correspond to urban and artificial surfaces, such as buildings, roads, and other infrastructure. Bare vegetation includes sparsely vegetated or non-vegetated surfaces, such as exposed soil, sand, and rocky areas. Water denotes permanent or seasonal surface water bodies, including rivers, lakes, and reservoirs. Herbaceous wetlands are areas dominated by non-woody vegetation that experience periodic or permanent inundation.

Correlation analysis

Pearson (linear) and Kendall (rank-based) correlation coefficients were used to determine the strength and direction of the VCI relationship with climatic factors. The analysis determines whether the variables show a direct linear relationship (linear) or a general pattern of increase or decrease (monotonic).

Analysis of variance

The statistical method analysis of variance (ANOVA) helps researchers identify meaningful differences between the average values of independent groups. The statistical method ANOVA enables researchers to study multiple groups through independent variables, making it perfect for environmental and agricultural research that involves various treatment conditions, different climatic areas, and time periods. ANOVA was used to evaluate whether VCI measurements, which represent vegetation health, showed meaningful differences between different environmental stressor levels and climatic variables. The main principle of ANOVA is to separate the total data variability into two parts: the effects of independent variables and random errors and natural fluctuations. The fundamental equation is as follows:

$$F = \frac{MS_{between}}{MS_{within}} = \frac{\frac{SS_{between}}{k - 1}}{\frac{SS_{within}}{N - k}} \qquad \text{Eq. (1)}$$

where $SS_{between}$ and SS_{within} are the sum of squares between and within groups, respectively, k is the number of groups, and N is the total number of observations. A higher F-value suggests that the group means are not equal, and the corresponding *P*-value determines the statistical significance (commonly using a 0.05 threshold). In this study, the ANOVA results were derived within a regression framework (regression-based ANOVA) using continuous predictors, and no discretization or binning was applied to the data. To reduce multicollinearity, the predictors were iteratively screened using a VIF threshold of 5, after which the model was fitted, and Type II ANOVA was used to evaluate the significance and relative contribution of the retained drivers.

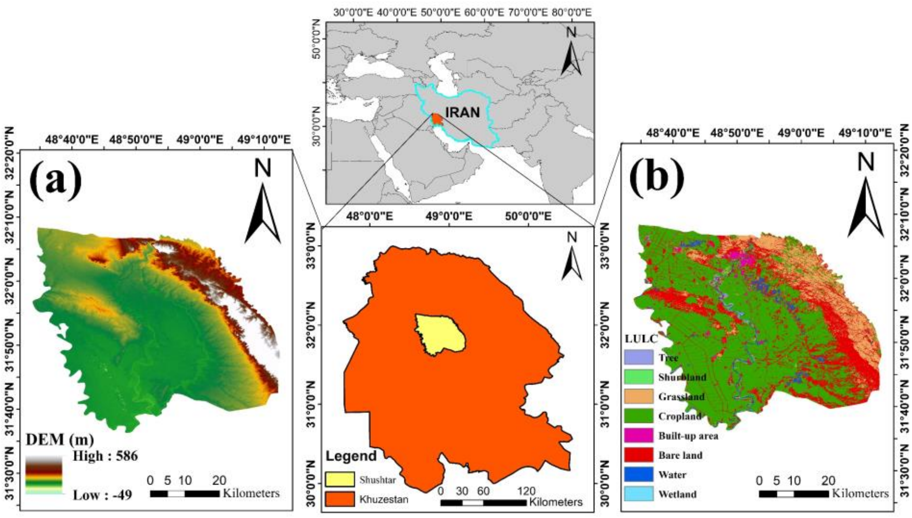

Fig. 2. The geographical location of the Shushtar. (a) Land use/cover map of Shushtar and (b) elevation map of Shushtar.

Multi linear regression model

The MLR model analyzed first-order linear relationships between vegetation condition (VCI) and ten environmental variables which included aerosol optical depth (AOD) and temperature and precipitation and soil water level near the surface, wind speed and surface pressure. The MLR model enables us to study dependent variables through linear combinations of independent predictors which produce standardized coefficients for interpreting factor contributions. The model required all variables to achieve zero mean and unit variance standardization before fitting because this step enabled coefficient comparison and scale effect reduction. Prior to implementing the MLR model, a Variance Inflation Factor (VIF) analysis was conducted to assess multicollinearity among the explanatory variables and to avoid the simultaneous inclusion of highly correlated predictors in the regression model.

SHAP analysis

The research applies SHapley Additive exPlanations (SHAP) to understand XGBoost model predictions which forecast VCI from AOD and climatic data. The XGBoost algorithm uses ensemble learning to train multiple decision trees sequentially through gradient descent optimization which corrects the residuals of previous trees. The SHAP method uses cooperative game theory to generate explainable predictions which reveal both global and local feature significance. Each feature is considered a player in a coalition contributing to the payout (i.e., the model prediction). The method demonstrates excellent performance when dealing with complex environmental systems that contain high-dimensional data with multicollinearity and non-stationarity. The model training process used 80% of the data but reserved 20% for independent testing (Bichri et al., 2024) to assess its ability to generalize. The hyperparameters set their optimal values through grid search and 5-fold cross-validation to minimize overfitting while enhancing model generalization. This procedure systematically evaluated combinations of key hyperparameters to identify the optimal model configuration while minimizing prediction error and preventing overfitting. The SHAP method enables us to understand model predictions by showing how each input variable affects the output. The XGBoost model allowed to evaluate its performance against MLR models and detect VCI patterns that extend beyond linear relationships.

Data

The research combines data from satellite-based and reanalysis products through the Google Earth Engine (GEE) Python API which provides free access to process and download the information. This study integrates multi-source remote sensing and reanalysis datasets to characterize environmental drivers and vegetation condition. Aerosol-related atmospheric information was obtained from the MODIS Terra and Aqua MCD19A2 product, a MAIAC-based Level-2 aerosol dataset that provides spatially consistent AOD observations for representing atmospheric aerosol loading. Meteorological and surface variables were derived from the ERA5-Land reanalysis dataset, which offers a physically consistent representation of land–atmosphere interactions by combining observations with land surface modeling to provide key climatic and hydrological parameters such as air temperature, precipitation, solar radiation, evapotranspiration, and soil moisture. Vegetation condition was assessed using atmospherically corrected Landsat Collection 2 Level-2 surface reflectance products acquired from Landsat 5 TM, Landsat 7 ETM+, Landsat 8 OLI, and Landsat 9 OLI-2. The NDVI was computed using the near-infrared (NIR) and red bands, specifically NIR (Band 4) and Red (Band 3) for Landsat 5 and Landsat 7, and NIR (Band 5) and Red (Band 4) for Landsat 8 and Landsat 9 (Eq. (2)). The resulting NDVI time series was subsequently used to derive the Vegetation Condition Index (VCI) (Eq. (3)). The spatial and temporal resolutions of each dataset are summarized in Table 1. All datasets used in this study were collected for the period 2001–2024 in the Shushtar sub-basin and were aggregated to monthly mean values to ensure temporal consistency across data sources. The VCI was derived from the monthly mean NDVI values by normalizing each month’s NDVI relative to its corresponding long-term minimum and maximum values, thereby enabling a standardized assessment of the vegetation condition.

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} - \mathrm{Red}} \qquad \text{Eq. (2)}$$

$$\mathrm{VCI} = \frac{\mathrm{NDVI} - \mathrm{NDVI}_{\mathrm{min}}}{\mathrm{NDVI}_{\mathrm{max}} - \mathrm{NDVI}_{\mathrm{min}}} \times 100 \qquad \text{Eq. (3)}$$

Table 1. Detailed summary of the temporal and spatial characteristics of the employed datasets

Dataset	Temporal resolution	Spatial resolution
MODIS MAIAC AOD (Terra & Aqua)	Daily (1 day)	~1.1 km
ERA5-Land reanalysis	Hourly	~9 km
Landsat 5, 7, 8, and 9	16-day revisit	~30 m

RESULTS AND DISCUSSION

Data distribution and analysis

The distributional analysis of the selected variables shows differences in their statistical characteristics and degrees of deviation from normality, which have important implications for modeling vegetation condition (Fig. 3). Variables such as wind speed and dewpoint temperature exhibit distributions that are relatively symmetric and closely aligned with the fitted normal curves, indicating more stable variability and fewer extreme values over the study period. In contrast, precipitation, AOD, and volumetric soil water show right-skewed distributions, reflecting the discontinuous and episodic nature of rainfall events, aerosol loading peaks, and soil moisture pulses in semi-arid region. Evaporation and potential evaporation display broader and slightly multimodal distributions, suggesting the influence of seasonal cycles and varying atmospheric demand regimes. Solar radiation and air temperature exhibit wider ranges and mild bimodality, consistent with strong seasonal contrasts between cooler and warmer periods. The Vegetation Condition Index (VCI) shows an approximately bell-shaped distribution with slight skewness, indicating that most observations fall within moderate vegetation condition states, while extreme stress or optimal conditions occur less frequently. Building on these distributional patterns, the deviations from normality suggest that linear modeling assumptions may not be uniformly appropriate across all environmental predictors. Variables characterized by skewness and multimodality are likely to capture episodic or threshold-driven processes that exert non-proportional influence on vegetation dynamics, particularly during extreme climatic conditions. In contrast, more normally distributed variables may represent background climatic controls that regulate vegetation condition more gradually over time. These results highlight the importance of employing robust or nonlinear modeling approaches.

Correlation among variables

Overall, the two correlation measures exhibited consistent patterns, although Pearson correlations generally showed higher magnitudes owing to their sensitivity to linear relationships, whereas Kendall correlations captured monotonic but potentially nonlinear associations and were less influenced by outliers (Fig. 4).

Regarding vegetation greenness, NDVI demonstrated moderate correlations with several environmental variables. In the Pearson analysis, NDVI showed a positive correlation with volumetric soil moisture ($r \approx 0.43$) and potential evaporation ($r \approx 0.40$), indicating that improved soil water availability and favorable surface energy conditions enhance canopy greenness. Conversely, NDVI was negatively correlated with air temperature ($r \approx -0.40$), solar radiation ($r \approx -0.31$), evaporation ($r \approx -0.42$), and AOD; $r \approx -0.31$. These relationships suggest that increased thermal and radiative stress, as well as elevated aerosol loading, may suppress photosynthetic activity by enhancing evapotranspiration demand or reducing the effective incoming radiation. The Kendall correlations for NDVI followed similar trends but with lower magnitudes ($\tau \approx -0.25$ to 0.27), reflecting monotonic but nonlinear responses and the influence of seasonal and interannual variability.

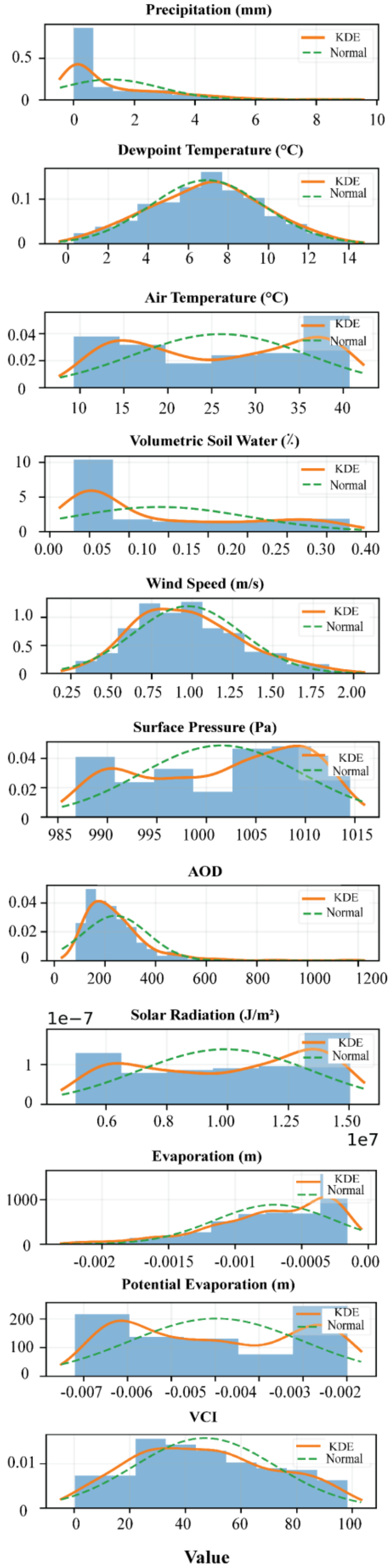

Fig. 3. Probability distributions of selected variables and vegetation condition index.

In contrast, VCI exhibits substantially weaker correlations with most environmental drivers in both Pearson and Kendall analyses. Pearson correlation coefficients between VCI and individual climatic or atmospheric variables are generally close to zero ($|r| < 0.10$), such as precipitation ($r \approx 0.03$), air temperature ($r \approx 0.09$), soil moisture ($r \approx 0.03$), and wind speed ($r \approx -0.01$), while slightly higher but still weak correlations are observed with AOD ($r \approx -0.26$). Kendall correlations confirm these weak associations, with τ values typically below ± 0.20. This behavior is primarily attributed to the formulation of VCI, which normalizes NDVI relative to its long-term monthly minimum and maximum, thereby reducing the direct influence of absolute environmental variability and emphasizing relative vegetation stress conditions.

Several factors may explain the weak correlations observed between some environmental drivers and vegetation indices, particularly the VCI. (1) Vegetation responses to environmental forcing are often nonlinear and may involve time lags that cannot be fully captured by zero-lag correlation analysis. (2) Vegetation conditions are influenced by the combined interaction of multiple drivers rather than a single variable, leading to diminished pairwise correlations despite strong multivariate effects. (3) Human interventions, such as irrigation and land management practices, can decouple vegetation conditions from climatic variability, especially in agricultural regions. Additionally, the normalization procedures inherent in the VCI calculation suppress the magnitude of correlations by design, as the index represents relative anomalies rather than absolute vegetation states. The measurement uncertainty in satellite-derived variables and spatial averaging effects may further attenuate the correlation strength.

VIF and MI tests

Prior to fitting the MLR model, multicollinearity among the environmental predictors was assessed using the Variance Inflation Factor (VIF), with a threshold value of 5 adopted to identify problematic collinearity (Fig. 5). The VIF analysis revealed severe multicollinearity for potential evaporation, air temperature, and solar radiation, with VIF values exceeding the selected threshold, indicating strong redundancy and shared variance among these energy-related variables. Consequently, these predictors were excluded from the final MLR model to avoid the simultaneous inclusion of highly correlated variables. In contrast, volumetric soil water, surface pressure, precipitation, evaporation, dewpoint temperature, AOD, and wind speed exhibited VIF values below the threshold and were retained. This preprocessing step reduces coefficient instability and inflation, improves model robustness, and enhances the interpretability of the MLR results by ensuring that the estimated effects reflect independent contributions of environmental drivers to vegetation condition. For correct linear fitting, all variables were normalized using a standard scaler to account for differences in units and magnitudes, allowing the linear effects of predictors to be evaluated without bias from scale differences.

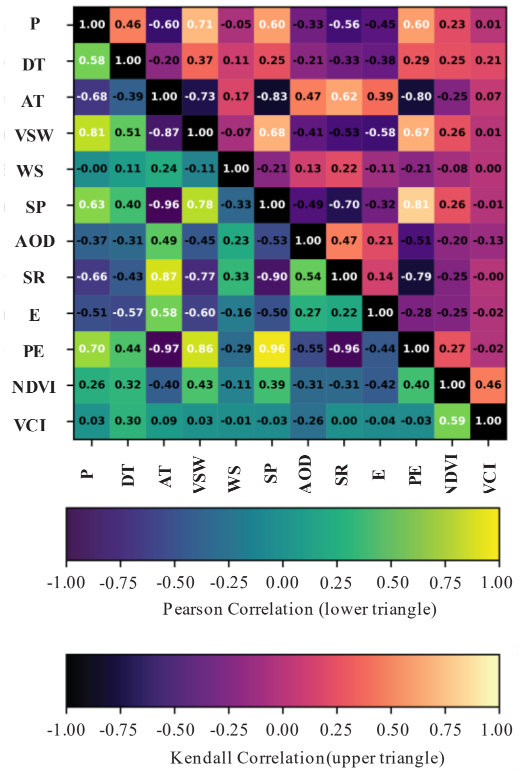

Fig. 4. Pearson and Kendall correlation results. Variables include Precipitation (P), Dewpoint Temperature (DT), Air Temperature (AT), Volumetric Soil Water (VSW), Wind Speed (WS), Surface Pressure (SP), Aerosol Optical Depth (AOD), Solar Radiation (SR), Evaporation (E), Potential Evaporation (PE), Normalized Difference Vegetation Index (NDVI), and Vegetation Condition Index (VCI).

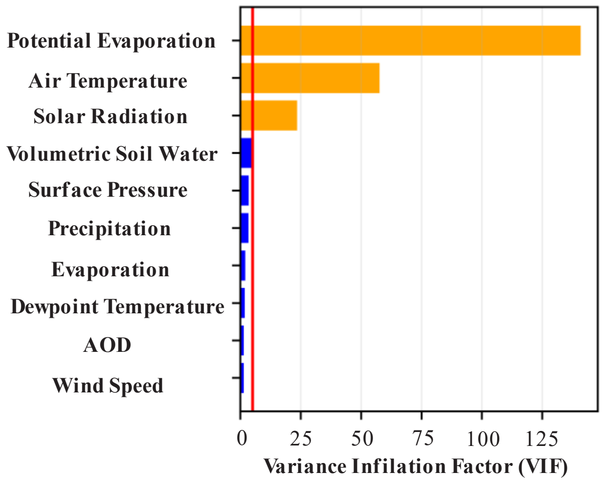

Fig. 5. Variance inflation factor (VIF) analysis of environmental predictors, with the red vertical line indicating the adopted VIF threshold of 5 for variable selection.

MI scores were computed using a non-parametric estimator without assuming linearity or specific data distributions. A minimum MI threshold of zero was adopted, meaning that only features exhibiting no measurable information gain with respect to the target were excluded. Under this setup, all input variables yielded positive MI values exceeding the defined threshold; therefore, no features were removed at this stage, and all variables were retained for subsequent modeling.

MLR model

The MLR model was fitted using the remaining seven environmental predictors. The model achieved an R^2 value of 0.29, an adjusted R^2 of 0.27, an RMSE of 19.06, and an MAE of 11. 47, reflecting a weak predictive performance that is expected, given the linear structure of the model and the inherently complex and nonlinear nature of vegetation–environment interactions.

The regression coefficients (Fig. 6) provide insights into the direction and relative magnitude of each predictor's linear influence on the VCI. Volumetric soil water exhibited the strongest positive coefficient, highlighting soil moisture availability as the dominant driver of vegetation conditions. Precipitation also had a positive contribution, although with a smaller magnitude, suggesting that its influence on VCI is partly indirect and mediated through soil moisture storage rather than an immediate vegetation response. Dewpoint temperature and wind speed displayed weak positive coefficients, indicating a limited but favorable association with vegetation conditions, likely through their roles in atmospheric moisture and energy exchange. In contrast, evaporation and AOD presented negative coefficients, implying that increased evaporative losses and higher aerosol loading tend to reduce vegetation conditions by intensifying water stress and limiting effective radiation. Surface pressure showed a near-zero coefficient, suggesting a negligible direct linear effect on VCI once other environmental drivers were accounted for.

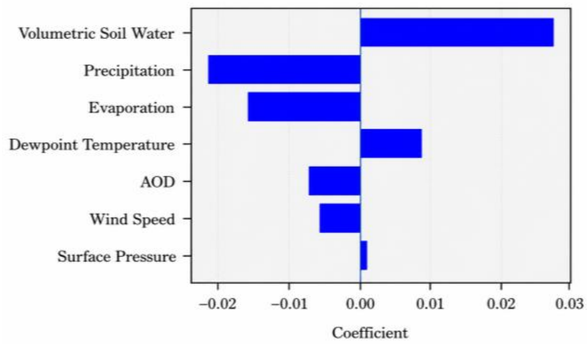

Fig. 6. Estimated regression coefficients of the multiple linear regression (MLR) model after multicollinearity removal using the vsriance infilteration factor (VIF) test.

ANOVA test

The results of the two-way Analysis of Variance (ANOVA) in Table 2 provide quantitative insight into the relative importance of environmental drivers influencing the Vegetation Condition Index (VCI). Among all predictors, AOD demonstrated the strongest explanatory power, exhibiting the highest sum of squares (9284.5) and F-value (20.11), with a highly significant *P*-value ($P < 0.01$). This indicates that atmospheric aerosol loading plays a dominant role in explaining the variability in vegetation conditions, likely through its influence on solar radiation attenuation and plant physiological stress.

Surface pressure and volumetric soil water also showed strong and statistically significant effects on VCI, with F-values of 18.19 and 16.37, respectively, and *P*-values below 0.01, underscoring the critical roles of atmospheric circulation patterns and soil moisture availability in regulating vegetation stress and productivity. Dew point temperature and air temperature further contributed significantly to VCI variability ($P < 0.01$), reflecting the importance of atmospheric moisture and thermal conditions for vegetation health.

In contrast, precipitation exhibited a weaker and marginally non-significant effect ($P = 0.06$), suggesting that short-term precipitation variability alone may not directly translate into an immediate vegetation response. Evaporation and wind speed showed the lowest explanatory power and a non-significant *P*-value ($P = 0.14$ and 0.7), indicating a limited direct influence on VCI within the analyzed framework.

Table 2. The summary of the result of Analysis of Variance (ANOVA) test

Predictor	Sum of square	F-value	*P*-value
Aerosol optical depth (AOD)	9284.5	20.11	0.00
Surface pressure	8395.1	18.19	0.00
Volumetric soil water	7555.7	16.37	0.00
Dewpoint temperature	5723.5	12.40	0.00
Precipitation	1693.6	3.67	0.06
Evaporation	997.9	2.16	0.14
Wind speed	69.9	0.15	0.70

The ANOVA results indicate that vegetation condition variability is primarily controlled by atmospheric and soil-related factors rather than direct hydrological inputs. The variables associated with aerosol loading, atmospheric pressure, soil moisture, and thermal–humidity conditions collectively exert a dominant influence on VCI, emphasizing the combined effects of radiation regulation, moisture availability, and atmospheric dynamics on vegetation stress. The comparatively weaker influence of precipitation and evaporation suggests that vegetation response is more strongly linked to integrated moisture and energy conditions than to short-term fluxes of these variables. These findings highlight the importance of considering both atmospheric composition and land–atmosphere interactions when assessing vegetation dynamics under changing environmental and climatic conditions.

SHAP analysis

The performance of the XGBoost model shows an improvement over the linear regression approach, highlighting its ability to capture complex and nonlinear relationships between environmental drivers and vegetation condition. During the training phase, the model achieved a coefficient of determination ($R^2 = 0.91$, adjusted $R^2 = 0.90$). The training RMSE (1.82) and MAE (1.36) further reflect a strong fit between the predicted and observed values. When evaluated on the test dataset, the model maintained robust predictive performance, with an R^2 of 0.86 and an adjusted R^2 of 0.84, confirming its generalization capability. The higher RMSE (9.32) and MAE (6.71) observed in the test phase are expected given the variability of unseen data and are indicative of normal generalization behavior rather than model instability.

To quantify the prediction uncertainty, a bootstrap resampling approach was applied to the model residuals. A total of 2000 bootstrap samples were generated to estimate the uncertainty associated with the model errors, from which a 95% confidence interval for the RMSE was derived (Fig. 7). In addition, the distribution of bootstrapped residual standard deviations was used to construct a residual-based 95% prediction band (± 1.96σ) around the 1:1 line. Most predictions fell within this uncertainty envelope, confirming the robustness and reliability of the model across varying vegetation conditions. The color-coded points further illustrate the prediction accuracy, with higher accuracy concentrated near the well-fitted line and lower accuracy associated with larger deviations.

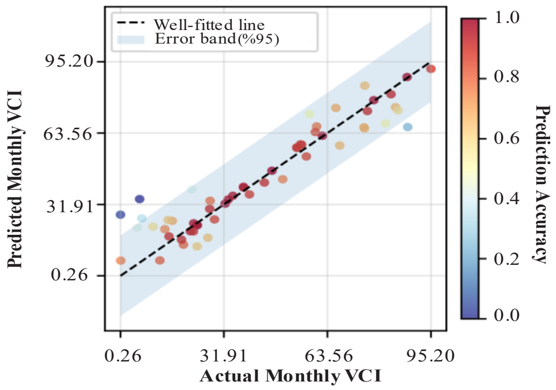

Fig. 7. Comparison between observed and predicted monthly vegetation condition index (VCI) values obtained from the optimized Extreme Gradiant Boosting (XGBoost) model.

The SHAP-based feature importance analysis reveals that surface pressure is the most influential environmental driver affecting vegetation condition, as indicated by the highest mean absolute SHAP value (Fig. 8). This highlights the role of large-scale atmospheric circulation and pressure-related controls on near-surface climate conditions relevant to vegetation dynamics. Evaporation ranks second in importance, emphasizing the strong influence of land–atmosphere exchange processes and atmospheric water demand on vegetation condition..AOD exhibits a moderate contribution, suggesting that aerosol-related radiative effects play a secondary but non-negligible role. In contrast, wind speed and volumetric soil water display relatively lower mean SHAP values.

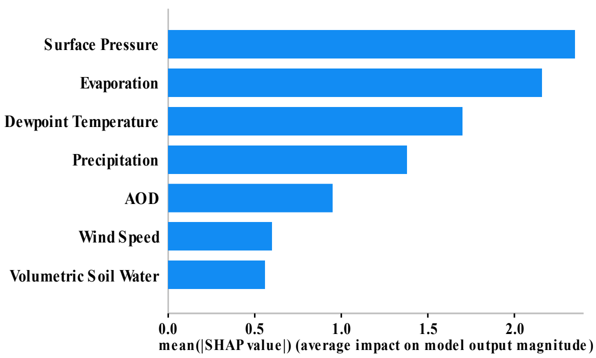

Fig. 8. Shapley Additive exPlanations (SHAP) feature importance ranking showing the mean absolute SHAP values for environmental drivers in the Extreme Gradiant Boosting (XGBoost) model.

Fig. 9, shows the SHAP beeswarm distributions for the dominant environmental drivers, highlighting both the magnitude and direction of their impacts on vegetation condition. Surface pressure exhibits the widest spread of SHAP values, confirming it as the most influential variable; higher surface pressure values are predominantly associated with positive SHAP contributions, whereas lower pressures tend to reduce the predicted VCI, reflecting the role of large-scale atmospheric stability and circulation patterns. Evaporation also shows a broad and structured distribution, indicating a strong but nonlinear influence, where moderate evaporation supports vegetation condition while high evaporative demand leads to negative impacts due to the enhanced water stress. Dewpoint temperature demonstrates a consistent positive relationship, with higher values generally contributing positively to VCI, underscoring the importance of atmospheric moisture availability. Precipitation displays a moderate spread with both positive and negative contributions, suggesting context-dependent effects influenced by timing and interaction with other hydrometeorological variables. AOD is mainly associated with negative SHAP values at higher concentrations, indicating a suppressive effect on vegetation condition likely related to the radiative attenuation. In contrast, wind speed and volumetric soil water exhibit relatively narrow SHAP distributions centered around zero, implying weaker direct effects and a stronger dependence on nonlinear interactions with other environmental drivers captured by the model.

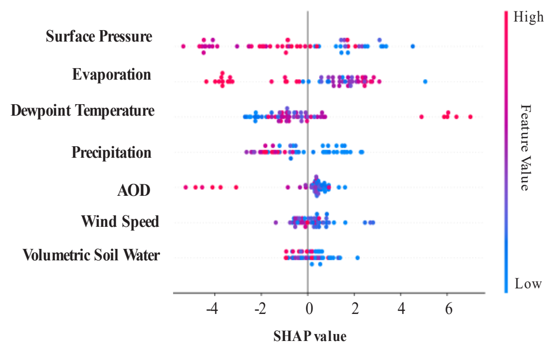

Fig. 9. Shapley Additive exPlanations (SHAP) summary (beeswarm) plot illustrating the distribution, direction of the effects of environmental drivers on Extreme Gradiant Boosting (XGBoost) model predictions of vegetation condition (VCI).

The SHAP dependence plots (Fig. 10) provide detailed insight into the direction, strength, and functional form of the relationships between individual environmental drivers and the predicted vegetation condition (VCI), while revealing nonlinear responses captured by the XGBoost model. Surface pressure exhibits an overall negative trend, with SHAP values decreasing as pressure increases, indicating that higher pressure regimes are generally associated with reduced vegetation condition in the study area. Evaporation also shows a clear negative relationship, suggesting that increased evaporative demand consistently suppresses VCI due to the enhanced surface drying and water stress. In contrast, dewpoint temperature displays a strong positive association with SHAP values, highlighting the beneficial role of higher atmospheric moisture availability in supporting vegetation condition. Precipitation demonstrates a weak to moderately negative trend, implying that its direct influence on VCI is limited and likely mediated through delayed soil moisture responses rather than immediate effects. AOD exhibits a pronounced negative relationship, with SHAP values declining sharply at higher aerosol concentrations, underscoring the adverse impact of aerosol loading through radiative attenuation and related stress mechanisms. Wind speed shows a generally negative pattern, where stronger winds tend to reduce VCI by enhancing evapotranspiration and mechanical stress on vegetation. Finally, volumetric soil water presents a relatively flat and weak dependence, indicating that its contribution to vegetation condition is largely indirect and governed by interactions with other environmental drivers within the nonlinear modeling framework.

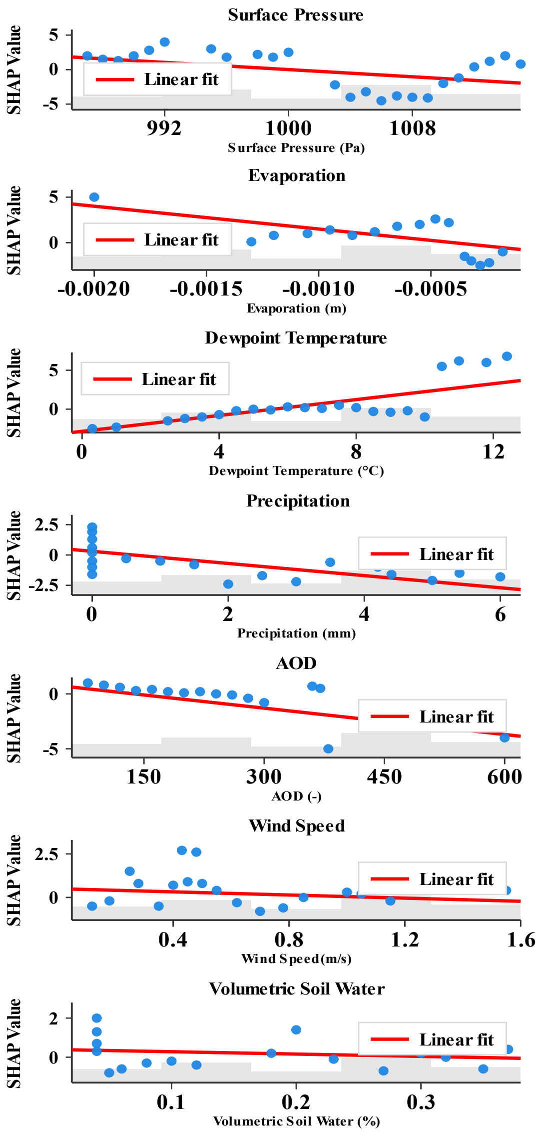

Fig. 10. Shapley Additive exPlanations (SHAP) dependence plots illustrating the marginal effects of individual environmental drivers on model predictions of vegetation condition (VCI).

CONCLUSION

This study demonstrates that the relative importance of environmental drivers controlling vegetation condition. Linear techniques, including regression-based ANOVA and MLR, identified a limited set of statistically significant predictors of VCI, primarily related to the moisture availability and atmospheric conditions. However, the low explanatory power of the MLR model ($R^2 \approx 0.29$) highlights a key limitation of linear frameworks: their restricted capacity to represent nonlinear responses, threshold effects, and interdependencies among environmental drivers. While these methods offer straightforward statistical approach, their reliance on assumptions of linearity, additivity, and homoscedasticity constrains their applicability in complex environmental systems. In contrast, the XGBoost model exhibited markedly superior predictive performance (R^2 = 0.86), underscoring the advantage of tree-based ML approaches in capturing nonlinear relationships and higher-order interactions. The SHAP-based interpretation further revealed that large-scale atmospheric controls, particularly surface pressure, alongside evaporation and moisture-related variables, exert dominant influences on vegetation condition. These findings suggest that vegetation dynamics are regulated not by isolated drivers, but by the combined effects of atmospheric circulation, evaporative demand, and moisture availability. Such interaction-driven controls are often underestimated in linear analyses, a limitation widely acknowledged in previous vegetation–climate studies employing regression-based approaches. Nevertheless, the applied ML framework is not without limitations. As prior studies have similarly reported enhanced performance of ML models in capturing the target responses to complex feature forcings, while also emphasizing the importance of interpretability tools to mitigate the "black-box" nature of these approaches. Additionally, the reliance on satellite-derived inputs introduces uncertainties related to sensor limitations, retrieval algorithms, and spatial resolution mismatches. Overall, this study highlights the complementary roles of linear and nonlinear modeling strategies: linear methods provide baseline statistical insights and hypothesis testing, whereas ML models offer superior predictive skill and a more realistic representation of complex vegetation–environment interactions. Future work should focus on integrating ground-based observations, incorporating spatiotemporal dependencies, and developing hybrid frameworks that combine process-based understanding with ML flexibility to improve attribution and prediction of vegetation dynamics under ongoing climate change.

FUNDING

This research had no funding.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Methodology: Amirhossein Haddadi and Behnam Rahimi; Programming: Amirhossein Haddadi; Remote sensing data gathering: Amirhossein Haddadi; Formal analysis: Amirhossein Haddadi; Investigation: Amirhossein Haddadi; Writing—original draft preparation: Amirhossein Haddadi; Writing—review: Amirhossein Haddadi; Revision writing: Behnam Rahimi; Literature Review: Behnam Rahimi.

DECLARATION OF COMPETING INTEREST

The authors declare no conflicts of interest.

ETHICAL STATEMENT

Not Applicable.

DATA AVAILABILITY

The data used in this study can be provided by the authors upon reasonable requests.

REFERENCES

Aboelghar, M., Arafat, S., Saleh, A., Naeem, S., Shirbeny, M., & Belal, A. (2010). Retrieving leaf area index from SPOT4 satellite data. *The Egyptian Journal of Remote Sensing and Space Science*, *13*(2), 121-127. https://doi.org/10.1016/j.ejrs.2010.06.001

Anyamba, A., Tucker, C. J., & Eastman, J. R. (2001). NDVI anomaly patterns over Africa during the 1997/98 ENSO warm event. *International Journal of Remote Sensing*, *22*(10), 1847-1860. https://doi.org/10.1080/01431160010029156

Bajgiran, P. R., Darvishsefat, A. A., Khalili, A., & Makhdoum, M. F. (2008). Using AVHRR-based vegetation indices for drought monitoring in the Northwest of Iran. *Journal of Arid Environments*, *72*(6), 1086-1096. https://doi.org/10.1016/j.jaridenv.2007.12.004

Bichri, H., Chergui, A., & Hain, M. (2024). Investigating the impact of train/test split ratio on the performance of pre-trained models with custom datasets. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *15*(2), 331-339. https://doi.org/10.14569/IJACSA.2024.0150235

Bojanowski, J. S., Sikora, S., Musiał, J. P., Woźniak, E., Dąbrowska-Zielińska, K., Ślesiński, P., Milewski, T., & Łączyński, A. (2022). Integration of Sentinel-3 and MODIS vegetation indices with ERA-5 agro-meteorological indicators for operational crop yield forecasting. *Remote Sensing*, *14*(5), 1238. https://doi.org/10.3390/rs14051238

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system, presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA.* New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Cheng, M., Zhong, L., Ma, Y., Wang, X., Li, P., Wang, Z., & Qi, Y. (2023). A new drought monitoring index on the Tibetan Plateau based on multisource data and machine learning methods. *Remote Sensing*, *15*(2), 512. https://doi.org/10.3390/rs15020512

Daniali, M., & Karimi, N. (2019). Spatiotemporal analysis of dust patterns over Mesopotamia and their impact on Khuzestan province, Iran. *Natural Hazards*, *97*(2), 259-281. https://doi.org/10.1007/s11069-019-03641-0

Domenikiotis, C., Spiliotopoulos, M., Tsiros, E., & Dalezios, N. R. (2004). Early cotton yield assessment by the use of the NOAA/AVHRR derived Vegetation Condition Index (VCI) in Greece. *International Journal of Remote Sensing*, *25*(14), 2807-2819. https://doi.org/10.1080/01431160310001632729

Dutta, D., Kundu, A., & Patel, N. R. (2013). Predicting agricultural drought in eastern Rajasthan of India using NDVI and standardized precipitation index. *Geocarto International*, *28*(3), 192-209. https://doi.org/10.1080/10106049.2012.679975

Hao, Z., & Singh, V. P. (2015). Drought characterization from a multivariate perspective: A review. *Journal of Hydrology*, *527*, 668-678. https://doi.org/10.1016/j.jhydrol.2015.05.031

Ji, L., & Peters, A. J. (2003). Assessing vegetation response to drought in the northern Great Plains using vegetation and drought indices. *Remote sensing of Environment*, *87*(1), 85-98. https://doi.org/10.1016/S0034-4257(03)00174-3

Klisch, A., & Atzberger, C. (2016). Operational drought monitoring in Kenya using MODIS NDVI time series. *Remote Sensing*, *8*(4), 267. https://doi.org/10.3390/rs8040267

Kogan, F. N. (1995). Application of vegetation index and brightness temperature for drought detection. *Advances in Space Research*, *15*(11), 91-100. https://doi.org/10.1016/0273-1177(95)00079-T

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions, presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA. Red Hook, NY: Curran Associates, Inc. https://doi.org/10.48550/arXiv.1705.07874

Mondal, S., Jeganathan, C., Sinha, N. K., Rajan, H., Roy, T., & Kumar, P. (2014). Extracting seasonal cropping patterns using multi-temporal vegetation indices from IRS LISS-III data in Muzaffarpur District of Bihar, India. *The Egyptian Journal of Remote Sensing and Space Science*, *17*(2), 123-134. https://doi.org/10.1016/j.ejrs.2014.09.002

Omia, E., Bae, H., Park, E., Kim, M. S., Baek, I., Kabenge, I., & Cho, B.-K. (2023). Remote sensing in field crop monitoring: A comprehensive review of sensor systems, data analyses and recent advances. *Remote Sensing*, *15*(2), 354. https://doi.org/10.3390/rs15020354

Patel, N. R., Parida, B. R., Venus, V., Saha, S. K., & Dadhwal, V. K. (2012). Analysis of agricultural drought using vegetation temperature condition index (VTCI) from Terra/MODIS satellite data. *Environmental Monitoring and Assessment*, *184*, 7153-7163. https://doi.org/10.1007/s10661-011-2487-7

Quiring, S. M., & Ganesh, S. (2010). Evaluating the utility of the Vegetation Condition Index (VCI) for monitoring meteorological drought in Texas. *Agricultural and Forest Meteorology*, *150*(3), 330-339. https://doi.org/10.1016/j.agrformet.2009.11.015

Segarra, J., Buchaillot, M. L., Araus, J. L., & Kefauver, S. C. (2020). Remote sensing for precision agriculture: Sentinel-2 improved features and applications. *Agronomy*, *10*(5), 641. https://doi.org/10.3390/agronomy10050641

Seiler, R. A., Kogan, F., & Wei, G. (2000). Monitoring weather impact and crop yield from NOAA AVHRR data in Argentina. *Advances in Space Research*, *26*(7), 1177-1185. https://doi.org/10.1016/S0273-1177(99)01144-8

Wang, J., Price, K. P., & Rich, P. M. (2001). Spatial patterns of NDVI in response to precipitation and temperature in the central Great Plains. *International Journal of Remote Sensing*, *22*(18), 3827-3844. https://doi.org/10.1080/01431160010007033