Developing a High-resolution Gross Primary Productivity Product using Sentinel-2 and Machine-learning Algorithms in the Golden Gate Highlands National Park, South Africa

Morena Mapuru[1]*, Mahlatse Kganyago[2], Katlego Mashiane[3], Sifiso Xulu[4]
[1]Department of Physical and Earth Sciences, School of Natural and Applied Sciences, Arid Region Water Research Centre, Sol Plaatje University, Kimberley, South Africa
[2]Department of Geography, Environmental Management and Energy Studies, University of Johannesburg, Johannesburg, South Africa
[3]Department of Geography, University of the Free State, Phuthaditjhaba, South Africa
[4]Department of Geography, University of South Africa, Florida Campus, Johannesburg, South Africa.
Corresponding author: johnmapuru97@gmail.com

DOI: https://dx.doi.org/10.4314/sajg.v15i1.4

Abstract

Gross Primary Productivity (GPP) is a key indicator of ecosystem functioning and carbon sequestration. GPP has been estimated using several tools, including remote sensing, one of which is the Moderate Resolution Imaging Spectroradiometer (MODIS). Despite its coarse spatial resolution limiting effectiveness in capturing finer landscape-scaled ecosystem and carbon dynamics, this tool is widely used. This study aims to develop the first high-resolution GPP product by downscaling MODIS GPP to Sentinel-2 spatial resolution using machine-learning algorithms and multi-source predictors in the Golden Gate Highlands National Park (GGHNP) in the eastern Free State of South Africa.

The objectives of the study were to (a) determine the relationship between GPP and selected predictor variables (viz.., Sentinel-2 spectral bands, vegetation indices, and topographical variables); (b) evaluate the performance of Random Forest (RF) and Support Vector Machine (SVM) algorithms in downscaling the MODIS GPP product to Sentinel-2 resolution under the selected predictor combinations; and (c) identify important predictors using RF variable importance analysis. The results showed strong positive relationships between GPP and the selected vegetation indices (r >0.6), including the Chlorophyll Red-edge Index (CI-Red-Edge), the Modified Soil-adjusted Vegetation Index (MSAVI), the Normalised Difference Red-edge Index (NDRE), and the Green NDVI (GNDVI). In contrast, the blue and green spectral bands were negatively correlated with GPP (r ~ -0.6). Moreover, the modelling results showed that the SVM model, comprising bands and indices (R^2 = 0.68, RMSE = 10.07 g C m^{-2}), was superior to the RF model (R^2 = 0.60, RMSE = 10.2 g C m^{-2}), but inferior when topographical variables were combined with spectral bands and vegetation indices. The key predictors identified through variable importance analysis included the blue and green bands, the Enhanced Vegetation Index (EVI), the CI-Red-edge, MSAVI, and NDRE indices. These results highlight the value of the high-resolution bands for Sentinel 2, as well as the red-edge and near-infrared bands, in modelling vegetation productivity. This study demonstrates the potential of combining high-resolution remote sensing data with machine-learning algorithms to estimate GPP in data-scarce mountainous regions and provides the first localised model for

generating vegetation productivity estimates in nature reserves such as the GGHNP. Overall, the approaches offer a scalable tool for improving ecological monitoring and understanding carbon dynamics in mountainous environments.

***Keywords**: Gross Primary Productivity, Remote Sensing, Machine Learning, Sentinel-2, Vegetation Indices*

1. Introduction

Gross Primary Productivity (GPP) represents the total amount of carbon dioxide assimilated by vegetation through photosynthesis and is a fundamental component of the global carbon cycle (Liao *et al*., 2023). Accurate and spatially explicit estimations of GPP are essential for understanding ecosystem productivity, the influence of climate change, and terrestrial carbon dynamics (Hari and Tyagi, 2022). Traditional methods for estimating GPP, such as flux towers and ground-based measurements, offer high temporal precision but are limited in spatial coverage (Sabzchi-Dehkharghani *et al*., 2024). Alternatively, satellite remote sensing provides cost-effective, synoptic, continuous, and scalable observations that can support GPP assessments at varying geographic scales (Maleki *et al*., 2020; Celis *et al*., 2023).

Recent advancements in high-resolution satellite imagery, particularly from the Sentinel-2 mission, have greatly expanded the potential for GPP modelling (Spinosa *et al*., 2023). Sentinel-2 provides fine spatial resolutions (10–20 m), three to five-day temporal frequencies, and up to 13 spectral bands, traversing the visible to shortwave infrared regions (Segarra *et al*., 2020). These spectral bands are highly responsive to key vegetation characteristics such as canopy structure, chlorophyll content, water dynamics, and photosynthetic activity, making them uniquely valuable for deriving vegetation indices and biophysical parameters critical to GPP estimation (Pabon-Moreno *et al*., 2022). Various vegetation indices derived from Sentinel-2 data, including the Normalised Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Normalised Difference Moisture Index (NDMI), Soil-adjusted Vegetation Index (SAVI), Modified Soil-adjusted Vegetation Index (MSAVI), and multiple red-edge-based indices, have demonstrated strong correlations with photosynthetic capacity and green biomass accumulation (Liu *et al*., 2021). These indices serve as reliable proxies for GPP and their widespread use has matured in remote sensing-based productivity models.

The commonly used NDVI is limited by saturation issues, and many recent studies have since shifted toward better indices such as EVI for estimating GPP. EVI offers improved sensitivity in high biomass regions and effectively reduces atmospheric and soil background noise (Zeng *et al*., 2022). In sparsely vegetated areas, where soil reflectance is high, indices such as the SAVI (Huete *et al*., 1988) and the MSAVI (Qi *et al*., 1994) are particularly valuable. Furthermore, red-edge-based indices such as the Chlorophyll Index Red Edge (ClRE) (Zhang *et al*., 2022) and the Red-edge Normalised Difference Vegetation Index (RENDVI) (Lin *et al*.,

2022) have demonstrated greater sensitivity and accuracy in estimating GPP, especially in areas of dense vegetation with high chlorophyll concentrations. Such indices are often more effective than their traditional counterparts. Beyond spectral data, topographic variables such as elevation, slope, and aspect have exerted a notable influence on local microclimates and vegetation productivity (Li *et al*., 2023). Topography influences the energy budget, thereby affecting factors such as solar radiation, soil moisture, and temperature gradients, all of which play a crucial role in plant physiological processes and growth (Huang *et al*., 2021). Incorporating both topographic and spectral information into remote sensing-based models can substantially improve the accuracy and spatial detail of GPP estimates, particularly in topographically diverse landscapes such as Golden Gate Highlands National Park (GGHNP).

Machine-learning techniques have revolutionised remote sensing data analytics and provide a powerful framework for capturing complex, nonlinear relationships with topographic variables and ground-based observations of ecosystem productivity. Random Forest (RF) and Support Vector Machine (SVM) algorithms stand out as high-performing and particularly well-suited for ecological modelling on account of their ability to manage high-dimensional datasets and to model nonlinear interactions without relying on strict statistical assumptions (Sarkar *et al*., 2022). For example, the application of RF algorithms has been successful in estimating terrestrial GPP in that they incorporate both topographic and spectral variables, thereby achieving satisfactory predictive accuracies (Sarkar *et al*., 2022). Similarly, global studies applying RF models based on light-use efficiency principles have demonstrated strong performance in predicting GPP (Wei *et al*., 2017). Yang *et al*. (2007), for example, developed a continental-scale GPP model by integrating MODIS and AmeriFlux data using an SVM approach. More details on various machine-learning applications on GPP are available from Zhang *et al*. (2019). These machine-learning models can be trained using coarse-resolution GPP products, such as MODIS GPP (MOD17A2H), and then applied to higher-resolution datasets, such as Sentinel-2, to enhance spatial detail in productivity estimates. Although this approach has shown considerable success in various contexts, its efficacy in topographically heterogeneous environments such as GGHNP is yet to be evaluated. The complexity of GGHNP not only presents challenges for accurately modelling ecosystem productivity through traditional techniques but also offers an opportunity to bridge GPP data gaps in this area under various experimental scenarios. Thus, this research represents the first attempt to develop a high-resolution GPP product on the basis of an integrated, data-driven approach.

To address the identified research gap, this study aimed to develop a high-resolution GPP product downscaled from MODIS to a Sentinel-2 10 m resolution. Firstly, several experiments were conducted using spectral bands, vegetation indices, topographical variables, and their combinations to discern the impact of various predictors on predictive accuracy. Secondly, the study evaluated the potential of RF and SVM algorithms for modelling GPP at finer resolutions.

Finally, the most influential predictors contributing to GPP estimation were identified using variable importance analysis.

2. Methods and materials

2.1. Description of the study area

The study was carried out in the Golden Gate Highlands National Park (latitude 28°30′22″S, longitude 28°37′0″E), located in the eastern Free State Province of South Africa (Figure 1). Covering approximately 340 km², the park lies at the base of the Maloti Mountains, with the highest peak reaching 2,829 metres above sea level (SANParks, 2020). The region experiences a dry June-July-August season, while rainfall varies between October and April, with an average annual precipitation of approximately 800 mm (Kay *et al*., 1993). Geologically, the park comprises five formations, four of which are sedimentary and one is igneous (SANParks, 2020). It falls within South Africa's grassland biome, specifically within the Drakensberg grassland and the mesic highland grassland bioregions (SANParks, 2020). The two dominant plant families in the park are *Poaceae* and *Asteraceae*. The grassland landscape is gently undulating, traversed by streams and rivers, primarily dominated by the *Eragrostis* species, Tristachya *leucothrix*, and *Themeda triandra*. The herbaceous layer is mainly composed of *Asteraceae* species (SANParks, 2020). The park is situated on the watershed between the Vaal and Orange River systems and plays a vital role in water provision. This region supplies high-quality water to the Gauteng region through the Lesotho Highlands Water Project and the Tugela-Vaal transfer scheme. Additionally, the Little Caledon and Klerkspruit Rivers, originating within the park, contribute approximately 30% to the total water supply in Southern Africa.

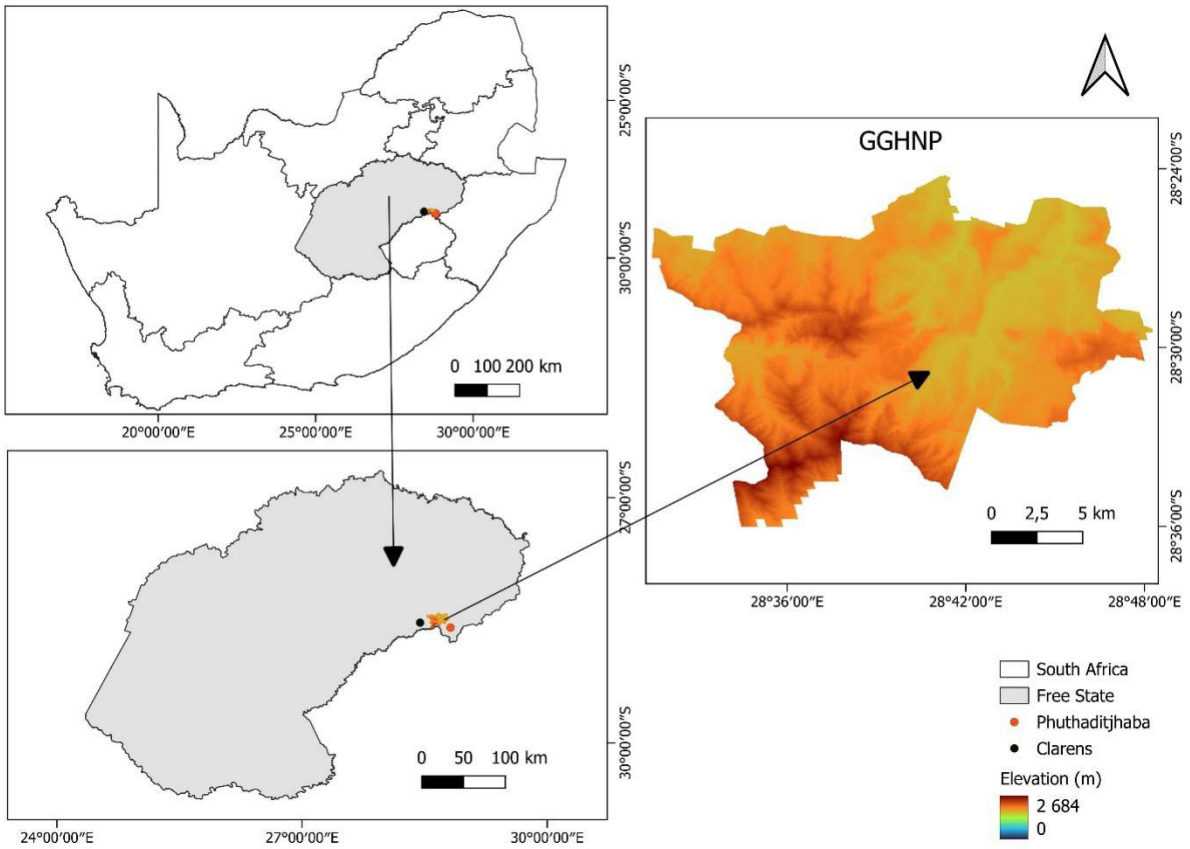

Figure 1: Location of the study area (Golden Gate Highlands National Park) in the context of Free State province and South Africa

2.2. Data sourcing and pre-processing

MODIS Gross Primary Productivity (GPP, MODIS/061/MOD17A2H) data at a spatial resolution of 500 m were obtained from Google Earth Engine (GEE). To match the spatial resolution of the predictors derived from the Sentinel-2 imagery, the data were resampled to a 10 m resolution using bilinear interpolation. Bilinear interpolation values were computed as a weighted average of the four nearest neighbouring pixels, thereby resulting in smoother spatial gradients and a reduction in the blocky appearance associated with the nearest-neighbour method. This approach preserves overall spatial patterns while minimising abrupt transitions and avoiding the introduction of unrealistic pixel values (Shankar *et al*., 2024). MODIS GPP data were collected from January to December 2024. Sentinel-2 Surface Reflectance (SR) images, covering a similar period, were also sourced from GEE. This timeframe was chosen to cover all phenological phases. All spectral bands were used, except for bands 1, 9, and 10, which are atmospheric bands and therefore unsuitable for vegetation analysis. Sentinel-2 bands 5, 6, 7, 11, and 12 (with 20 m native resolution) were resampled to a 10 m spatial resolution to ensure spatial consistency with the remaining spectral bands. A simple median-compositing approach was then applied to all images acquired during the study period to generate a representative composite. The median composite is robust to outliers and effectively reduces

the influence of anomalous observations, thereby preserving reflectance values that are closest to the central tendency (Mapuru *et al*., 2025). The resampled MODIS GPP was used as the response variable in this study. In addition, a Digital Elevation Model (DEM) with a spatial resolution of ~ 90 m was obtained from WorldClim (https://www.worldclim.org/data/worldclim21.html) and resampled using a bilinear interpolation technique to a 10 m resolution to derive the topographical variables, which included slope, aspect, and elevation indices, and the Terrain Ruggedness Index (TRI). Sentinel-2 SR imagery was used to calculate a range of vegetation and moisture indices, as detailed in Table 1.

The integration of Sentinel-2-derived vegetation and moisture indices is crucial for improving the prediction of GPP because these indices collectively capture canopy structure, biochemical properties, and environmental conditions that regulate photosynthesis. Greenness-based indices such as NDVI, EVI, SAVI, MSAVI, TVI, and GNDVI are effective proxies for leaf area index, canopy density, and photosynthetically active biomass, and have demonstrated strong relationships with flux tower-derived GPP across diverse ecosystems (Maluleke *et al*., 2024). Sentinel-2 red-edge indices, including NDRE, the Chlorophyll Red-edge Index (Cl-red edge), and the Red-edge Chlorophyll Index (RECI), exploit the sensor's narrow red-edge bands to enhance sensitivity to chlorophyll concentration and photosynthetic capacity, particularly in dense vegetation areas. The traditional indices tend to saturate under high Leaf Area Index and Biomass conditions (Croft *et al*., 2020; Peng *et al*., 2021). Pigment-specific indices such as the Anthocyanin Reflectance Index (ARI) provide additional information on plant stress and photoprotective responses, while moisture-related indices such as NDMI capture the water status of the vegetation − a key control on stomatal conductance, evapotranspiration, and carbon assimilation (Sett *et al*., 2024). Thus, the combined use of these Sentinel-2-based indices enables a more comprehensive representation of biophysical and physiological controls on GPP, resulting in more robust and accurate satellite-driven GPP estimations.

Table 1: Sentinel-2 spectral bands and indices and their respective spatial resolutions

Predictors	**Formulae**	**Spatial resolution** ***(Native Resolution)***
Bands		
Blue	$B2$	10 m
Green	$B3$	10 m
Red	$B4$	10 m
RedEdge 1	$B5$	10 m *(20 m)*
RedEdge 2	$B6$	10 m *(20 m)*
RedEdge 3	$B7$	10 m *(20 m)*
Near-Infrared (NIR)	$B8$	10 m
RedEdge-NIR	$B8A$	10 m *(20 m)*
Shortwave Infrared (SWIR1)	$B11$	10 m *(20 m)*
Shortwave Infrared (SWIR2)	$B12$	10 m *(20 m)*
Indices		
Chlorophyll RedEdge Index (Cl-red-edge)	$\frac{RedEdge}{Red} - 1$	10 m
Modified Soil -adjusted Vegetation Index (MSAVI)	$\frac{2.NIR + 1 - \sqrt{(2.NIR + 1)^2 - 8(NIR - Red)}}{2}$	10 m
Soil-aAdjusted Vegetation Index (SAVI)	$\frac{(NIR - Red).(1 + L)}{NIR + Red + L}$	10 m
Enhanced Vegetation Index (EVI)	$2.5.\frac{(NIR - Red)}{(NIR + 6.Red - 7.5.Blue + 1)}$	10 m
Normalised Difference Vegetation Index (NDVI)	$\frac{(NIR - Red)}{(NIR + Red)}$	10 m
Green Normalised Difference Vegetation Index (GNDVI)	$\frac{(NIR - Green)}{(NIR + Green)}$	10 m
Normalised Difference Moisture Index (NDMI)	$\frac{(NIR - SWIR)}{(NIR + SWIR)}$	10 m
Normalised Difference RedEdge (NDRE)	$\frac{(NIR - RedEdge)}{(NIR + RedEdge)}$	10 m
Transformed Vegetation Index (TVI)	$\sqrt{\frac{(NIR - Red)}{(NIR + Red)}} + 0.5$	10 m
RedEdge Chlorophyll Index (RECI)	$\frac{RedEdge}{Red} - 1$	10 m
Anthocyanin Reflectance Index (ARI)	$\frac{(Red - Blue)}{(Red + Blue)}$	10 m

2.3. Sampling, data extraction, and normality testing

To create the training and validation datasets, 100 random points were generated throughout the study area using the 'Random Point in Polygon' tool in QGIS Desktop 3.40.11 (https://qgis.org/download), ensuring a spatially representative sample. These points were then used to extract both the presence data and associated predictors via the 'Sample Raster Values' tool, which automatically retrieves raster values from each predictor layer for every point (Mapuru *et al*., 2024). The resultant dataset was exported to an Excel spreadsheet for subsequent analysis in an RStudio Integrated Development Environment (Version 4.4.2; https://posit.co/products/open-source/rstudio/?sid=1). Prior to model development, data normality was assessed to inform the selection of appropriate statistical and machine-learning techniques. The Shapiro-Wilk test, along with visual assessments using histograms and Q-Q plots, was employed to evaluate the data distribution (Khatun, 2021). The results confirmed that the dataset is approximately normally distributed, indicating that no transformation was necessary for modelling. To improve model reliability, multicollinearity among the predictors was examined. This was done using the *dredge* function in R, which systematically identifies highly correlated predictors that may affect model performance (Cushman *et al*., 2024). The analysis confirmed that no variables were redundant, allowing all predictors to be retained for model training.

2.4. Modelling

Gross Primary Productivity (GPP) was modelled using two widely used machine-learning algorithms: Random Forest (RF) and Support Vector Machine (SVM). To ensure a robust and unbiased evaluation of model performance, the dataset was divided into a training set (80%) and a validation set (20%). The models were also subjected to a 10-fold cross-validation, which assesses how well predictive models generalise to unseen data. This procedure is crucial for evaluating the performance and robustness of machine-learning models, thereby helping to prevent overfitting (Yates *et al*., 2023).

Random Forest, an ensemble-based algorithm, generates multiple decision trees using different random subsets of the dataset (Salman *et al*., 2024). For regression tasks, the final prediction is derived by averaging the outputs of all the trees (Breiman, 2001). This method helps reduce overfitting and increases prediction accuracy by introducing randomness in both data sampling and feature selection. In this study, RF was used to estimate GPP using topographical variables, spectral bands, and vegetation indices derived from Sentinel-2 imagery. The model was trained on the training dataset. To optimize the performance of the models, hyperparameters were set as follows: number of trees (ntrees = 1,000), node size = 5 (default value), *mtry* = default value and importance = T (TRUE) (Vincenzi *et al*., 2011, Mapuru *et al*., 2024). A variable importance analysis was also conducted to determine which spectral bands, indices, and topographic variables had the greatest influence on GPP prediction

using the Percentage Increase in Mean Squared Error (%InMSE) value (Mapuru *et al*., 2024). Model performance was assessed using the coefficient of determination (R^2) and the Root Mean Square Error (RMSE), where a higher R^2 and a lower RMSE indicate better predictive performance (Pang *et al*., 2022).

The Support Vector Machine (SVM), a supervised learning technique, operates by identifying an optimal hyperplane that best fits the data while minimising prediction errors (Gaye *et al*., 2021). For regression analysis, the algorithm seeks a function that fits the data within a defined margin, thereby helping to avoid overfitting. This study employed the Radial Basis Function (RBF) kernel, which is well-suited for capturing complex nonlinear relationships among spectral variables, vegetation indices, topographical variables, and GPP values. The SVM model was trained on the 80% training dataset, and key parameters, such as cost (C) =10, gamma (γ) = 0.5, and epsilon = 0.1, were carefully tuned to enhance predictive performance by controlling model complexity, spatial sensitivity, and tolerance to noise. The evaluation of the model was conducted using the remaining 20% of the validation dataset, with R^2 and RMSE used as performance indicators (Tatachar, 2021).

Finally, the study compared the predictive capabilities of the RF and SVM models by examining how well each algorithm explained the variability in GPP. These machine-learning algorithms were chosen because they excel at handling complex, non-linear relationships between remote sensing data and topographical factors with high accuracy, often outperforming traditional models, especially in varied terrain where topography and vegetation indices change rapidly. RF handles high-dimensional data well and is robust, while SVM offers strong generalization, making both ideal for capturing the intricate spatial patterns of GPP in challenging environments, such as mountainous parks.

3. Results

3.1. Relationship between GPP and Predictors

The study quantified the relationship between Gross Primary Productivity (GPP) and the selected predictor variables using correlation analysis. The results are displayed in Figure 2. Several patterns are noteworthy, where strong positive correlations ($r > 0.6$; $p < 0.05$) with GPP were observed for the CI-Red-edge index, MSAVI, NDRE, GNDVI, EVI, and ARI (Figures 2a and 2b). In contrast, the green and blue bands exhibited strong negative correlations ($r \sim -0.5$ to -0.6; $p<0.05$) (Figure 2b), while aspect exhibited moderate negative correlations ($r \sim -0.3$; $p <0.05$) with GPP. Weak positive correlations were found with SWIR (1 and 2), red, NIR, Red-Edge-1, -2, and -3, Red-Edge-NIR, and slope (Figure 2a). Meanwhile, elevation, RECI, terrain ruggedness, NDMI, SAVI, and NDVI showed weak negative correlations with GPP (Figure 2a).

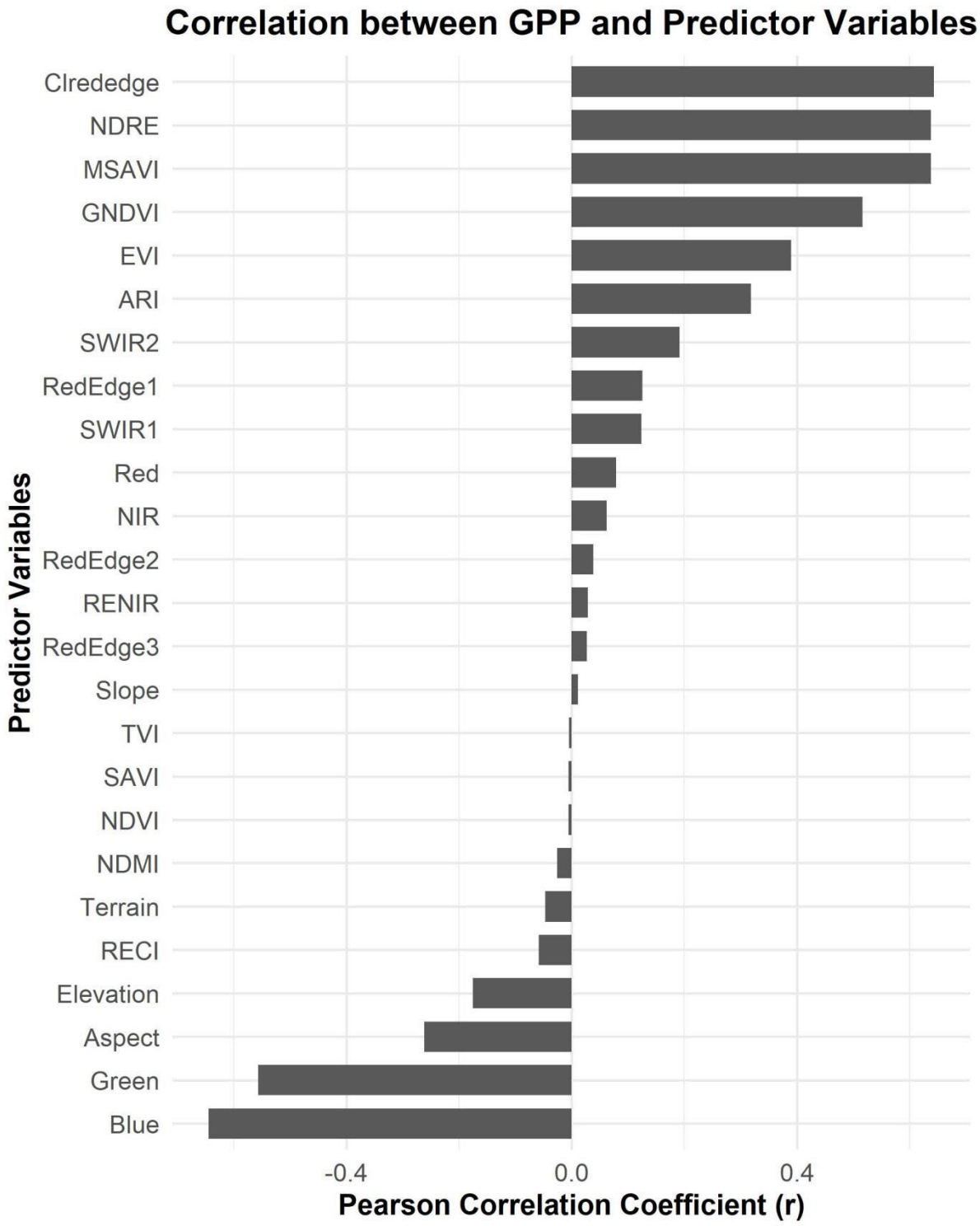

Figure 2a: The relationship between GPP and all predictors (spectral bands, spectral indices, and topographic variables)

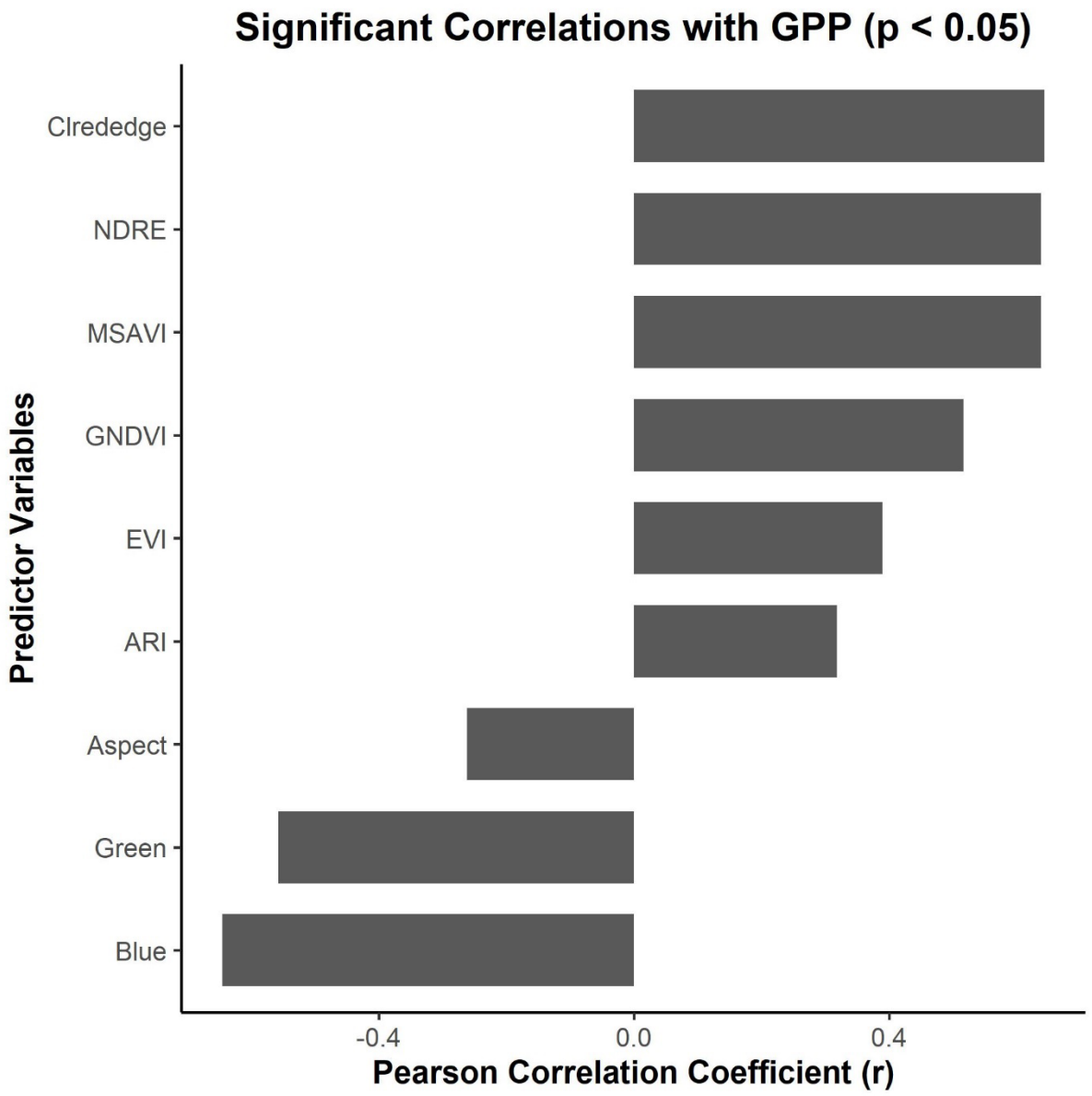

Figure 2b: Predictors with significant correlations with GPP ($p<0.05$)

3.2. Model performance

The performance of the Random Forest (RF) and Support Vector Machine (SVM) models in predicting Gross Primary Productivity (GPP) was evaluated using a variety of predictors, including spectral bands, vegetation indices, topographical variables, and their combinations (see Table 2). Model performance was assessed using the coefficient of determination (R^2) and the Root Mean Square Error (RMSE). The SVM model consistently outperformed the RF model when predicting GPP using only spectral bands, only indices, or a combination of both (see Table 2). However, both models exhibited poor performance when topographical variables were used as predictors on their own. Notably, the RF model showed superior results when combining topographical variables with either spectral bands or indices, or when integrating all predictors together (see Table 2). Overall, the SVM exhibited the best performance when combining bands with indices to an R^2 of 0.68 and an RMSE of 10.07 g C m^{-2}. On the other

hand, RF displayed similar but relatively low performance when topographical variables were combined with indices (R^2 = 0.6, RMSE = 10.2 g C m^{-2}).

Table 2: Model predictive performance per set of predictors. The best models (as a trade-off between R^2 and RMSE [g C m^{-2}]) are highlighted in bold.

Predictors	**RF**		**SVM**	
	R^2	RMSE	R^2	RMSE
Bands (*p* =10)	0.32	14.27	**0.39**	**13.28**
Indices (*p* = 11)	0.45	13.95	**0.52**	**13.45**
Bands and Indices (*p* = 21)	0.56	11.98	**0.68**	**10.07**
Topography (*p* = 4)	0.02	15.68	**0.02**	**14.87**
Topography and Bands (*p* = 14)	**0.25**	**21.43**	0.15	22.77
Topography and Indices (*p* = 15)	**0.60**	**10.2**	0.48	11.57
All Predictors (*p* = 25)	**0.46**	**13.12**	0.39	14.1

3.3. Variable importance

The Random Forest (RF) analysis of variable importance, based on the Percentage Increase in Mean Squared Error (%InMSE), identified the blue band, EVI, MSAVI, green band, Red-Edge2, and NDRE as the most influential predictors of GPP, exhibiting %InMSE >5. The other predictors, such as Aspect, CI-Red-Edge, Red-Edge3, RENIR, Slope, Elevation, SWIR2, and SAVI, had a relatively moderate influence on the model performance with %InMSE closer to 3, while other variables had %InMSE values closer to 0 (Figure 3).

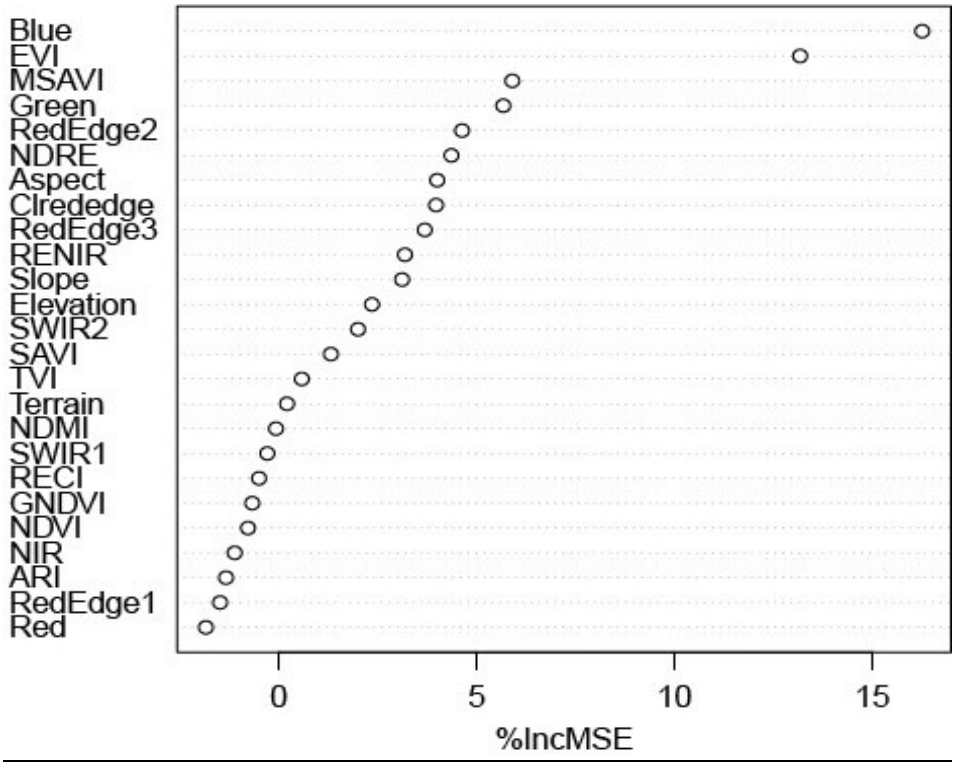

Figure 3: Random Forest (RF) variable importance using Percentage Increase in Mean Squared Error (%InMSE)

3.4. GPP distribution maps from the best-performing models

Figure 4 presents the spatial distribution of GPP predicted by the best-performing SVM and RF models. Both maps reveal pronounced spatial variability in GPP across the study area, indicating strong heterogeneity in vegetation productivity. The SVM-derived GPP map (Figure 4a), generated using spectral bands and vegetation indices, shows distinct spatial contrasts, with higher GPP values concentrated along drainage networks, valley bottoms, and areas characterised by denser vegetation cover. Lower GPP values are predominantly observed in upland and more sparsely vegetated regions, particularly across areas with rugged terrain. The spatial patterns are sharply defined, reflecting the sensitivity of the model to spectral variability and the condition of the vegetation.

The RF-derived GPP map (Figure 4b), produced using topographic variables in combination with vegetation indices, exhibits broadly similar spatial patterns. Elevated GPP values are again evident in low-lying areas and along riparian zones, while reduced productivity dominates higher-elevation and topographically-complex regions. Compared to the SVM output, the RF map displays smoother spatial transitions and a more gradual variation in GPP values, suggesting the moderating influence of terrain variables on predicted productivity.

Overall, both models produce ecologically meaningful GPP distributions, with high productivity associated with moisture and highly vegetated landscapes and low productivity

linked to topographical constraints. The consistency between the two maps supports the robustness of the GPP predictions, while differences in spatial sharpness reflect the distinct modelling approaches and predictor sets used.

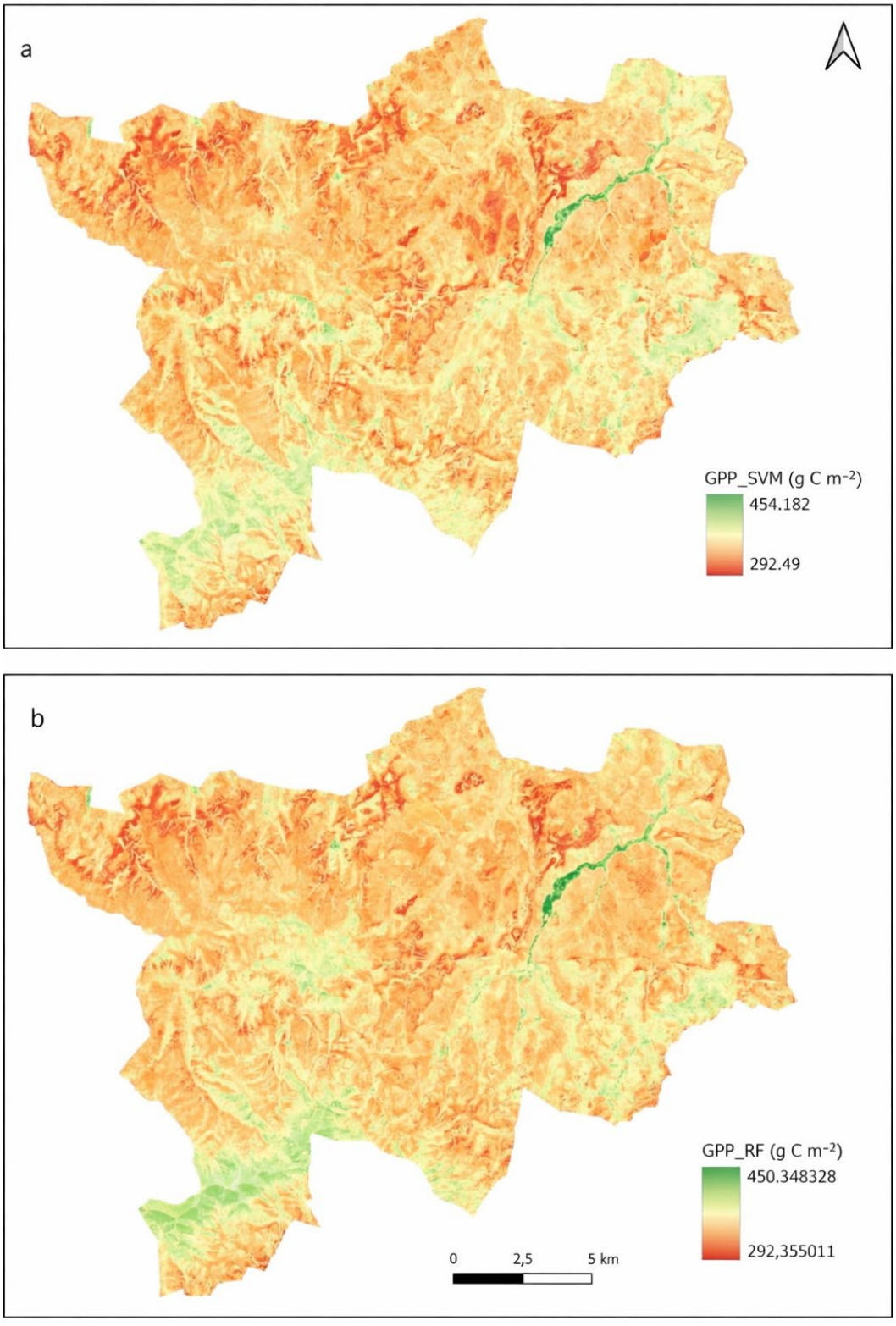

Figure 4: GPP distribution maps produced by the best-performing models: (a) SVM using spectral bands and vegetation indices; (b) RF using topographic variables and vegetation indices.

4. Discussion

This study successfully modelled the relationships between GPP and various predictors, and the results are generally consistent with the literature on remote sensing of ecosystem productivity. Strong positive correlations were found for the Cl-Red-Edge, MSAVI, NDRE, and GNDVI indices, thereby emphasizing the value of red-edge and near-infrared sensitive indices in detecting canopy chlorophyll content and biomass (Lin *et al*., 2019; Maleki *et al*., 2020). These indices capture subtle variations in vegetation structure and function, which are critical to accurately estimating photosynthetic capacity and, by extension, GPP.

Conversely, the observed strong negative correlations between GPP and the blue and green bands corroborates similar studies such as Zhang *et al*. (2015), highlighting the sensitivity of these bands to atmospheric scattering and non-vegetative surfaces, thereby reducing their reliability as stand-alone predictors of vegetation productivity. The limited explanatory power of commonly used indices such as NDVI and SAVI in this study also mirrors the recent works of Mallick *et al*. (2024) and Liu *et al*. (2023), who noted that NDVI tends to saturate in areas of high biomass and may fail to capture subtle variations in photosynthetic activity across more productive landscapes.

Topographic variables, such as elevation, aspect, and terrain ruggedness, showed weak negative correlations with GPP, reflecting their indirect and often contextually dependent influence on vegetation through microclimate, soil properties, and water availability (Hua *et al*., 2021; Ding *et al*., 2023). As opposed to GPP, elevation is one of the most studied topographic variables. Higher elevations are often associated with lower temperatures, shorter growing seasons, and reduced soil fertility, all of which can limit vegetation growth and productivity. For example, in alpine grasslands, the start of the growing season is delayed at higher elevations, leading to a shorter growing season and lower GPP (Hua *et al*., 2021). However, the relationship between elevation and GPP is not always straightforward. In some regions, such as the Virunga Volcanoes Massif, optimal vegetation growth is observed at intermediate elevations (2000–2800 m), where temperatures and soil moisture conditions are more favourable (Kayiranga *et al*., 2017). This highlights the contextually dependent nature of the relationship between elevation and GPP.

Aspect, which refers to the direction in which a slope faces, influences the amount of solar radiation a site receives, thereby affecting the microclimate and productivity of vegetation. South-facing slopes in the northern hemisphere generally receive more solar radiation, leading to warmer and drier conditions, which, depending on the ecosystem characteristics, can either enhance or reduce GPP. For example, in the Tianshan Mountains, south-facing slopes tend to have earlier vegetation green-up dates compared to north-facing slopes, indicating higher productivity in sunnier areas (Ding *et al*., 2023). In contrast, in the alpine tundra of the Rocky Mountains, north-facing slopes often support more diverse and productive plant communities

owing to the cooler and moister conditions, which are favourable for certain alpine species (Malanson *et al*., 2024). This underscores the importance of aspect in creating microrefugia for plant communities in a warming climate.

Terrain ruggedness, which refers to the complexity of the landscape, influences water redistribution and soil properties. In rugged terrains, water tends to accumulate in lower-lying areas, leading to higher levels of soil moisture and potentially higher GPP in these regions. For example, in the Atlantic Forest of Brazil, greater topographic variability is associated with greater soil fertility and higher GPP, as the uneven terrain promotes the retention of nutrients and water (Rodrigues *et al*., 2021). However, in some cases, rugged terrain can also lead to water scarcity on the upper slopes, thereby reducing GPP. In the Three Parallel Rivers Region of the Qinghai-Tibet Plateau, the interaction between slope and aspect determines the spatial distribution of vegetation, with steeper slopes often having lower NDVI values owing to limited water availability (Wang *et al*., 2021). This reinforces the notion that while topography shapes vegetation patterns on alarger scale, its direct influence on productivity metrics, such as GPP, is often secondary to spectral and physiological indicators.

Based on the characteristics of the dataset, the selection of features, and the tuning of the model, the performance of RF and SVM in predicting GPP can vary significantly. Overall, this study found that RF performed well when topographical variables were combined with spectral bands (R^2 = 0.25 and RMSE = 21.43 g C m^{-2}) or indices (R^2 = 0.60 and RMSE = 10.2 g C m^{-2}), or when all available variables (R^2 = 0.46 and RMSE = 13.12 g C m^{-2}) were used together. Likewise, Sarkar *et al*. (2022) reported improved model accuracy when topographic and remote sensing variables were integrated, with an R^2 = 0.82 and RMSE = 1.52 g C m-1d^{-1}. On the other hand, SVM performed better when using bands (R^2 = 0.39 and RMSE = 0.32 g C m^{-2}) and indices (R^2 = 0.52 and RMSE = 0.45 g C m^{-2}) individually, or in their combination (R^2 = 0.68 and RMSE = 0.56 g C m^{-2}). These findings highlight the importance of matching model selection and feature combinations to the specific characteristics of the data to enhance predictive performance.

The identification of the blue band, green band, MSAVI, NDRE, and Red-Edge2 as top predictors of GPP highlights the value of integrating both traditional spectral bands and narrowband vegetation indices. This finding reinforces the conclusions of studies such as Imran *et al*. (2020), which advocate for the use of narrow bands, red-edge, and broad-band vegetation indices for detailed monitoring of the biophysical and biochemical traits of plants. The prominence of EVI and MSAVI reflects their improved sensitivity in high-biomass areas compared to NDVI, a notable characteristic in various vegetation monitoring contexts (Imran *et al*., 2020). The strong importance of Red-Edge2 and NDRE further underscores the growing recognition of red-edge wavelengths in capturing physiological stress and chlorophyll dynamics, thereby aligning with the conclusions of studies by Saini (2023).

5. Broader Implications

The findings of this study have broad implications for ecological monitoring and modelling. They suggest that remote sensing-based productivity estimates can be substantially improved by incorporating vegetation indices sensitive to chlorophyll and canopy structures, particularly those in the red-edge and near-infrared range. This has important applications for precision agriculture, carbon accounting, and climate change monitoring, where accurate and timely GPP estimation is essential. Moreover, the study supports the growing consensus that ensembles of machine-learning methods, particularly SVM models, offer robust frameworks for integrating diverse data sources to predict complex ecological variables. As remote sensing data become increasingly fine-grained and multidimensional, such modelling approaches will be vital in transforming high-volume data into actionable ecological measures.

6. Limitations of the study

This study has several limitations that should be acknowledged. Firstly, the accuracy of the GPP estimates is dependent on the quality of the remote sensing inputs; residual atmospheric effects, cloud contamination. Furthermore, sensor noise may introduce uncertainty into the spectral bands and derived vegetation indices.

Secondly, the models incorporate a comprehensive list of predictors, primarily spectral, vegetation, and topographic variables. However, other important drivers of GPP, such as soil properties and detailed meteorological factors, could have improved the accuracy of the results. Furthermore, data constraints, in terms of availability, temporal and spatial coverage, and resolution, often limit modelling.

Thirdly, although cross-validation was applied to reduce overfitting, the performance of the RF and SVM models may still be influenced by parameter selection and the representativeness of the training data. Consequently, model transferability across different regions or time periods may be limited, although this has not been tested in this study. Finally, the GPP maps represent modelled estimates rather than direct measurements and should be interpreted as relative spatial patterns of productivity rather than as absolute values. The lack of a reliable distribution in the case of the Eddy Covariance Flux Towers in developing regions presents a major limitation, hindering reliable validation of the derived products. Nonetheless, the downscaling approach adopted here is promising for operational applications in data-scarce mountainous areas.

7. Recommendations

This study used downscaled MODIS-derived GPP as a response variable for modelling. While this approach yielded promising results, future research should evaluate the robustness

and accuracy of models based on both remote sensing-derived GPP and *in-situ* GPP measurements. The current model incorporated spectral bands, vegetation indices, and topographical variables as predictors. However, it is anticipated that incorporating additional variables such as climate and soil data, both of which are closely linked to carbon cycling and sequestration, will enhance the predictive performance of the model. Therefore, it is recommended that future studies consider the inclusion of these additional variables.

8. Conclusion

This study aimed to develop a high-resolution Sentinel-2 Gross Primary Productivity (GPP) product using various spectral, vegetation indices, and topographical variables, based on machine-learning algorithms. Specifically, various experimental scenarios were designed to reveal the influence of the various combinations on the GPP modelling performance. Moreover, the effectiveness of the Random Forest (RF) and Support Vector Machine (SVM) models in predicting GPP was evaluated, and important variables were identified. The results revealed that vegetation indices such as NDRE, MSAVI, and Cl-Red-edge exhibited strong positive correlations with GPP, underscoring their significance in capturing plant physiological activity. In contrast, topographical variables showed weak correlations with GPP, indicating a limited direct influence. The RF model outperformed the SVM when using combinations of topographic variables with spectral bands, topographic variables with indices, or when all predictors were included together. The best performance was achieved with an R^2 of 0.60 and an RMSE of 10.2 for the combination of topographic variables and indices. While the SVM demonstrated competitive accuracy with individual bands, indices, and combined bands and indices, its best results were obtained when both bands and indices ($R^2 = 0.68$, RMSE = 10.07) were used together. Key predictors identified through RF importance analysis included the blue and green bands, MSAVI, NDRE, and Red-Edge2, which further highlighted the value of red-edge and near-infrared data in GPP modelling. These findings align with existing literature on the importance of integrating spectral data and machine-learning techniques for monitoring ecosystem productivity. This research provides a practical framework to enhance remote sensing-based GPP estimation and offers valuable insights for ecological modelling and agricultural and climatic studies. The findings offer valuable insights into the capability of Sentinel-2 and topographic datasets to capture spatial variability in ecosystem productivity, while highlighting the efficacy of machine-learning approaches for scalable and efficient environmental monitoring.

9. Code and Data Availability:

The data that support the findings of this article can be requested from the corresponding author at Johnmapuru97@gmail.com.

10.Disclosures

The authors declare no conflicts of interest.

11.Acknowledgements

The financial assistance of the University of South Africa (UNISA) towards this research is hereby acknowledged. The opinions expressed, and the conclusions arrived at are those of the authors and are not necessarily to be attributed to UNISA. Mahlatse Kganyago received funding from the BRICS Multilateral Joint Call, administered by the National Research Foundation. This work is based on the research supported in part by the National Research Foundation of South Africa (Ref: BRIC231103160523).

12.Authors Contributions

Conceptualization: M.M.; methodology: M.M. and M.K; software: M.M; validation: S.X., K.M. and M.K; formal analysis: M.M; investigation: M.M., S.X and M.K.; resources: K.M. and M.M; data curation: M.M; writing and preparation of the original draft: M.M; writing and editing of the review: S.X., M.K. and K.M; visualization: M.M; funding acquisition: S.X. All authors have read and agreed to the published version of the manuscript.

13.References

Breiman, L., 2001. Random forests. *Machine Learning*, 45, pp.5-32.

Cao, D., Huang, X., Liu, G., Tian, L., Xin, Q. and Yang, Y., 2025. Evaluating the performance of satellite-derived vegetation indices in gross primary productivity (GPP) estimation at 30 m and 500 m spatial resolution. *Remote Sensing*, 17(19), p.3291.

Celis, J., Xiao, X., White Jr, P.M., Cabral, O.M. and Freitas, H.C., 2023. Improved modeling of gross primary production and transpiration of sugarcane plantations with time-series Landsat and Sentinel-2 images. *Remote Sensing*, 16(1), p.46.

Croft, H., Chen, J.M., Luo, X., Bartlett, P., Chen, B. and Staebler, R.M., 2020. Leaf chlorophyll content as a proxy for gross primary productivity in vegetation models. *Agricultural and Forest Meteorology*, 280, p. 107768.

Cushman, S.A., Kaszta, Z.M., Burns, P., Hakkenberg, C.R., Jantz, P., Macdonald, D.W., Brodie, J.F., Deith, M.C. and Goetz, S., 2024. Simulating multi-scale optimization and variable selection in species distribution modelling. *Ecological Informatics*, 83, p.102832.

Ding, C., Li, Y., Xie, Q., Li, H. and Zhang, B., 2023. Impacts of terrain on land surface phenology derived from harmonized Landsat 8 and Sentinel-2 in the Tianshan Mountains, China. *GIScience and Remote Sensing*, 60(1), p.2242621.

Florist, A., 2022. Reply on RC1. https://doi.org/10.5194/egusphere-2022-640-ac1

Gaye, B., Zhang, D. and Wulamu, A., 2021. A tweet sentiment classification approach using a hybrid-stacked ensemble technique. *Information*, 12(9), p.374.

Hari, M. and Tyagi, B., 2022. Terrestrial carbon cycle: tipping edge of climate change between the atmosphere and biosphere ecosystems. *Environmental Science: Atmospheres*, 2(5), pp.867-890.

Hua, X., Ohlemüller, R. and Sirguey, P., 2021. Topographical effects on the timing of the growing season in alpine grasslands.

Huang, E., Chen, Y., Fang, M., Zheng, Y. and Yu, S., 2021. Environmental drivers of plant distributions at global and regional scales. *Global Ecology and Biogeography*, 30(3), pp.697-709.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, 25(3), pp.295-309.

Imran, A.B., Khan, K., Ali, N., Ahmad, N., Ali, A. and Shah, K., 2020. Narrow band-based and broadband-derived vegetation indices using Sentinel-2 Imagery to estimate vegetation biomass. *Global Journal of Environmental Science and Management*, 6(1), pp.97-108.

Kay, C., Bredenkamp, G. J., and Theron, G. K., 1993. The plant communities of the Golden Gate Highlands National Park in the north-eastern Orange Free State. *South African Journal of Botany*, 59(4), 442-449.

Kayiranga, A., Ndayisaba, F., Nahayo, L., Karamage, F., Nsengiyumva, J.B., Mupenzi, C. and Nyesheja, E.M., 2017. Analysis of climate and topography impacts on the spatial distribution of vegetation in the Virunga Volcanoes Massif of East-Central Africa. *Geosciences*, 7(1), p.17.

Khatun, N., 2021. Applications of normality test in statistical analysis. *Open Journal of Statistics*, 11(01), p.113.

Li, Y., Gong, J. and Zhang, Y., 2023. Investigating the relationship between topographic factors and vegetation spatial patterns in the alpine plateau: a case study in the Southwestern Tibetan Plateau. *Remote Sensing*, 15(22), p.5356.

Liao, Z., Zhou, B., Zhu, J., Jia, H. and Fei, X., 2023. A critical review of methods, principles and progress for estimating the gross primary productivity of terrestrial ecosystems. *Frontiers in Environmental Science*, 11, p.1093095.

Lin, S., Hao, D., Zheng, Y., Zhang, H., Wang, C. and Yuan, W., 2022. Multi-site assessment of the potential of fine resolution red-edge vegetation indices for estimating gross primary production. *International Journal of Applied Earth Observation and Geoinformation*, 113, p.102978.

Lin, S., Li, J., Liu, Q., Li, L., Zhao, J. and Yu, W., 2019. Evaluating the effectiveness of using vegetation indices based on red-edge reflectance from Sentinel-2 to estimate gross primary productivity. *Remote Sensing*, 11(11), p.1303.

Liu, F., Wang, C. and Wang, X., 2021. Can vegetation index track the interannual variation in gross primary production of temperate deciduous forests?. *Ecological Processes*, 10, pp.1-13.

Liu, Y., Xu, W., Hong, Z., Wang, L., Ou, G., Lu, N. and Dai, Q., 2023. Integrating three-dimensional greenness into RSEI improved the scientificity of ecological environment quality assessment for forests. *Ecological Indicators*, 156, p.111092.

Malanson, G.P., Fagre, D.B., Butler, D.R. and Shen, Z., 2024. Alpine plant communities and current topographic microrefugia vary with regional climates. *Geomorphology*, 458, p.109241.

Maleki, M., Arriga, N., Barrios, J.M., Wieneke, S., Liu, Q., Peñuelas, J., Janssens, I.A. and Balzarolo, M., 2020. Estimation of gross primary productivity (GPP) phenology of a short-rotation plantation using remotely sensed indices derived from Sentinel-2 images. *Remote Sensing*, 12(13), p.2104.

Mallick, K., Verfaillie, J., Wang, T., Ortiz, A.A., Szutu, D., Yi, K., Kang, Y., Shortt, R., Hu, T., Sulis, M. and Szantoi, Z., 2024. 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 Sensing of Environment*, 307, p.114123.

Maluleke, A., Feig, G., Brümmer, C., Rybchak, O. and Midgley, G., 2024. Evaluation of selected Sentinel-2 remotely sensed vegetation indices and MODIS GPP in representing productivity in semi-arid South African ecosystems. *Journal of Geophysical Research: Biogeosciences*, 129(4), p.e2023JG007728.

Mapuru, M.J., Xulu, S., Gebreslasie, M. and Daemane, E.M., 2024. Exploring environmental factors that influence the distribution of poplar trees. *African Journal of Ecology*, 62(3), p.e13310.

Mapuru, M., Xulu, S., Gebreslasie, M. and Sadiki, M., 2025. Detecting and mapping invasive *Populus alba* species in mountainous ecosystems using Sentinel-2 imagery and random forest classification. *Journal of Sensors*, 2025(1), p.3138385.

Peng, Y., Gitelson, A.A. and Sakamoto, T., 2013. Remote estimation of gross primary productivity in crops using MODIS 250 m data. *Remote Sensing of Environment*, 128, pp.186-196.

Pabon-Moreno, D.E., Migliavacca, M., Reichstein, M. and Mahecha, M.D., 2022. On the potential of Sentinel-2 for estimating gross primary production. IEEE *Transactions on Geoscience and Remote Sensing*, 60, pp.1-12.

Pang, A., Chang, M.W. and Chen, Y., 2022. Evaluation of random forests (RF) for regional and local-scale wheat yield prediction in southeast Australia. *Sensors*, 22(3), p.717.

Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H. and Sorooshian, S., 1994. A modified soil-adjusted vegetation index. *Remote Sensing of Environment*, 48(2), pp.119-126.

Rodrigues, A.C., Villa, P.M., Ferreira-Júnior, W.G., Schaefer, C.E.R. and Neri, A.V., 2021. Effects of topographic variability and forest attributes on fine-scale soil fertility in late-secondary succession of Atlantic Forest. *Ecological Processes*, 10, pp.1-9.

Sabzchi-Dehkharghani, H., Biswas, A., Meshram, S.G. and Majnooni-Heris, A., 2024. Estimating gross and net primary productivities using earth observation products: a review. Environmental Modeling and Assessment, 29(1), pp.179-200.

Saini, R., 2023. Integrating vegetation indices and spectral features for vegetation mapping from multispectral satellite imagery using AdaBoost and random forest machine-learning classifiers. *Geomatics and Environmental Engineering*, 17(1), pp.57-74.

Salman, H.A., Kalakech, A. and Steiti, A., 2024. Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024, pp.69-79.

SANParks., 2020. Golden Gate Highlands National Park management plan, South Africa.

Sarkar, D.P., Shankar, B.U. and Parida, B.R., 2022. Machine-learning approach to predict terrestrial gross primary productivity using topographical and remote sensing data. *Ecological Informatics*, 70, p.101697.

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

Sett, T., Nikam, B.R., Singh, H. and Purohit, S., 2024. Hydrological variability in Indian forest ecosystem: Analysis of drought resilience, recovery and water-use efficiency in moist and dry deciduous forests. In *Urban Forests, Climate Change and Environmental Pollution: Physio-*

Biochemical and Molecular Perspectives to Enhance Urban Resilience (pp. 793-820). Cham: Springer Nature Switzerland.

Shankar, B.U., Sarkar, D.P. and Parida, B.R., 2024. Modeling GPP with machine learning using multisource features based on Fluxnet data. In 2024 IEEE *Ninth International Conference for Convergence in Technology* (I2CT) (pp. 1-5). IEEE.

Spinosa, A., Fuentes-Monjaraz, M.A. and El Serafy, G., 2023. Assessing the use of Sentinel-2 data for spatio-temporal upscaling of flux tower gross primary productivity measurements. *Remote Sensing*, 15(3), p.562.

Tatachar, A.V., 2021. Comparative assessment of regression models based on model evaluation metrics. *International Research Journal of Engineering and Technology* (IRJET), 8(09), pp.2395-0056.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G.A. and Torricelli, P., 2011. Application of a random forest algorithm to predict spatial distribution of the potential yield of *Ruditapes philippinarum* in the Venice lagoon, Italy. *Ecological Modelling*, 222(8), pp.1471-1478.

Wang, C., Wang, J., Naudiyal, N., Wu, N., Cui, X., Wei, Y. and Chen, Q., 2021. Multiple effects of topographic factors on spatio-temporal variations of vegetation patterns in the Three Parallel Rivers Region, Southeast Qinghai-Tibet Plateau. *Remote Sensing*, 14(1), p.151.

Wei, S., Yi, C., Fang, W. and Hendrey, G., 2017. A global study of GPP focusing on light-use efficiency in a random forest regression model. *Ecosphere*, 8(5), p.e01724.

Yang, F., Ichii, K., White, M.A., Hashimoto, H., Michaelis, A.R., Votava, P., Zhu, A.X., Huete, A., Running, S.W. and Nemani, R.R., 2007. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment*, 110(1), pp.109-122.

Yates, L.A., Aandahl, Z., Richards, S.A. and Brook, B.W., 2023. Cross validation for model selection: a review with examples from ecology. *Ecological Monographs*, 93(1), p.e1557.

Zeng, Y., Hao, D., Huete, A., Dechant, B., Berry, J., Chen, J.M., Joiner, J., Frankenberg, C., Bond-Lamberty, B., Ryu, Y. and Xiao, J., 2022. Optical vegetation indices for monitoring terrestrial ecosystems globally. *Nature Reviews: Earth and Environment*, 3(7), pp.477-493.

Zhang, H., Li, J., Liu, Q., Lin, S., Huete, A., Liu, L., Croft, H., Clevers, J.G., Zeng, Y., Wang, X. and Gu, C., 2022. A novel red-edge spectral index for retrieving the leaf chlorophyll content. *Methods in Ecology and Evolution*, 13(12), pp.2771-2787.

Zhang, K., Liu, N., Chen, Y. and Gao, S., 2019, Comparison of different machine learning methods for GPP estimation using remote sensing data. In IOP Conference Series: *Materials Science and Engineering*, 490 (6), 062010).

Zhang, Q., Cheng, Y.-B., Lyapustin, A., Wang, Y., Zhang, X., Suyker, A. E., Verma, S. B., Shuai, Y. and Middleton, E. M. 2015. Estimation of crop gross primary production (GPP). II: Do scaled MODIS vegetation indices improve performance? *Agricultural and Forest Meteorology*, 200, 1-8