Proceedings of CONF-MCEE 2026 Symposium: Geomaterials and Environmental Engineering
DOI: 10.54254/2755-2721/2026.MH35286

Carbon Sink Variations and Driving Factors Based on XGBoost-SHAP Analysis in the Beijing-Tianjin-Hebei Region

Yiquan Liu

School of Earth System Science, Tianjin University, Tianjin, China
2561734367@qq.com

Abstract. The Beijing-Tianjin-Hebei (named Jing-Jin-Ji in China, JJJ), as a key economic and industry development region in China, plays a vital role in achieving China's carbon neutrality goals. Understanding the dynamic changes in JJJ's carbon sink and its driving factors is crucial for policy formulation on carbon emissions. In this study, we use multi source satellite observations, eXtreme gradient boosting model (XGBoost) and SHapley Additive exPlanations (SHAP) to explore the spatiotemporal variations of gross primary productivity (GPP) and net primary productivity (NPP), and the impact of environmental factors on them. The results show that both GPP and NPP exhibited significant upward trends ($p < 0.001$) in BTH, with increasing rates of 9.79 g C m-2 year-1 and 5.21 g C m-2 year-1, respectively. The spatial distribution of GPP and NPP exhibits significant similar north-high, south-low heterogeneity. High-value are concentrated in the forest and mixed forest ecosystems, while low-value are primarily located in farmlands and urban construction. High-performance XGBoost models (R2 equals 0.72 for GPP and 0.84 for NPP) were established and employed to implement SHAP analysis. We found that the primary limiting environmental factor affecting GPP is relative humidity, followed by third-layer soil moisture and temperature. Temperature is the primary limiting driver for NPP, followed by temperature-radiation interactions and third-layer soil moisture. This study provides scientific evidence for understanding the carbon cycle processes within ecosystems in the JJJ region and their response to climate change, offering significant reference value for regional carbon sink management.

Keywords: Jing-Jin-Ji, Carbon Sink, XGBoost, SHAP

1. Introduction

Global warming, driven by increasing atmospheric carbon dioxide (CO_2) concentrations from industrial emissions, is the most obvious consequence of anthropogenic activity's impact on the earth. Terrestrial ecosystems are vital for regulating climate change and mitigating global warming by enhancing carbon sinks. Gross primary productivity (GPP) and net primary productivity (NPP) are widely used core metrics for quantifying ecosystem's carbon uptake and sequestration potential. Accurate estimation of GPP and NPP, coupled with a profound understanding of their spatiotemporal variability, is crucial for clarifying regional carbon cycling processes, assessing

DOI: 10.54254/2755-2721/2026.MH35286

dynamic changes in carbon budgets, and characterizing ecosystem responses to environmental modifications [1].

The Beijing-Tianjin-Hebei (named Jing-Jin-Ji in China, JJJ) region, located at the northern edge of the North China Plain and consisting of Beijing, Tianjin, and Hebei Province, is one of China's most important political and economic centers. This region features pronounced heterogeneity in topography and supports diverse ecosystem types, including forests, grasslands, croplands, and wetlands. Over recent decades, rapid urbanization, intensive industrial development, and large-scale agricultural expansion have subjected local ecosystems to immense pressure, compromising the capacity and resilience of regional carbon sinks. Therefore, comprehensively assessing the spatiotemporal dynamics of GPP and NPP in the JJJ region and their driving mechanisms is crucial for formulating effective ecological conservation strategies and supporting China's carbon neutrality goals [2].

Most previous large-scale GPP/NPP studies have relied on process-based models, particularly models based on light-use efficiency frameworks such as CASA, VPM, EC-LUE and BEPS. While these models are valuable for representing biophysical processes, they may struggle to capture complex and nonlinear interactions among climate variability and ecosystem functions [3]. To more accurately capture the impacts of climate change and human activities on ecosystem carbon sinks, the tree-based ensemble learning algorithm eXtreme Gradient Boosting (XGBoost), a tree-based ensemble learning algorithm with strong performance in nonlinear regression and built-in regularization to reduce overfitting, has been widely used. To overcome the limitations of pure predictive outcomes and enhance interpretability, the SHarpley Additivity exPlanation (SHAP) based on cooperative game theory was integrated into the XGBoost-based research framework. SHAP provides a consistent framework to quantify the marginal contribution of each predictor to model outputs, enabling attribution of GPP/NPP variability to individual environmental drivers and revealing how multiple factors jointly regulate carbon uptake across space and time [4,5].

In this study, we combine satellite products and meteorological observations spanning from 2001 to 2024 to estimate regional GPP and NPP with an XGBoost-based modeling framework in the JJJ. We then apply SHAP to diagnose the dominant environmental controls on carbon sink dynamics and to interpret their relative importance and potential interactions. By characterizing the spatiotemporal patterns of ecosystem productivity and identifying key drivers, our works aim to provide a scientific basis for carbon sink assessment and management in the JJJ, and to support evidence-based policies for ecological protection and climate mitigation.

2. Methodology and datasets

2.1. Study area

The Beijing–Tianjin–Hebei region (36°03′N–42°40′N, 113°27′E–119°50′E) covers approximately 218000 km^2. This region has a temperate monsoon climate with distinct seasons: hot, wet summer and cold, dry winter. Mean annual temperature ranges from 8 to 14 °C, and annual precipitation is from 400 to 800 mm, concentrated mainly in summer in this region. Based on MODIS land-cover data (Fig. 1), Cropland is the dominant land use type and is concentrated across the central and southern plains, and forests are mainly distributed in the northern and western mountains [2].

DOI: 10.54254/2755-2721/2026.MH35286

2.2. Datasets

The GPP and NPP products are from MODIS/Terra Net Primary Production Gap-Filled Yearly L4 Global 500m SIN Grid V061 (https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod17a3hgf-061). The land-cover product is from MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid. Meteorological and soil moisture variables (https://www.earthdata.nasa.gov/data/catalog/lpcloud-mcd12q1-006). Precipitation, air temperature, relative humidity, solar radiation, and soil moisture are obtained from ECMWF reanalysis data (ERA5/ERA5-Land). Soil moisture is represented using volumetric soil water in four layers: 0–7 cm (L1), 7–28 cm (L2), 28–100 cm (L3), and 100–289 cm (L4) (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview). All datasets undergo quality control, spatial co-registration, and resampling, and are standardized to a common 500 m spatial resolution and annual temporal resolution for model training and analysis.

2.3. XGBoost model

Extreme Gradient Boosting (XGBoost) is an efficient and robust ensemble learning algorithm that builds an additive model by iteratively optimizing an objective function and combining multiple weak learners. Its key advantages include: (1) explicit L1/L2 regularization to control model complexity and reduce overfitting; and (2) second-order approximation of the loss function to improve optimization efficiency and accuracy. In this study, precipitation, temperature, relative humidity, solar radiation, multi-layer soil moisture (L1–L4), and selected interaction terms (temperature with radiation and temperature with precipitation) are used as input features. Separate XGBoost regression models are developed for GPP and NPP prediction [4].

2.4. SHapley Additive exPlanations analyze

SHapley Additive exPlanations (SHAP), based on cooperative game theory, provides a unified and theoretically consistent framework for attributing model predictions to individual features. SHAP quantifies the importance of different features through SHAP values, where each SHAP value represents the average marginal contribution of each feature relative to the baseline prediction. In this study, positive SHAP values indicate that a feature increases the predicted GPP/NPP, whereas negative values indicate a decrease. By aggregating SHAP values across samples, we could get feature importance rankings and dependence relationships, enabling the direction and strength of impacts of environmental factors on GPP and NPP [5].

2.5. Model evaluation metrics

Model performance is evaluated using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and Bias. R^2 reflects the proportion of variance explained by the model, while RMSE and MAE quantify prediction errors. Bias measures systematic over- or under-estimation. The calculation formulas for these metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} \tag{1}$$

DOI: 10.54254/2755-2721/2026.MH35286

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{3}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{4}$$

where y_i is the true value (observed value), $\hat{y}_i$ is the predicted value, $\overline{y}$ is the mean of the true, and n is the number of observations.

3. Results

3.1. Spatiotemporal variations of GPP and NPP

Figure 1 maps land-cover types and the spatial distribution of multi-year mean GPP and NPP across the JJJ. Overall, productivity hotspots are concentrated in the forest belts in the north and west, especially in deciduous broadleaf and mixed forests. In contrast, relatively low values are observed across croplands and urban/built-up areas. For croplands, the lower long-term carbon sink signal is consistent with a shorter effective growing period and the annual export of a large fraction of aboveground biomass (and associated carbon) through harvesting. Urban areas show the weakest carbon fixation capacity due to extensive impervious surfaces and limited vegetated cover.

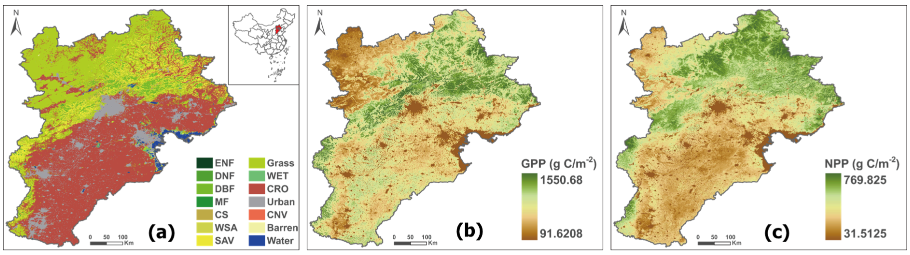

Figure 1. (a) Land cover types and spatial distributions of multi-year average (b) GPP and (c) NPP in the JJJ from 2001 to 2024

The interannual variations in regional GPP and NPP from 2001 to 2024 are shown on Figure 2. Both GPP and NPP exhibit upward trends over the study period, indicating an overall enhancement of ecosystem carbon uptake and sink. The mean annual increase is 9.79 g C m^{-2} yr^{-1} for GPP and 5.21 g C m^{-2} yr^{-1} for NPP. Mean GPP over the study period is 697.44 g C m^{-2} yr^{-1}, ranging from 542.06 g C m^{-2} yr^{-1} (2001) to 835.47 g C m^{-2} yr^{-1} (2022). Mean NPP is 341.09 g C m^{-2} yr^{-1}, increasing from 241.33 g C m^{-2} yr^{-1} in 2001 to 420.04 g C m^{-2} yr^{-1} in 2022. Despite the overall increase, both GPP and NPP show substantial annual variability. Such variability likely reflects climate fluctuations which can strongly modulate ecosystem functions and carbon sinks.

DOI: 10.54254/2755-2721/2026.MH35286

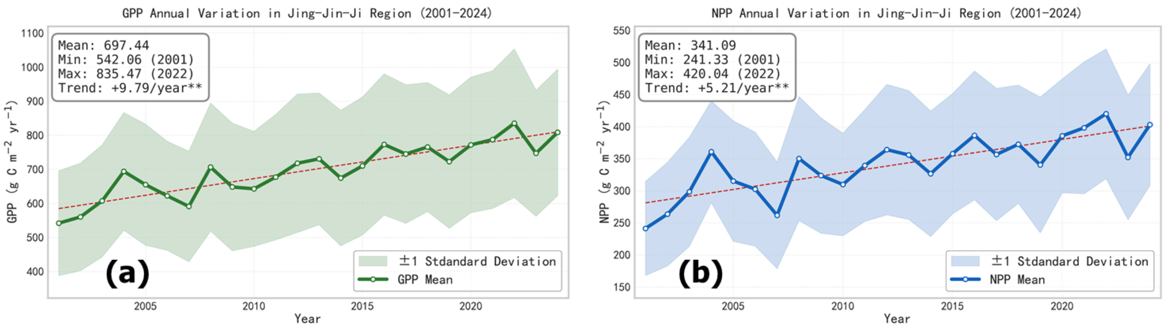

Figure 2. Annual variation of (a) GPP and (b) NPP in the JJJ from 2001 to 2024

3.2. Model performance of XGBoost

Figure 3 evaluates the predictive skill of the XGBoost models for GPP and NPP. For GPP, the model achieves R^2 = 0.72, RMSE = 102.3 g C m^{-2} yr^{-1}, MAE = 75.2 g C m^{-2} yr^{-1}, and Bias = 0.02 g C m^{-2} yr^{-1}, indicating that it explains ~72% of the variability with only a small positive bias. Most points lie close to the 1:1 line, and the highest point density is concentrated between 400 and 900 g C m^{-2}, consistent with the dominant GPP range in the region. The NPP model performs even better, with R^2 = 0.84, RMSE = 39.8 g C m^{-2} yr^{-1}, MAE = 29.37 g C m^{-2} yr^{-1}, and Bias = −0.02 g C m^{-2} yr^{-1}. The tighter clustering around the 1:1 line indicates improved agreement between predicted and observed NPP.

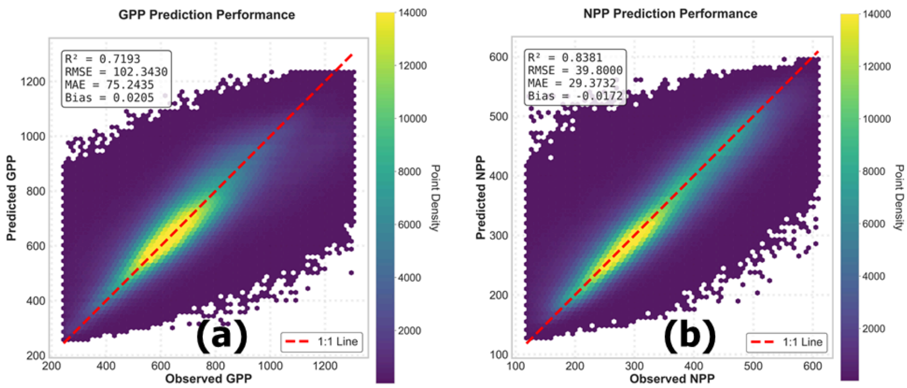

Figure 3. Prediction performance of XGBoost model for (a) GPP and (b) NPP

3.3. SHAP interpretation of GPP drivers

Figure 4 presents the SHAP-based attribution of key drivers of GPP. The SHAP summary plot ranks variables by importance and shows the direction of their effects. Relative humidity emerges as the dominant predictor, with the widest SHAP spread (approximately −0.15 to 0.10), highlighting the strong role of atmospheric moisture conditions in regulating photosynthesis. Soil moisture in the mid-depth layer (Soil Moisture L3) ranks second, underscoring the importance of root-zone water availability. Temperature ranks third; at higher temperatures, SHAP values tend to be negative, suggesting heat stress or increased physiological constraints that reduce GPP. The dependence plots reveal pronounced nonlinear responses. Precipitation shows a unimodal pattern: GPP benefits from increasing precipitation up to a moderate range (about 400–800 mm), while further increases may reduce effective radiation (e.g., via cloudiness) and dampen photosynthetic gains. Radiation generally enhances GPP, but the marginal benefit diminishes at very high radiation levels. Relative

DOI: 10.54254/2755-2721/2026.MH35286

humidity exhibits an optimal intermediate range (roughly 5–8 in the plotted scale), with reduced GPP contribution under both low and high humidity. Temperature shows positive contributions within a moderate window (5–12 °C) and negative contributions beyond that. Soil moisture effects are also nonlinear across depths: intermediate layers (L2–L3) exert stronger controls than the shallowest (L1) and deepest (L4) layers.

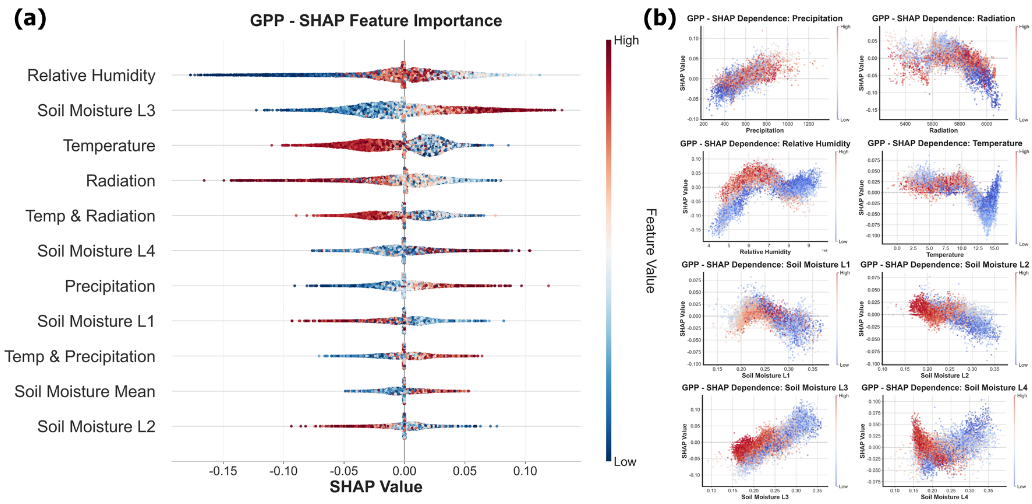

Figure 4. Result of SHAP (a) feature importance and (b) dependence for GPP

3.4. SHAP interpretation of NPP drivers

Figure 5 summarizes the SHAP results for NPP. In contrast to GPP, temperature is the most influential factor for NPP, consistent with its dual control on photosynthesis and autotrophic respiration. The temperature and radiation interaction (Temp with Radiation) is ranked second, indicating that the joint effect of thermal and light conditions is critical for net biomass accumulation. Soil moisture (Soil Moisture L3 and Soil Moisture Mean) also plays a major role, while radiation and relative humidity contribute at similar levels. Dependence plots indicate that precipitation has a nonlinear "increase-then-decrease" relationship with NPP, with the turning point occurring at higher precipitation (around 1000 mm) than for GPP. Radiation shows a strong positive association with NPP, reflecting enhanced net carbon gain under higher light availability. Relative humidity exhibits an inverted-U response, with an optimal range of approximately 6–10. NPP appears more temperature-sensitive: SHAP values are positive within an optimal interval (7.5–12.5 °C) and become negative outside this range. Soil moisture generally contributes positively to NPP across all depths, with the strongest effect in the surface layer (L1) and progressively weaker influence at greater depths.

DOI: 10.54254/2755-2721/2026.MH35286

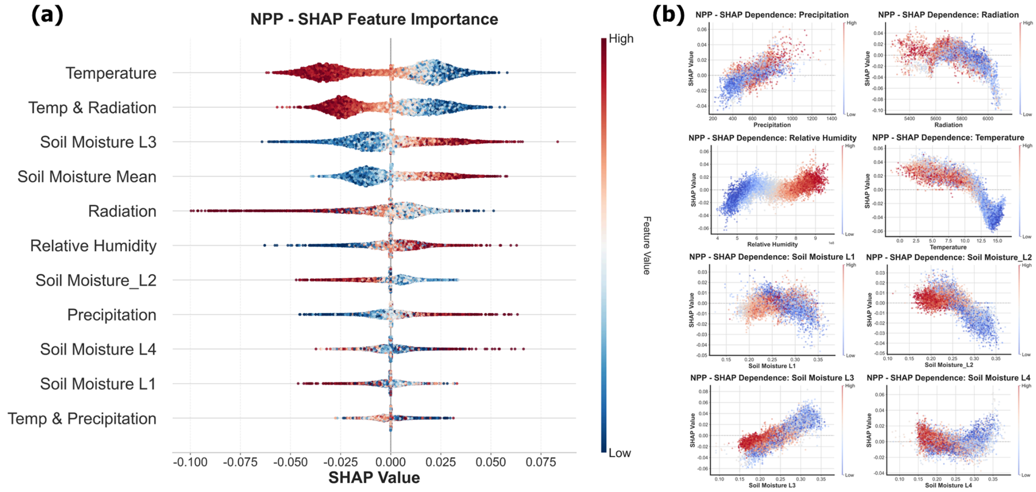

Figure 5. Result of SHAP (a) feature importance and (b) dependence for NPP

4. Discussion

4.1. Spatiotemporal patterns of carbon uptake in the JJJ

This study reveals pronounced spatiotemporal heterogeneity in ecosystem carbon uptake across the Beijing–Tianjin–Hebei region, as reflected by both GPP and NPP. Spatially, carbon sink strength is generally higher in the north and west and lower in the central and southeastern plains. This "north–high/south–low" pattern is closely aligned with the region's topographic gradient and land-cover configuration: mountainous areas support extensive forest ecosystems that typically exhibit higher leaf area, longer-lived biomass pools, and more stable water supply than intensively managed croplands and urban land, thereby enabling greater gross and net carbon assimilation. In contrast, cropland-dominated plains and urbanized areas show comparatively weak carbon sink signals, driven by shorter effective growing periods, harvest-related biomass export, and limited vegetated cover in built-up landscapes.

Temporally, both GPP and NPP show significant increasing trends from 2001 to 2024, indicating that regional carbon uptakes and sinks have been strengthened during the study period in JJJ. Multiple interacting environmental factors may collectively contribute to this outcome: climate change, the CO_2 fertilization effect, and changes in land cover and vegetation [6]. The complex interactions among factors make it extremely difficult to isolate their individual contributions to ecosystem carbon sinks. Rigorous attribution analysis must explicitly account for land-use changes, CO_2 emission processes, and climate-vegetation interactions [7].

4.2. Environmental controls and mechanisms inferred from SHAP

Results of SHAP values indicate that both GPP and NPP are jointly regulated by multiple environmental factors, but the dominant control factors differ between the two carbon fluxes, consistent with their physiological definitions. The generation of GPP requires active water consumption by vegetation, making water-related variables primary limiting factors, including relative humidity and soil moisture content. This aligns with the geographical characteristics of the JJJ region, situated in a transitional zone between semi-arid and semi-humid climates, where GPP exhibits extreme sensitivity to atmospheric moisture and soil moisture. Moreover, water limitation

DOI: 10.54254/2755-2721/2026.MH35286

directly impacts stomatal conductance and canopy photosynthesis from a mechanistic perspective, constraining leaf carbon assimilation even under favorable radiation and temperature [8]. Temperature dominates the NPP variation because NPP is influenced by both photosynthetic and respiratory processes in vegetation. Respiration intensifies dramatically under global warming, offsetting the carbon dioxide assimilated through photosynthesis and altering net carbon accumulation in vegetation. The high SHAP value assigned to the temperature-radiation interaction term for NPP further indicates that climate impacts on carbon sinks are not a simple summation but likely involve coupled effects. Enhancing radiation could increase vegetation carbon uptake by boosting photosynthetic radiation, but higher temperatures increase carbon losses from vegetation respiration, thereby reducing the carbon sink benefits derived from enhanced radiation [8].

The differential effects of soil moisture content at varying depths on GPP and NPP reveal hydrological regulation mechanisms in JJJ. The more pronounced effect of deep soil moisture on GPP indicates that sustained water supply within the root distribution zone in the JJJ region is crucial for maintaining photosynthetic capacity, particularly during dry periods. Conversely, the stronger contribution of shallow soil moisture to NPP suggests that near-surface water availability may be more tightly linked to net biomass accumulation at annual scales This relationship may be achieved through its influence on regulating respiration, supplying soil nutrients, and affecting phenological processes. All interpretations based on SHAP results, while reasonable, should be treated with caution. SHAP values can reflect direct effects but struggle to capture significant associations between variables that remain obscured. For instance, the interactive influence between temperature and radiation can be reflected, but the connection between soil water at different depths cannot be captured because it is not included in the analysis [9].

4.3. Limitations

Several limitations should be considered when interpreting the results. First, MODIS GPP/NPP products are model-based estimates and may contain systematic biases related to input meteorology, biome parameterization, and stress scalars. Future work should incorporate eddy-covariance flux observations and independent benchmarks to validate and bias-correct productivity estimates. Second, the current predictor set emphasizes meteorology and soil moisture but does not explicitly incorporate atmospheric CO2 concentration, nitrogen deposition, disturbances, irrigation, fertilization, cropping intensity, or urban expansion metrics. Excluding these factors can lead to omitted-variable bias and potentially misattribute variability to climate variables [8]. Third, XGBoost captures nonlinear relationships effectively, but SHAP only provides attribution within the model. Correlated predictors and spatiotemporal autocorrelation can complicate interpretation. Finally, annual datasets improve comparability but may obscure short-term extremes, such as heatwaves, drought, and floods, which can drive productivity anomalies. Seasonal scale analysis and extreme-event diagnostics would provide a more complete picture of climate sensitivity.

5. Conclusions

Using multi source satellite products and XGBoost–SHAP framework, this study quantified the spatiotemporal dynamics of GPP and NPP in the Beijing–Tianjin–Hebei region during 2001–2024 and diagnosed the dominant environmental controls. The main findings are:

(1) GPP and NPP exhibit clear geographic heterogeneity, with higher carbon sink strength in northern and western mountainous forest ecosystems and lower values in croplands and urban areas.

DOI: 10.54254/2755-2721/2026.MH35286

GPP ranges from 91.62 to 1550.68 g C m^{-2}, and NPP ranges from 31.51 to 769.83 g C m^{-2}, with hotspots concentrated in forest belts.

(2) Both GPP and NPP increase significantly from 2001 to 2024, with mean linear trends of 9.79 g C m^{-2} yr^{-1} for GPP ($p < 0.001$) and 5.21 g C m^{-2} yr^{-1} for NPP ($p < 0.001$), indicating sustained enhancements of regional carbon uptake and sink in the JJJ.

(3) XGBoost achieves strong predictive performance for both carbon fluxes (R^2 = 0.7190 for GPP and R^2 = 0.8381 for NPP) in JJJ, demonstrating that it can capture complex nonlinear relationships between environmental factors and ecosystem productivity.

(4) SHAP value results indicate that relative humidity and soil moisture are the primary constraints on GPP, while temperature and its interaction with radiation are key drivers of NPP. The relationship between environmental factors and carbon sink indicators exhibits strong nonlinear characteristics, highlighting the importance of comprehensively considering various environmental factors when assessing the response of carbon sinks to environmental conditions.

Overall, our findings provide a comprehensive and explainable assessment of carbon sink dynamics in the Beijing-Tianjin-Hebei region, offering scientific basis for implementing ecosystem carbon management and climate adaptation strategies in this region.

References

[1] Yue, X. et al. Large potential of strengthening the land carbon sink in China through anthropogenic interventions. Sci. Bull. S2095927324003876 (2024) doi: 10.1016/j.scib.2024.05.037.

[2] Wu, X., Zhang, Y. & Li, X. Exploring the Relationship between Urbanization and Eco-Environment Using Dynamic Coupling Coordination Degree Model: Case Study of Beijing–Tianjin–Hebei Urban Agglomeration, China. Land 13, 850 (2024).

[3] Sitch, S. et al. Trends and Drivers of Terrestrial Sources and Sinks of Carbon Dioxide: An Overview of the TRENDY Project. Glob. Biogeochem. Cycles 38, e2024GB008102 (2024).

[4] 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 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). doi: 10.1145/2939672.2939785.

[5] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in Proceedings of the 31st International Conference on Neural Information Processing Systems 4768–4777 (Curran Associates Inc., Red Hook, NY, USA, 2017).

[6] Zhu, Z. et al. Greening of the Earth and its drivers. Nat. Clim. Change 6, 791–795 (2016).

[7] Zhang, B., Tian, L., Zhao, X. & Wu, P. Feedbacks between vegetation restoration and local precipitation over the Loess Plateau in China. Sci. China Earth Sci. 64, 920–931 (2021).

[8] Harris, N. L. et al. Global maps of twenty-first century forest carbon fluxes. Nat. Clim. Change 11, 234–240 (2021).

[9] Wankmüller, F. J. P. et al. Global influence of soil texture on ecosystem water limitation. Nature 635, 631–638 (2024).