Development of the Grass LAI and CCC remote sensing-based models and their transferability using Sentinel-2 data in heterogeneous grasslands

Philemon Tsele [1,*], Abel Ramoelo [2], Mcebisi Qabaqaba [2]

[1] *Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria 0028, South Africa;*

[2] *Centre for Environmental Studies, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria 0028, South Africa;*

*Correspondence: philemon.tsele@up.ac.za; Tel.: +27-12-420-4939

Estimation of biophysical variables such as leaf area index (LAI) and canopy chlorophyll content (CCC) at high spatiotemporal resolution is important for managing natural and heterogeneous environments. However, accurate estimation of biophysical variables particularly over heterogeneous environments remains a challenge. The objective of the study was to develop locally parameterised grass LAI and CCC empirical models using the Sentinel-2 variables combined with the Stepwise multiple linear regression (SMLR) and Random forest (RF) at the Golden Gate Highlands National Park (GGHNP) and Marakele National Park (MNP) in South Africa. Results showed that in MNP, SMLR yielded better LAI estimation with root mean squared error (RMSE) of 0.67 $m^2.m^{-2}$ and mean adjusted error (MAE) of 0.54, explaining 48% of LAI variability, when bands and indices are combined. In contrast, RF gave better CCC estimation i.e. RMSE and MAE of 17.08 $\mu g.cm^{-2}$ and 13.18 respectively, explaining about 40% of CCC variability with Sentinel-2 bands only. In GGHNP, the RF models provided the best estimates of both

LAI and CCC compared to SMLR models. Furthermore, the CCC and LAI estimation models of GGHNP showed improved model accuracies when 50% and 75% of the MNP field samples were transferred to the GGHNP models. In contrast, the CCC and LAI estimation models of MNP showed a decline in model performance across all scenarios where the GGHNP field samples were transferred to the MNP models. These findings have significant implications for the development of locally parameterised types of models that can provide improved and consistent site-specific accurate estimates of grass biophysical parameters over heterogeneous environments.

Keywords: Leaf area index (LAI), Canopy chlorophyll content (CCC), Sentinel-2 imagery, indices

Introduction

Estimation of vegetation biophysical variables is important for understanding vegetation condition, structure, growth status and gross primary productivity. In light of biodiversity loss and ecosystem restoration, these variables can be used to assess and monitor vegetation state and biodiversity at large. The leaf area index (LAI) defined as the one-sided leaf area per unit of horizontal surface area (Jonckheere et al. 2004) is an important indicator of plant canopy structure and growth, and also forms an essential input in climate models to determine ecosystem productivity. Another biophysical variable called the leaf chlorophyll content (LCC) carries valuable information about vegetation physiology and could be regarded as a key indicator of plant health status. Accurate measurements of LCC can be helpful for precision management of natural resources and agricultural fields (Bei et al. 2019). Furthermore, the canopy chlorophyll content (CCC), which refers to the overall amount of chlorophyll *a* and *b* pigments in a compact group of plants per unit ground area (Gitelson et al. 2005) is derived

from the product of the LCC, $\mu g.cm^{-2}$ and the corresponding LAI, $m^2.m^{-2}$ in a subplot (Darvishzadeh et al. 2008). CCC is an important indicator of vegetation health condition, plant species diversity and forage quality assessment (Ali et al. 2020). These variables i.e. LAI, and CCC form part of the essential biodiversity variables (BON 2015) and are also listed and ranked as some of the top 30 biodiversity metrics measured from space, using satellite remote sensing (Skidmore et al. 2021).

Heterogeneous ecosystems like the grasslands and savannah of South Africa, are characterised by native grasses of different mixture of grass and tree species often distributed across varying terrain slopes, soils and geology types (Masemola, Cho, and Ramoelo 2016; Ramoelo et al. 2015). Such an environment, analogous to rangelands, is favourable for livestock production and grazing by animals (Svinurai et al. 2021). Therefore, it is critical to (i) assess areas where there is a change in response to climate and/or anthropogenic effects, (ii) monitor the functional status and diversity of the rangeland vegetation communities in-order to enhance ecosystem productivity and stability, guided by effective resource management strategies and policies. These aforementioned processes are measurable through biophysical variables such as LAI, LCC and CCC which can be estimated in the field or through modelling procedures applied to remotely sensed imagery (Chuvieco 2016). Recently, there has been a successful attempt to generate vegetation biophysical products that provide modelled estimates of e.g. LAI, CCC and fractional vegetation cover (FVC) at high spatiotemporal resolutions of Sentinel-2 data (Weiss, Baret, and Jay 2020).

The retrieval accuracy of these biophysical variables (especially LAI and CCC) in heterogeneous environments characterised by diversity of land cover, species diversity and varying terrain slopes remains a notable concern and an area for further investigation (Darvishzadeh et al. 2008; Cho, Ramoelo, and Math 2014; Brown et al. 2021). For example, in the heterogeneous grasslands of South Africa, inadequate retrieval accuracies of grass LAI and

CCC from the Sentinel-2 Level 2 Prototype Processor (SL2P) (Weiss, Baret, and Jay 2020) were reported in a recent validation study (Tsele et al. 2022). Brown, Ogutu, and Dash (2019) assessed the accuracy of the Sentinel-2 derived LAI and CCC biophysical variables over a deciduous broadleaf forest site in Southern England. The study reported moderate inaccuracies, and suggested using alternative model inversion methods such as the invertible forest reflectance model (INFORM) that are optimised for forest environments. A virtually similar study by Filipponi (2021) modelled LAI estimates from both Sentinel-2 and Landsat-8 imagery over croplands, grasslands, broadleaved and needleleaf forests in Italy. The study found a general underestimation of LAI over the aforementioned test site classes, however higher overestimations were noted in the grasslands. Furthermore, Ali et al. (2020) compared statistical and physically-based methods in estimating CCC using Sentinel-2 data and various vegetation indices (VIs) over a heterogeneous mixed mountain forest. It was found that, although both methods had comparable prediction accuracies of CCC, the statistical methods gave the lowest prediction error coupled with the highest coefficient of determination (R^2). Overall, these studies show that the performance of the models used to estimate the biophysical parameters may lack generality in heterogenous canopies at regional to global level. This could be an indication for the need to further explore and develop locally parameterised types of models, looking at empirical and/or inversion of the physically-based models. However, in agricultural environments which are largely characterised by homogeneous cover, the models have demonstrated satisfactory performance (Kganyago et al. 2020; Hu et al. 2020; Kganyago, Mhangara, and Adjorlolo 2021) and the potential to be transferred to other sites (Kganyago, Adjorlolo, and Mhangara 2022).

While numerous studies have evaluated the performance of Sentinel-2 data in estimating vegetation biophysical variables (Delegido et al. 2011; Clevers and Gitelson 2013; Ramoelo and Cho 2018; Sun et al. 2019; Guerini Filho, Kuplich, and Quadros 2020; Andreatta et al.

2022) there is limited effort towards examining the performance of Sentinel-2 bands in conjunction with red-edge based indices to estimate both LAI and CCC variables over heterogenous ecosystems. Few studies reported varying degrees of accuracy when evaluating the performance of Sentinel-2 data for predicting LAI and CCC over a heterogenous grassland environment using statistical approaches, for example Schwieder et al. (2020) and Sakowska, Juszczak, and Gianelle (2016). Furthermore, to our knowledge based on available literature, the stepwise multiple linear regression (SMLR) and random forest (RF) which are known to be parametric linear and non-parametric non-linear methods respectively, have not been widely tested in the context of heterogenous grass biophysical-parameters estimation and monitoring from Sentinel-2 imagery. It is worth exploring whether these methods could be potential alternatives to current operational approaches, especially in providing improved and consistent site-specific accuracy of grass biophysical parameter estimation.

The notion of transferrable models across geographic sites for estimation of vegetation biophysical parameters is important, given the excessive costs associated with field data collection in order to obtain extensive training sets. This could affect the potential to estimate these parameters at regional to global level with acceptable accuracies. In particular, radiative transfer models (RTMs) have minimum reliance on *in-situ* data and are known to be robust and transferrable because they use the physical laws (Goel 1987) to accurately describe the spectral variation of canopy reflectance as a function of viewing and illumination geometry, canopy, including leaf and soil background characteristics (Darvishzadeh et al. 2011). However, it has been reported that RTMs still require local parameterization in order to simulate multispecies canopies accurately, especially in heterogenous environments (Atzberger et al. 2015; Darvishzadeh et al. 2008; Combal et al. 2003; Bsaibes et al. 2009). In contrast, statistical or empirical models often lack transferability to other sites (Verrelst et al. 2015) and also, their robustness and accuracy of the modelled relationships depends on the properties of the acquired

field data (Atzberger et al. 2015). Therefore, this study explored locally parameterised types of SMLR and RF based empirical models on whether the models can be improved through transfer scenarios involving different proportions of field samples from different sites. For example, this notion was successfully tested by (Kganyago, Adjorlolo, and Mhangara 2022) in an agricultural environment whereby, the transferability of empirical models and training samples between two study sites in South Africa that share similar site characteristics (i.e. crop types) and imagery acquisition conditions was investigated, for improving the accuracy retrievals of crop LCC, CCC and LAI biophysical parameters. Based on our observations of the available field-sample measurements of grass CCC and LAI between the target sites in this study, an assumption was made that the range of values of these biophysical parameters in the target sites is not far apart, and therefore the notion of transferability can be tested. The aim of this study was to evaluate the Sentinel-2 spectral reflectance bands and various VIs for the estimation of grass biophysical parameters during peak productivity over a heterogeneous grassland environment in South Africa. The study objectives were to:

(i) Evaluate and compare the LAI and CCC estimation capability using SMLR and RF as well as Sentinel-2 data in heterogeneous environments,

(ii) Identify which Sentinel-2 derived variables (bands and/or VIs) are important to estimate grass LAI and CCC in heterogeneous environments,

(iii) Evaluate the effect of transferring varying proportions of field samples on improving model accuracy from one site to the other.

Material and methods

Study area

Two heterogenous study sites were selected in two South African National Parks namely, the Golden Gate Highlands National Park (hereafter, GGHNP) located in the Free State province

near the Lesotho border, and Marakele National Park (hereafter, MNP) located in the Waterberg mountains of the Limpopo province, as shown in Figure 1. The study sites were selected based on certain landscape attributes, such as biomes and vegetation communities (Mucina and Rutherford 2006). For example, about 36 grass species were identified within the visited plots in GGHNP (Figure 1) and the most dominant were *Eragrostis curvula*, *Elionurus muticus*, *Aristida adscensionis*, *Stiburus alopecuroide*, *Sporobolus africanus*, *Heteropogon contortus*, *Tristachya leucothrix*, *Microchloa caffra*, *Themeda triandra*, *Urochloa decumbens*, *Helichrysum rugulosum* and *Helichrysum pilosellum*. Similarly, more than 30 grass species were identified across the visited plots in MNP such as (to name a few), *Hyperthelia dissoluta, Eragrostis lehmanniana, Themeda triandra, Digitaria eriantha, Miscanthus junceus, Digitaria Brazzae, Aristida diffusa, Eragrostis racemosa, Schizachyrium jeffisi and Panicum natalense*.

Furthermore, the GGHNP and MNP sites are mountainous and characterised by surface height variation i.e. elevations that range between approximately 1639 m to 2815 m and 976 m to 2091 m respectively, estimated from the 30 m resolution Shuttle Radar Topography Mission (SRTM) data acquired from the United States Geological Survey (USGS) Earth Explorer (https://earthexplorer.usgs.gov/). Both sites fall within the summer rainfall region of South Africa. In particular, the GGHNP receives average rainfall of approximately 700 mm per year (Mucina and Rutherford 2006) whereas, the MNP site can receive average rainfall of up to around 630 mm annually (Van Staden and Bredenkamp 2005).

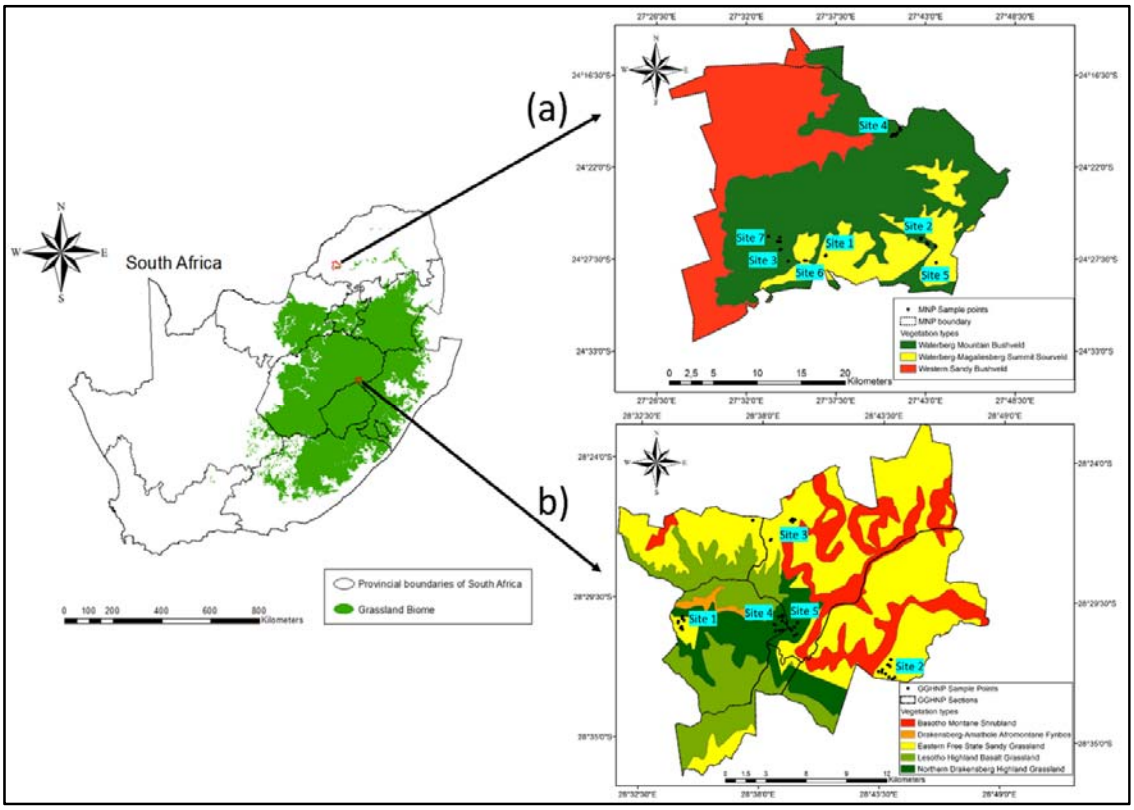

Figure 1: The location of the two selected study sites in South Africa (a) and (b) where LAI, CCC and FVC sample field measurements were taken. Site (a) represents the Marakele National Park (MNP) whereas site (b) represents the Golden Gate Highlands National Park (GGHNP) on 8 – 10 April 2021 and 18 – 21 March 2021 respectively.

The dominant underlying geology in the GGHNP includes mudstone, fine-to-medium sandstone and basalt, based on the national geology map developed by the Council for Geosciences (CG 1997). In the same site, the soils include shallow to deep sandy soil that is extremely gravelly as well as a clay-rich subsoil (https://data.isric.org/; Van Engelen and Dijkshoorn 2013). On the other hand, in MNP the geology is largely characteristic of sandstone and mudstone, followed by granule stone, siltstone and diabase (CG 1997). The soils in MNP range from shallow-gravel soil to low activity clayed soil (https://data.isric.org/; Van Engelen and Dijkshoorn 2013).

Schematic workflow

Figure 2 show a schematic workflow summarizing the various phases of the methodology that were implemented in this study. These phases are discussed in subsequent sections of the paper.

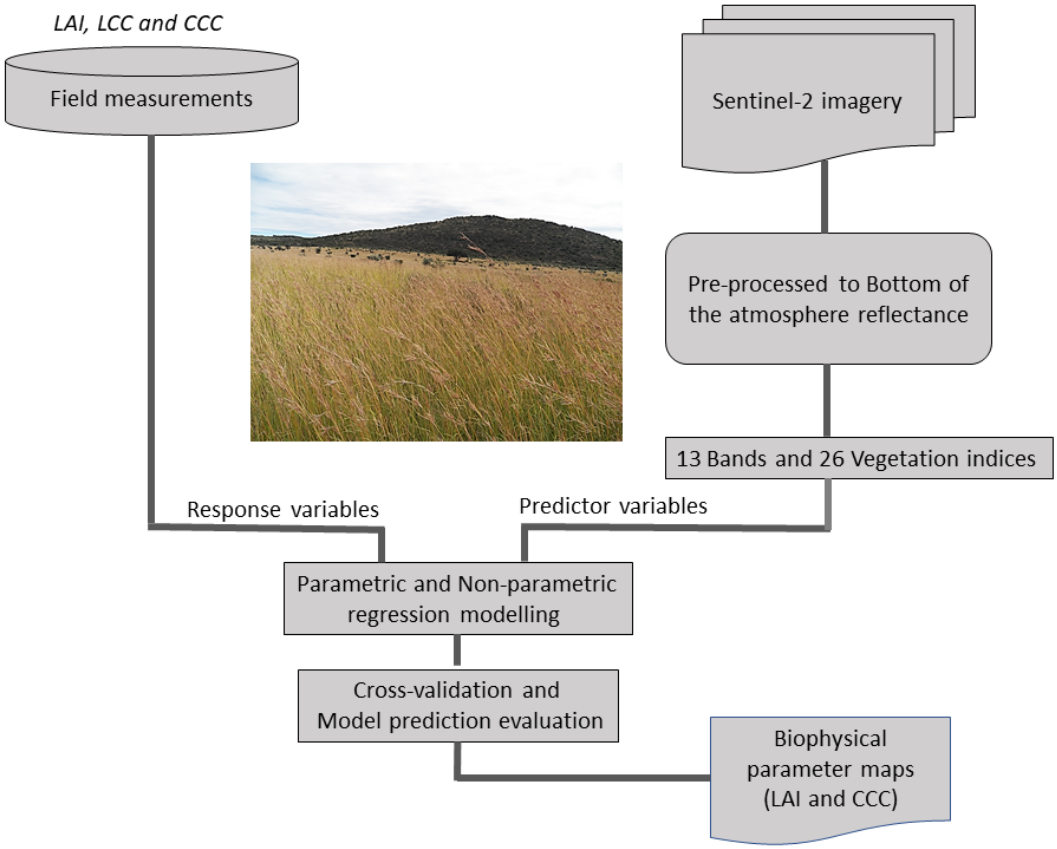

Figure 2: Methodological flowchart developed in this study for estimation of grass LAI and CCC in heterogenous natural environment in South Africa.

Remotely-sensed imagery

Sentinel-2 images were acquired free of charge from the European Space Agency data hub (https://scihub.copernicus.eu/dhus/#/home) on the 27th of March 2021 and the 9th of April 2021. The selection of the images was such that they (i) are free from any cloud obscuration (ii) covered the two study sites and (iii) had acquisition dates that were very close (i.e. <= 6 days) to the field data collection dates. Sentinel-2 multispectral imager (MSI) data has 13

spectral bands, characterised by fine spatial resolutions in the range 10-60 m, that cover large geographic areas (i.e. 120km × 120km per scene) at high a temporal resolution of up-to 5 days (Table 1). The Sentinel-2 images were pre-processed to surface reflectance or Bottom of the Atmosphere (BOA) reflectance i.e. Level-2A using the SNAP Sentinel-2 atmospheric correction tool, Sen2Cor, version 2.8 (Louis et al. 2016).

Table 1: Resolution characteristics of the Sentinel-2 MSI data. The spectral bands (and VIs discussed later) were used as input predictor variables in this study for regression modelling.

Band number	Band description	Central wavelength (nm)	Bandwidth	Spatial resolution (m)	Temporal resolution
B1	Coastal aerosol	443	20	60	5 days
B2	Blue	490	65	10	
B3	Green	560	35	10	
B4	Red	665	30	10	
B5	Red edge$_1$	705	15	20	
B6	Red edge$_2$	740	15	20	
B7	Red edge$_3$	783	20	20	
B8	Near infrared	842	115	10	
B8A	Narrow near infrared	865	20	20	
B9	Water vapour	945	20	60	
B10	Cirrus	1375	30	60	
B11	Shortwave infrared$_1$	1610	90	20	
B12	Shortwave infrared$_2$	2190	180	20	

Furthermore, the Sentinel-2 BOA images were resampled to the spatial resolution of 20 m using the resampling function within the SNAP software version 8.0 (https://step.esa.int/main/download/snap-download/). This spatial resolution is such that the 20 m x 20 m pixels in sampled geographic areas correspond to single field plots of size of 20 m x

20 m that contains two subplots, each of size 1 m x 1 m. In particular, the total number of sampled locations were 80 and 68 in GGHNP and MNP, respectively. The sampling strategy involved a combination of stratified and purposive sampling methods (Tsele et al. 2022). Random samples were initially taken across different grass vegetation communities and varying slope terrains spanning the crests, valleys and low to mid-slopes. However, when in the field, there were certain inaccessible areas, which led to the use of purposive sampling where re-placement of the sampled locations was done, close to the randomized points. Lastly, the geographical coordinates of the field sublots were used to extract corresponding Sentinel-2 20 m resolution pixels of spectral reflectance and VIs for modelling LAI and CCC in both the GGHNP and MNP sites.

Vegetation indices used

VIs are simple band mathematical expressions that capitalize on varying spectral information between bands, in order to enhance the radiometric signal of the target feature while suppressing that of other features (Chuvieco 2016). In remote sensing of vegetation, many studies have successfully demonstrated that VIs can be used as an optical measure of greenness, and also as a proxy measure of vegetation biochemical and biophysical variables such as leaf nitrogen content (LNC), LAI, CCC, green biomass and FVC, for example Ramoelo et al. (2012); Ramoelo and Cho (2018); Masemola, Cho, and Ramoelo (2016); Ali et al. (2020); Guerini Filho, Kuplich, and Quadros (2020), Andreatta et al. (2022). Furthermore, such studies found VIs beneficial in developing statistical and/or physically-based models for retrieval of biophysical vegetation attributes. In this study, 26 VIs were computed based on Sentinel-2 bands (Table 2) and evaluated as prediction variables for estimation of LAI and CCC in a heterogeneous grassland environment. Majority of the VIs used in this study, included at least one Sentinel-2 red-edge band as shown in Table 2 because, their inclusion have shown to have

the potential for improved estimation of vegetation biophysical variables such as the LAI, CCC and LNC, for example Delegido et al. (2011); Clevers and Gitelson (2013); Sun et al. (2019); Ali et al. (2020).

Table 2: Vegetation indices (VIs) and bands that were used as input predictor variables in this study during modelling procedures. All VIs were computed using the Sentinel-2 spectral bands. The bands (B_i) highlighted in **bold** represent the red-edge bands of Sentinel-2 data.

Index	Name	Formula based on Sentinel-2 bands	Citation
SR1	Simple ratio 1	B8/B4	Jordan (1969)
SR2	Simple ratio 2	B2/**B6**	Henrich et al. (2012)
SR3	Simple ratio 3	B4/**B5**	
SR4	Simple ratio 4	B2/**B5**	
SR5	Simple ratio 5	**B5**/B4	
SR6	Simple ratio 6	**B6**/B5	
SR7	Simple ratio 7	**B7**/B4	
SR8	Simple ratio 8	B8/**B5**	
SR9	Simple ratio 9	B8A/**B5**	
SR10	Simple ratio 10	**B5**/B3	
SR11	Simple ratio 11	**B5**/B2	
SR12	Simple ratio 12	**B5**/B9	
NDVI1	Normalized difference VI1	(B8-B4)/(B8+B4)	Rouse Jr et al. (1973)
NDVI2	Normalized difference VI2	(**B7**-B4)/(**B7**+B4)	Henrich et al. (2012)
NDVI3	Normalized difference VI3	(**B7**-B3)/(**B7**+B3)	
NDVI4	Normalized difference VI4	(B3-**B5)**/(B3+**B5**)	
NDVI5	Normalized difference VI5	(B9-**B5**)/(B9+**B5**)	
NDVI6	Normalized difference VI6	(**B5**-B3)/(**B5**+B3)	

RE	Red edge	(B4+**B7**)/2	Horler, DOCKRAY, and Barber (1983), Horler et al. (1983)
REP	Red edge position	700 + 40*[(RE-**B5**)/(**B6**-**B5**)]	Dawson and Curran (1998)
CIRE	Chlorophyll index red edge	(B9/**B5**) -1	Gitelson, Keydan, and Merzlyak (2006)
CRE	Chlorophyll red edge	(**B7**/**B5**)$^{(-1)}$	Gitelson, Gritz, and Merzlyak (2003)
LCI	Leaf chlorophyll index	(B8-**B5**)/(B8+B4)	Datt (1999)
MCARI	Modified Chlorophyll Absorption in Reflectance Index	((**B5**-B4)-0.2(**B5**-B3))×(**B5**/B4)	Daughtry et al. (2000)
PNDVI	Pan NDVI	(B9-(B3+**B5**+B2))/(B9+(B3+**B5**+B2))	Wang et al. (2007)
RBNDVI	Red-Blue NDVI	(B9-(**B5**+B2))/(B9+(**B5**+B2))	Wang et al. (2007)

Stepwise multiple linear regression

The stepwise multiple linear regression (SMLR) fits field observed biophysical variable (e.g. LAI or CCC) using a linear combination of predictor variables (e.g. spectral reflectance bands and/or VIs). The fitting can be described using the following, generic first-order multiple linear regression equation (Eberly 2007):

$$B = \beta_0 + \beta_1 a + \beta_2 b.... + \beta_x c + \varepsilon \quad (1)$$

where B is the LAI or CCC; a, b and c are the predictor variables, β_0, β_1, β_2 *and* β_x are the unknown coefficients and ε is the random error. The required assumptions prior to using the equation were observed. Equation 1 was recursively applied using both the forward and backward selection procedures available in the 'olsrr' package version 0.5.3 in R-studio (Hebbali and Hebbali 2017) to find optimal models based on a variable set of important predictors (i.e. Sentinel-2 bands and/or VIs) for estimating grass LAI and CCC in the MNP and GGHNP sites. In particular, the Akaike information criterion (AIC) within the 'olsrr' package provided the means for optimal model selection by estimating the quality of each model (Hebbali and Hebbali 2017).

Random Forest

Random forest (RF) is an ensemble machine-learning algorithm (Breiman 2001) that builds an assortment of multiple decision trees. RF is an extension of the Classification and Regression Trees (CART) algorithm (Breiman et al. 2017) and has been found in other studies to be potentially more accurate and relatively robust to outliers, when compared to other non-linear non-parametric methods such as the individual decision trees and neural networks (Mutanga, Adam, and Cho 2012; Chen et al. 2014; Rodriguez-Galiano et al. 2012; Liang et al. 2016). For every tree that is grown in a RF, a new training set of size m is randomly selected with replacement from the original training set of size M (where $m < M$). The proportion of samples that is not selected in the original training set, is left out-of-bag (OOB) and used to estimate the model performance and variable importance. Furthermore, for each node of the tree, there are X input variables (e.g. spectral bands) from which only x number of variables of out the X are randomly selected for determining the optimal split at that node for growing a forest of trees. The unclassified pixel is run through each of the generated trees, and each tree would then

classify this pixel into one of the *Y* classes (as defined in the training data set). Finally, the pixel would be assigned to the class that had the most classifications i.e. majority vote.

In this study, the *variable selection using random forests* (VSURF) package available in R-statistics software (Genuer, Poggi, and Tuleau-Malot 2015) was used for estimating LAI and CCC with Sentinel-2 data in MNP and GGHNP. A variable selection process was executed for three modelling scenarios encompassing the Sentinel-2 bands, Sentinel-2 derived VIs, and the combination of bands and VIs in-order to identify the important variables in estimating grass LAI and CCC in heterogeneous environments. Furthermore, the implementation of the RF algorithm using the randomForest package in R statistical software version 4.1.3, was fine-tuned using the caret package (Kuhn 2008) in R. The caret package provides various functions that expedite the development of predictive models by offering tools for tasks such as data splitting, pre-processing, feature selection, model tuning, and variable importance selection. To determine the optimal model, the root mean square error (RMSE), R^2, and mean absolute error (MAE) were calculated during parameter tuning, with the minimum root mean squared error (RMSE) value being used as the selection criterion.

Cross-validation for SMLR and RF regression, and evaluation of model prediction accuracies

A rigorous measure of model error requires a set of data points or observations that were not utilised in model calibration. In instances where few sample data points exist, splitting the points into validation and calibration datasets may lead to having few cases in each dataset. In contrast to the split-sample approach, a cross-validation procedure can be implemented (Snee 1977). Cross-validation is a method that is based on dividing the observations into different or equally-sized sub-datasets, and each time the method does calibration using an empirical function, it leaves out one or more observations at a time for testing purposes (Chuvieco 2016). In this study, given that there were relatively few ground observations in both study sites i.e. MNP (68 samples) and GGHNP (80 samples), the observations in the sample dataset for each site were not split into training and validation datasets. Furthermore, the cross-validation resampling method (Snee 1977) was used to validate the SMLR and RF fitted models. In this

study, the observations in the dataset for each site were randomly divided into $k = 10$ equal-sized sub-datasets. We defined 5 iterative validation steps and, in each step, the k sub-datasets were used only once as a validation dataset for model testing. The results from each of these iterative steps, were assessed using statistical performance metrics such as the R^2, RMSE, Relative root mean squared error (RRMSE) and MAE.

The prediction accuracies of the SMLR and RF models were evaluated with the R^2, RMSE, RRMSE and MAE. These represent some of the standard performance metrics (Chai and Draxler (2014)) that are widely used in numerous studies involving the estimation of vegetation biophysical and/or biochemical parameters, for example Ali et al. (2021); Kganyago, Mhangara, and Adjorlolo (2021); Verrelst et al. (2015); Ramoelo and Cho (2018); Guerini Filho, Kuplich, and Quadros (2020); Richter et al. (2012); Darvishzadeh et al. (2008). The R^2 shown in Equation 1 was computed for each model to measure the goodness of fit. This was followed by the computation of RMSE shown in Equation 2 which indicate the amount of error expressed in the units of the biophysical variable of interest i.e. $m^2.m^{-2}$ for LAI and $\mu g.cm^{-2}$ for CCC. RMSE can range from 0 to ∞ and a lower value (closer to 0), indicate an accurate model (Chai and Draxler 2014). Additionally, the RRMSE shown in Equation 3 was used to facilitate comparison of model accuracies between different variables, where model accuracy was regarded as either excellent (RRMSE<10%), good (10%<RRMSE<20%), fair (20%<RRMSE<30%) or inadequate (RRMSE>30%) (Jamieson, Porter, and Wilson 1991; Heinemann et al. 2012). Furthermore, MAE shown in Equation 4 was also used a supplementary metric to RMSE to evaluate model error. The combination of MAE and RMSE metrics gave a representation of the variation in model error distribution, which can be normally- or uniformly distributed (Chai and Draxler 2014).

$$R^2 = 1 - \frac{\sum (e_k^N - \overline{e}_k)^2}{\sum (e_k - \overline{e}_k)^2} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N} (e_k - m_k)^2}{n}} \quad (3)$$

$$RRMSE = \frac{RMSE}{\overline{m}_k} \times 100 \quad (4)$$

$$MAE = \frac{1}{n} \sum_{k=1}^{n} | e_k - m_k | \quad (5)$$

where m_k is the observed biophysical variable i.e. LAI or CCC, and e_k is the model predicted biophysical variable i.e. LAI or CCC, $\overline{m}_k$, and $\overline{e}_k$ denotes the respective means of observed and model predicted biophysical variables, *n* is the sample size, and *N* is the number of errors.

Results and Discussion

Field measurements of biophysical variables

Table 3 show the summary statistics of the grass biophysical variables and terrain attributes over the two study sites. Generally, the field measurements across the subplots, resembled an approximately normal distribution, which was inferred from the proximity of the respective mean and median values per variable. The difference between these two basic statistical measures i.e. measures of central tendency, was considered in this study as a natural test for data distribution symmetry (Gastwirth 1971).

Table 3: Summary statistics of selected terrain attributes and measured biophysical variables of grassland sample subplots. The statistical parameters, CV denotes the coefficient of variation, and StDev the standard deviation.

Site	Measured variables	No. of Subplots	Min.	Max.	Mean	Median	StDev	CV
GGHNP	LAI ($m^2.m^{-2}$)	80	0.61	6.24	2.26	2.02	1.24	0.55
	CCC ($\mu g.cm^{-2}$)	80	7.24	162.61	46.01	37.05	32.26	0.70
	FVC	80	0.47	1.00	0.86	0.87	0.11	0.13
	Grass height (cm)	80	5.00	34.00	11.92	11.00	5.48	0.46
	Elevation (m)	80	1832.20	2102.41	1966.04	1960.56	78.36	0.04
	Slope (°)	80	0.49	12.04	3.67	3.26	2.59	0.70
	Aspect (°)	80	0.00	357.51	174.36	135.00	130.72	0.75
MNP	LAI ($m^2.m^{-2}$)	68	0.47	5.00	1.90	1.90	0.84	0.44
	CCC ($\mu g.cm^{-2}$)	68	9.29	132.59	42.37	42.72	20.98	0.50
	FVC	68	0.25	0.97	0.62	0.60	0.18	0.28
	Grass height (cm)	68	4.50	37.00	16.38	15.75	8.30	0.51
	Elevation (m)	68	1307.59	1893.29	1470.54	1389.27	167.68	0.11
	Slope (°)	68	0.34	14.12	3.61	3.16	2.65	0.73
	Aspect (°)	68	0.00	354.81	131.63	68.11	125.11	0.95

Table 3 data for GGHNP (which is mainly a grassland environment) suggests the sampled subplots were characterised by high vegetation cover with little variability according to the respective mean, standard deviation and CV of the FVC. This is also corroborated by the high LAI values reaching a maximum of 6.24. The CCC suggests the grasses in the sampled areas were on average green and healthy. However, the grass height, LAI and CCC showed moderately high variability (i.e. according to CV that ranged from approximately 46% to 70%)

compared to the FVC biophysical variable. This variability is representative of the various vegetation communities across the sections of the park, and in particular the heterogenous grassland environment. In addition, the CCC's variability could also be controlled by the different soils and climate within GGHNP, for example Li et al. (2018).

On the other hand, the grasses in MNP (which is generally a mixed savanna and grassland environment) had on average, moderately-high vegetation cover marked by FVC of 62%. The CCC was on average, moderate and showed little variability (StDev and CV of about 21% and 50% respectively) across the subplots in MNP compared to in GGHNP (Table 3). However, the LAI and grass height appeared to have a relatively high variability (i.e. CV of about 44% and 51% respectively) across the subplots indicating widespread vegetation structural differences and heterogeneity of the mountainous savanna and grassland environment.

Furthermore, Table 3 include statistical information on the altitude, slope and aspect particularly where our subplots were located. This provides important information on terrain variability where our subplots were located. It is evident that our study sites are mountainous, characterised by high altitude ranges (Table 3). Considerable care during sampling design was taken to ensure that our subplots were located in fairly homogeneous surroundings characterised by varying slope terrains as this can be seen on the slope values (Table 3). On average the two sites, at least where our subplots were located had south to southeast facing slopes i.e. aspect (Table 3).

LAI model performance

Table 4 and Table 5 show the LAI modelling results from the SMLR and RF methods tested on different modelling scenarios involving the Sentinel-2 MSI bands, VIs and the combination of bands and VIs. For each modelling scenario, the selection of the best performing LAI model

was based on firstly, the lowest RMSE and MAE followed by RRMSE. The LAI modelling results in MNP (Table 4) show that SMLR had a better estimation capability i.e. RMSE of 0.67 $m^2.m^{-2}$ explaining about 48% of LAI variability, when the bands and VIs are combined.

Table 4: LAI predictive modelling results in MNP using stepwise multiple linear regression (SMLR) and Random Forest (RF).

Model Scenarios	**R^2**	**RMSE ($m^2.m^{-2}$)**	**RRMSE (%)**	**MAE**
SMLR				
LAI and Bands only	0.33	0.72	35.68	0.57
LAI and Indices only	0.48	0.67	31.85	0.54
LAI, Bands and Indices	**0.48**	**0.67**	**27.31**	**0.54**
RF				
LAI and Bands only	0.38	0.68	24.36	0.53
LAI and Indices only	0.33	0.71	31.44	0.58
LAI, Bands and Indices	0.34	0.74	31.34	0.59

This performance is marginally followed by the RF model using only the bands, which yielded RMSE of 0.68 $m^2.m^{-2}$ explaining approximately 38% of LAI variability in MNP. In particular, notwithstanding the low MAE values, both these models demonstrated a fair or reasonable LAI estimation capability in MNP with RRMSE values of 27.31% (SMLR) and 24.26% (RF) respectively. However, in GGHNP, the RF model based on a combination of bands and VIs show a relatively better LAI estimation capability with RMSE of 0.93 $m^2.m^{-2}$ when compared to SMLR (Table 5).

Table 5: LAI predictive modelling results in GGHNP using stepwise multiple linear regression (SMLR) and Random Forest (RF).

Model Scenarios	**R^2**	**RMSE ($m^2.m^{-2}$)**	**RRMSE (%)**	**MAE**
SMLR				
LAI and Bands only	0.33	1.05	43.12	0.90
LAI and Indices only	0.33	1.07	40.92	0.85
LAI, Bands and Indices	0.39	0.95	35.95	0.77
RF				
LAI and Bands only	0.32	1.04	32.28	0.85
LAI and Indices only	0.44	0.95	20.81	0.75
LAI, Bands and Indices	**0.43**	**0.93**	**22.24**	**0.74**

In addition, the RF model explained approximately 43% of LAI variability and showed a fair predictive performance according to lower RRMSE and MAE values (Table 5). In overall, these results (Table 4 and Table 5) highlighted the importance of using the combination of bands and VIs as model predictor variables to achieve better LAI estimation capability in MNP and GGHNP. In contrast, using the bands and VIs separately for LAI estimation in MNP and GGHNP revealed inadequate model accuracies with RRMSE's generally greater than 30%.

Table 6: CCC predictive modelling results in MNP using stepwise multiple linear regression (SMLR) and Random Forest (RF).

Model Scenario	R^2	RMSE (µg.cm^{-2})	RRMSE (%)	MAE
SMLR				
CCC and Bands only	0.27	18.41	42.15	15.13
CCC and Indices only	0.35	17.82	38.69	14.13
CCC, Bands and Indices	0.33	18.32	31.32	15.20
RF				
CCC and Bands only	**0.40**	**17.08**	**26.16**	**13.18**
CCC and Indices only	0.24	19.16	37.68	15.35
CCC, Bands and Indices	0.31	18.43	37.66	14.83

CCC model performance

Results in Table 6 and In GGHNP, the RF model again demonstrated better estimation capability of CCC in comparison to the SMLR modelling results. In particular, the RF model using both spectral bands and VIs as predictor variables, yielded a better predictive performance with low RMSE and MAE values of 21.15 µg.cm-2 and 16.72 respectively, capturing relatively higher CCC variability (R2 = 0.53) compared to SMLR in GGHNP (Table 7). Furthermore, the RF model based only on VIs as predictor variables, became the second-best performing model in predicting CCC in GGHNP. Therefore, the RF models based on a combination of bands and VIs, and band only, provided reasonable CCC prediction accuracies i.e. RRMSE's of 23.25% and 23.13% compared to SMLR in GGHNP, respectively (Table 7).

show the relative predictive performance of SMLR and RF models tested on different modelling scenarios involving the Sentinel-2 spectral bands and/or VIs for estimation of CCC in MNP and GGHNP. For each modelling scenario, selection of the optimal CCC model was based on the lowest RMSE and MAE followed by RRMSE. The CCC modelling results in

MNP (Table 6) show that RF yielded a better predictive performance with RMSE and MAE values of 17.08 μg.cm^{-2} and 13.18 respectively, explaining about 40% of CCC variability when only the Sentinel-2 bands are used. Additionally, this model had a fair CCC estimation capability over MNP with RRMSE of approximately 26.16% compared to the SMLR and other modelling scenarios. This performance is marginally followed by the SMLR model using only the VIs as predictor variables, which had RMSE and MAE values of 17.82 μg.cm^{-2} and 14.13 respectively, explaining about 35% of CCC variability in MNP. Notwithstanding the encouraging performance of SMLR, the SMLR CCC prediction accuracies in MNP based on all three modelling scenarios were found to be inadequate i.e. RRMSE's > 30% (Table 6).

In GGHNP, the RF model again demonstrated better estimation capability of CCC in comparison to the SMLR modelling results. In particular, the RF model using both spectral bands and VIs as predictor variables, yielded a better predictive performance with low RMSE and MAE values of 21.15 μg.cm^{-2} and 16.72 respectively, capturing relatively higher CCC variability (R^2 = 0.53) compared to SMLR in GGHNP (Table 7). Furthermore, the RF model based only on VIs as predictor variables, became the second-best performing model in predicting CCC in GGHNP. Therefore, the RF models based on a combination of bands and VIs, and band only, provided reasonable CCC prediction accuracies i.e. RRMSE's of 23.25% and 23.13% compared to SMLR in GGHNP, respectively (Table 7).

The SMLR CCC results in GGHNP based on all three modelling scenarios, yielded inadequate prediction accuracies with relatively high RRMSE's reaching approximately 57.49%. Interestingly, there is a notable drop in the RRMSE to 47.73% for the SMLR, which suggests that, the inclusion of both the bands and VIs in the modelling process may improve the prediction accuracy of CCC in GGHNP. This is evident when observing the RF model results based on the bands and VIs modelling scenario (Table 7).

Selection of important predictor variables

The packages used for SMLR and RF algorithms, namely the AIC (Hebbali and Hebbali 2017) and VSURF (Genuer, Poggi, and Tuleau-Malot 2015) respectively, have built-in variable selection measures that were executed for different modelling scenarios, in order to find optimal models based on a variable set of important predictors i.e. Sentinel-2 bands and/or VIs. Table 8 and Table 9 show results of the chosen optimal models and their corresponding variable sets of important variables for estimating grass LAI and CCC in the MNP and GGHNP heterogeneous sites. For LAI estimation, the model scenario involving the combination of bands and VIs was common for both sites, but the important variables differ (Table 8). However, for estimation of CCC, different modelling scenarios involving Sentinel-2 bands only and both the bands and VIs combined, yielded optimal models in MNP and GGHNP sites respectively, with each model having a different set of important variables (Table 9).

Table 7: Estimation of LAI best models based on important variables in MNP and GGHNP

Model Scenario	**R^2**	**RMSE ($m^2.m^{-2}$)**	**RRMSE (%)**	**MAE**	**Important variables**
SMLR for LAI in MNP					
LAI, Bands and Indices	0.48	0.67	27.31	0.54	B1, B2, B7, B8, B12, SR1, SR4, SR5, SR12, NDVI2, NDVI3, NDVI4, NDVI5, RBNDVI, RE, MCARI, LCI
RF for LAI in GGHNP					
LAI, Bands and Indices	0.43	0.93	22.24	0.74	PNDVI, REP, CIRE, SR9, B11

For LAI estimation in MNP, SMLR gave a better performance whereas in GGHNP, RF performed better (Table 8). A set of important variables for the RF LAI model included the Shortwave infrared$_1$ band coupled with red-edge based indices i.e. PNDVI, REP, CIRE, SR9. Furthermore, for the SMLR LAI model had a relatively longer list of important variables comprising the Sentinel-2 bands (i.e. coastal aerosol, blue, red edge$_3$, near infrared and shortwave infrared$_2$) and indices such as SR1, SR4, SR5, SR12, NDVI2, NDVI3, NDVI4, NDVI5, RBNDVI, RE, MCARI and LCI. Both LAI optimal models by SMLR and RF in MNP and GGHNP respectively (Table 8), did not have any overlap between their important variables; thus, each LAI model per site had a unique set of important predictor variables.

Table 8: Estimation of CCC best models based on important variables in MNP and GGHNP.

Model Scenario	**R^2**	**RMSE (µg.cm^{-2})**	**RRMSE (%)**	**MAE**	**Important variables**
RF for CCC in MNP					
CCC and Bands only	0.40	17.08	26.16	13.18	B12, B1, B8A, B4
RF for CCC in GGHNP					
CCC, Bands and Indices	0.53	21.15	23.25	16.72	CIRE, B11, SR6, B5, REP, SR9

For CCC estimation in MNP and GGHNP, RF gave a better predictive performance in both sites compared to SMLR (Table 9). In MNP, a set of important variables for the RF CCC optimal model included only a few Sentinel-2 bands i.e. shortwave infrared$_2$, coastal aerosol, narrow near infrared and red. While in GGHNP, the RF CCC optimal model had a set of important variables comprising the Sentinel-2 bands (i.e. shortwave infrared$_2$ and red edge$_1$) and red-edge based indices such as CIRE, SR6, REP and SR9 (Table 9). Both RF CCC optimal

models in MNP and GGHNP respectively, did not have any overlap between their important predictor variables; thus, each RF CCC model per site had a unique set of important variables. Given that a similar pattern was also observed with LAI models for the two sites, this could be an indication of the heterogeneous nature of the sites in-terms of grass species diversity, diversity of land cover and varying terrain slopes (Tsele et al. 2022).

Interestingly, these results suggest that in GGHNP alone, the RF models can provide better estimates of both LAI and CCC (Table 8 and Table 9). In addition, common important variables namely, CIRE, B11, REP and SR9 were found to be among the ideal predictors of both LAI and CCC in GGHNP according to the RF models (Table 8 and Table 9). Whereas in MNP, similar important variables (i.e. B12 and B1) were chosen according to the SMLR and RF models in the estimation of LAI and CCC, respectively. Furthermore, the Wilcoxon rank sum test (Rey and Neuhäuser 2011) was used to evaluate the statistical significance of the difference between the resulting accuracies of the prediction models in Table 8 and Table 9. The resulting values of the Wilcoxon test exceeded the significance level of 0.05 for all the models, meaning there is no statistically significant difference between the median of the CCC estimated in GGHNP and the CCC estimated in MNP i.e. p-value = 0.51. Similarly, there is no statistically significant difference between the median of the LAI estimated in GGHNP and that estimated in MNP i.e. p-value = 0.99.

LAI and CCC prediction maps of MNP and GGHNP

Figure 3 show the realistic patterns of LAI and CCC spatial predictions across the GGHNP influenced by numerous variables such as rainfall, temperature, seasons, underlying soil and geology types, topography and vegetation type. The LAI spatial variations suggest that the region is characterised largely by high biomass. For example, all sections of the GGHNP were estimated to have, on average LAI values > 2 with the largest LAI mean values of about 2.7,

2.6 and 2.5 corresponding to the top-left section (locally named as Little Serengeti), top-right section (locally identified as Witkrans) and bottom-right section (locally known as Heuweltop) respectively (Figure 3). A similar pattern of spatial variations of CCC is also evident across the sections.

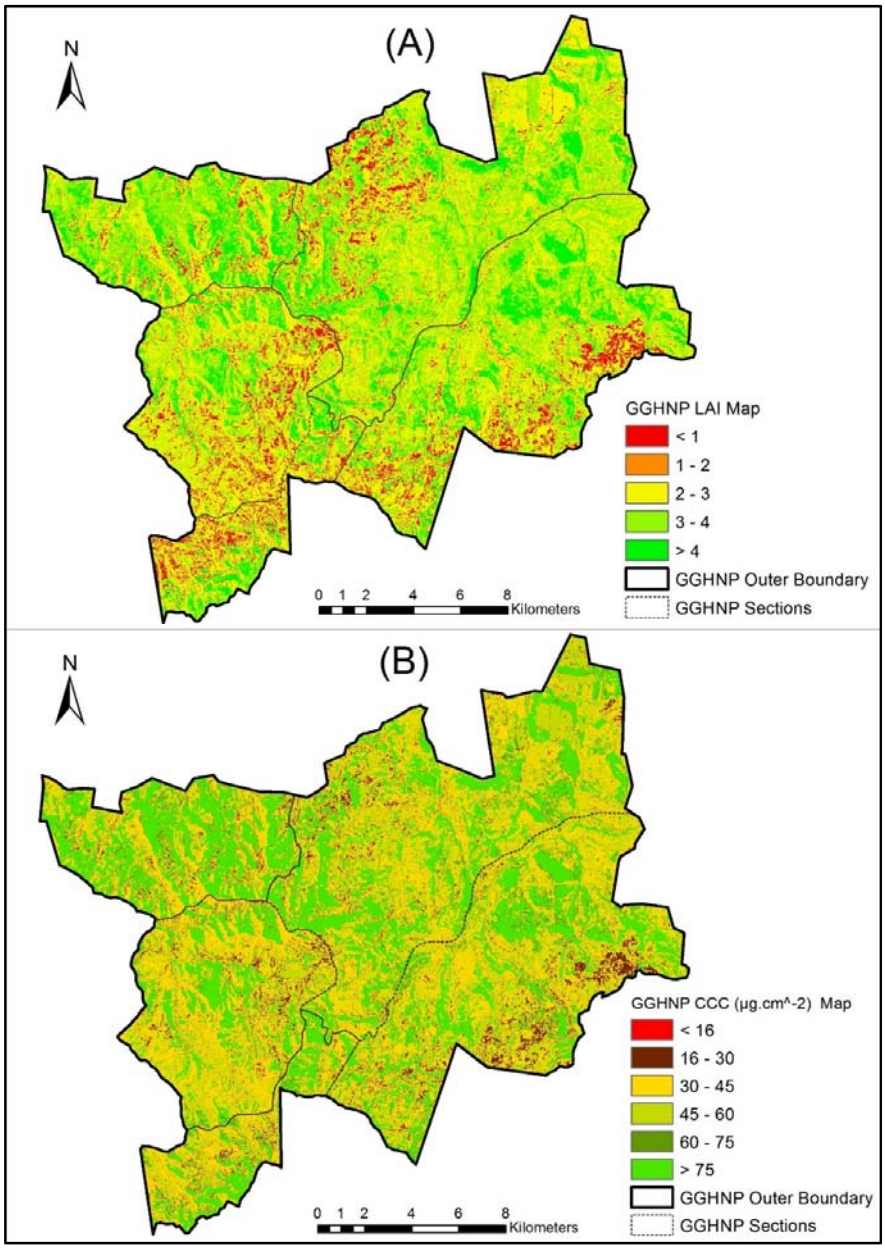

Figure 3: The LAI (A) and CCC (B) distribution maps in GGHNP estimated using the best performing models based on the RF algorithm (i.e. in Table 8 and Table 9 respectively) during peak productivity of the grassland.

The vegetation types found in these sections are (in the order of dominance) the eastern Free State sandy grassland, Basotho montane shrubland, Lesotho highland basalt grassland and northern Drakensberg highland grassland (Mucina and Rutherford 2006). Interestingly, the areas that appeared reddish on the map with lower LAI values < 1 (Figure 3) were found to coincide with mostly basaltic lava and mudstone underlying geology types according to the South African national geology map (CG 1997). Whereas, higher LAI values coincided largely with the fine-grained sandstone underlying geology type. Furthermore, higher grass LAI values > 3 were mostly in regions dominated by shallow, deeper and gravelly soils (Leptosols), whereas lower grass LAI values occurred largely in regions dominated by silt, clay and loam soils (https://data.isric.org/). The estimated CCC followed the LAI trends in these regions (Figure 3). These estimations in GGHNP could be used to identify and monitor potential hotspots where the grazers are most likely to be found. In addition, overgrazed areas coupled with the seasonal effects on the varying concentrations of vegetation biophysical variables (like LAI and CCC) can also be monitored.

Figure 4 show the realistic patterns of LAI and CCC estimations across the MNP region which could be influenced by numerous variables such as seasons, soil type, underlying geology, elevation and vegetation type. For example, in the western part of MNP which is dominated by the sandy bushveld vegetation type (Mucina and Rutherford 2006), clay-rich subsoil (ferric lixisols) and mudstone geology was modelled to have, on average lower LAI values (<= 2) with variable CCC in the range 18 to 81 ($\mu g.cm^{-2}$). This LAI estimate may suggest the area in the western region has low biomass, thus could be characterised by low volume grazing. However, the central and eastern parts of MNP that are largely characterised by moderate to high elevation (i.e. ~1024 to 2091 m), mountain bushveld vegetation type (Mucina and Rutherford 2006), sandstone and siltstone geology types and shallow-gravel soil, were modelled to have, on average higher LAI values of about 3.

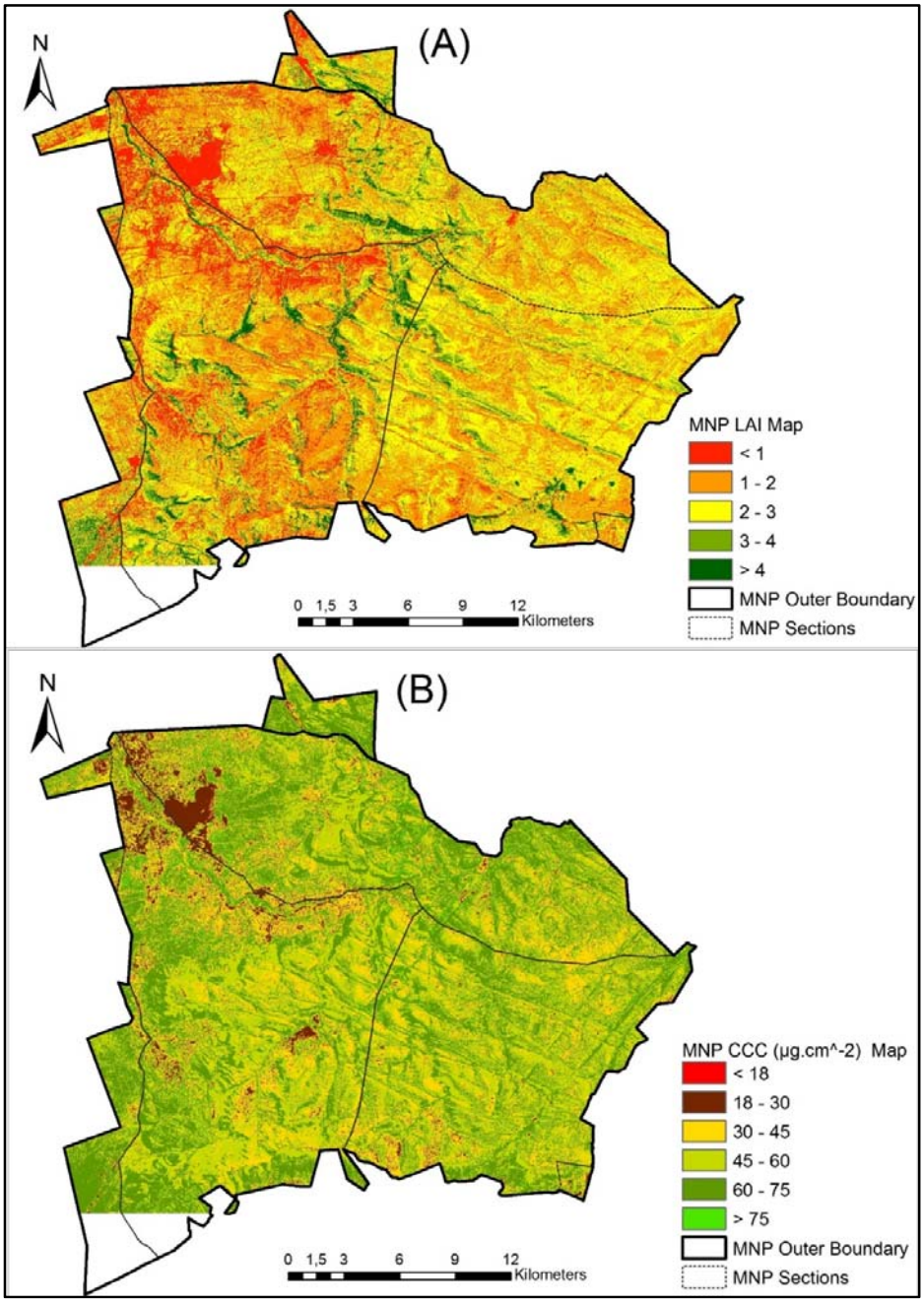

Figure 4: The LAI (A) and CCC (B) distribution maps in MNP estimated using the best performing models based on the SMLR algorithm (i.e. in Table 8) and the RF algorithm (i.e. in Table 9) respectively, during peak productivity of the grassland.

Most areas that were estimated to have the highest LAI > 4 (Figure 4), appeared to have some linear disconnected patterns suggesting that it may be in the water-logged areas and wetlands i.e. channelled-valley bottom wetlands. This suggestion was confirmed by use of the DEM map (not shown). CCC also appeared very high (> 60 $\mu g.cm^{-2}$) along the aforementioned areas. These estimations (Figure 4) can be used to infer that the central, eastern including a section of the northern parts of MNP are characterised by high volume grazing or at least that is where most grazers are located. This observation is yet to be confirmed with animal location data by linking it with the maps (Figure 4).

Assessing the effect of transferring varying proportions of field samples on model accuracy

Table 10 – Table 13 show statistical modelling results based on varying proportions of field samples that were applied on the best performing LAI and CCC estimation models presented earlier in Table 8 and Table 9. The CCC and LAI estimation models of GGHNP showed improved model accuracies when 50% and 75% of the MNP field samples were transferred to the GGHNP models. In particular, the CCC estimation model of GGHNP showed minor improvements with RMSEs of 21.06 $\mu g.cm^{-2}$ and 21.12 $\mu g.cm^{-2}$ and RRMSEs of 23.14% and 23.24% when the varying proportions of 50% and 75% of the MNP field samples were used, respectively (Table 10) compared to the original CCC estimation model with RMSE of 21.15 $\mu g.cm^{-2}$ and RRMSE of 23.25% (Table 9). Virtually similar improvements could notably be seen in the LAI estimation model of GGHNP with RMSEs of 0.85 and 0.89 corresponding to RRMSEs of 21.14% and 21.36% when the varying proportions of 75% and 50% of the MNP field samples were used, respectively (Table 11) compared to the original LAI estimation model with RMSE of 0.93 and RRMSE of 22.24% (Table 8). However, when transferring less (< 50%) or more (>75%) of the MNP field samples, the accuracies of the GGHNP CCC and LAI models tend to decline i.e. higher RMSEs and RRMSEs coupled with lower R^2 values.

This could be an indication of the dynamics in the two study sites, in that they have different types of grasses, fire regimes, and also dissimilar dominant biomes i.e. GGHNP is mainly grassland and MNP is both grassland and mesic savanna (Mucina and Rutherford (2006)).

Table 9: The best performing RF model for CCC in GGHNP was used (Table 8). A 70%/30% testing and validation split only for the scenario GGHNP: 100% and MNP: 100%. Cross validation was used for the rest of the scenarios with varying proportions of samples.

Field samples used %	R^2	RMSE (µg.cm^{-2})	RRMSE (%)	MAE
GGHNP: 100% MNP: 100%	0.32	22.43	53.23	15.70
GGHNP: 100% MNP: 75%	0.36	21.12	23.24	16.39
GGHNP: 100% MNP: 50%	0.45	21.06	23.14	16.47
GGHNP: 100% MNP: 25%	0.43	22.48	24.30	17.45

In contrast to the improved model results (Table 10 and Table 11), the CCC and LAI estimation models of MNP showed a decline in model performance across all scenarios where the GGHNP field samples were transferred to the MNP models (Table 12 and Table 13). Notwithstanding the aforementioned decline, the scenario where 100% of the GGHNP field samples were transferred to the MNP models had better RSME values (for both CCC and LAI estimations) compared to other scenarios with varying proportions of samples. For example, the scenarios with proportions of 50% and 75% GGHNP field samples, further lowered the MNP model performance, evident in the increasing RSME values in the approximate range of 24.97-25.16 µg.cm^{-2} and 0.92-0.95 for CCC and LAI estimations, respectively (Table 12 and

Table 13) compared to the original CCC (17.08 μg.cm^{-2}) and LAI (0.67) estimation models of MNP (Table 8 and Table 9).

Table 10: The best performing RF model for LAI in GGHNP was used (**Error! Reference source not found.**). A 70%/30% testing and validation split was performed only for the scenario GGHNP: 100% and MNP: 100%. Cross validation was used for the rest of the scenarios with varying proportions of samples.

Field samples used %	R^2	RMSE ($m^2.m^{-2}$)	RRMSE (%)	MAE
GGHNP: 100% MNP: 100%	0.32	0.97	47.67	0.72
GGHNP: 100% MNP: 75%	0.38	0.85	21.14	0.66
GGHNP: 100% MNP: 50%	0.38	0.89	21.36	0.70
GGHNP: 100% MNP: 25%	0.39	0.92	21.91	0.74

Furthermore, a notable drop is evident in the R^2 values and suggests using proportions of GGHNP field samples below 100% in the MNP models, did not positively contribute in capturing the variability the LAI and CCC across the MNP site. In overall, the decline in model performance across all scenarios where the GGHNP field samples were transferred to the MNP models may be attributed to the dynamic nature of the two study sites. For example, the GGHNP field samples were transferred into the MNP models that represent a site comprising both the grassland and savanna biomes. Whereas, the MNP field samples were transferred into the GGHNP models that represent a site covered by only the grassland biome. This may have been the reason for the improved model accuracies in GGHNP when 50% and 75% of the MNP grass samples were transferred to the GGHNP models.

Table 11: The best performing RF model for CCC in MNP was used (Table 8). A 70%/30% testing and validation split only for the scenario MNP: 100% and GGHNP: 100%. Cross validation was used for the rest of the scenarios with varying proportions of samples.

Field samples used %	R^2	RMSE ($\mu g.cm^{-2}$)	RRMSE (%)	MAE
MNP: 100% GGHNP: 100%	0.19	19.0	47.26	15.15
MNP: 100% GGHNP: 75%	0.15	24.97	35.12	18.14
MNP: 100% GGHNP: 50%	0.18	25.16	37.06	18.25
MNP: 100% GGHNP: 25%	0.25	20.51	30.52	15.79

Table 12: The best performing SMLR model for LAI in MNP was used (**Error! Reference source not found.**). A 70%/30% testing and validation split only for the scenario MNP: 100% and GGHNP: 100%. Cross validation was used for the rest of the scenarios with varying proportions of samples.

Field samples used %	R^2	RMSE ($m^2.m^{-2}$)	RRMSE (%)	MAE
MNP: 100% GGHNP: 100%	0.26	0.85	43.51	0.65
MNP: 100% GGHNP: 75%	0.20	0.95	23.0	0.72
MNP: 100% GGHNP: 50%	0.20	0.92	22.58	0.71
MNP: 100% GGHNP: 25%	0.19	0.98	25.49	0.73

Conclusions

This paper developed locally parameterised empirical models to evaluate the Sentinel-2 spectral reflectance bands and various VIs for the estimation of grass LAI and CCC during peak productivity over heterogeneous environments in two South African National Parks. Our findings show that SMLR yielded better LAI estimation in MNP when selected bands and indices are combined as predictor variables. Whereas, for LAI estimation in GGHNP, RF gave a better performance based on a unique set of important predictor variables such as the PNDVI, REP, CIRE, SR9 and B11 compared to SMLR in MNP. Furthermore, RF yielded better predictive performance in the estimation of CCC in both MNP and GGHNP sites. These results suggest that in GGHNP alone, the RF models can provide better estimates of both LAI and CCC. The resulting values of the Wilcoxon test exceeded the significance level of 0.05 for all the models, meaning there is no statistically significant difference in their predictive performance in estimating LAI and CCC. The generated prediction maps of GGHNP and MNP showed realistic spatial patterns of LAI and CCC estimates influenced numerous variables such as climate, seasons, topography, vegetation type as well as soil and geology types.

Furthermore, the CCC and LAI estimation models of GGHNP showed improved model accuracies when 50% and 75% of the MNP field samples were transferred to the GGHNP models. In contrast, the CCC and LAI estimation models of MNP showed a decline in model performance across all scenarios where the GGHNP field samples were transferred to the MNP models. The relative performance in model accuracies may be attributed to the dynamic nature of the two study sites. Nonetheless, this study showed that locally parameterised empirical models can be improved through transfer scenarios involving different proportions of field samples from different sites, based on the assumption that the range of field sample values of biophysical parameters between the sites are not far apart. Overall, these findings prompt the

need for further development of locally parameterised types of models over heterogenous ecosystems.

Author Contributions: Conceptualisation, P.T. and A.R.; methodology, P.T. and A.R.; Formal analysis, P.T. and M.Q; validation, P.T., A.R, and M.Q.; writing—original draft preparation, P.T.; writing—review and editing, A.R.; project administration, P.T.;

Funding: This research was funded by Research development programme of the University of Pretoria, as well as the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) of the National Research Foundation (NRF) of South Africa, grant number 118590.

Acknowledgments: The Sentinel-2 data used in this study were downloaded from the European Space Agency Copernicus Open Access Hub. We sincerely thank the field assistants (namely Phomolo Seriba, Katlego Mashiane, Steven Khosa and Brian Mabunda) in the Golden Gate Highlands National Park and Marakele National Park for their collaborative effort in collecting grass LAI, FVC and LCC ground measurements. We also thank the anonymous reviewers for their valuable and insightful comments.

Conflicts of Interest: The authors declare no conflict of interest.

Data availability statement: We understand that the publication of the data is becoming a good practice in research. However, we plan to share all our data in future, but at this stage we are still going to further analyse it for locally parameterised types of models, looking at both empirical and the inversion of the physically-based models.

References

Ali, Abebe Mohammed, Roshanak Darvishzadeh, Andrew Skidmore, Tawanda W Gara, and Marco Heurich. 2021. "Machine learning methods' performance in radiative transfer model inversion to retrieve plant traits from Sentinel-2 data of a mixed mountain forest." *International journal of digital earth* 14 (1):106-20.

Ali, Abebe Mohammed, Roshanak Darvishzadeh, Andrew Skidmore, Tawanda W Gara, Brian O'Connor, Claudia Roeoesli, Marco Heurich, and Marc Paganini. 2020. "Comparing methods for mapping canopy chlorophyll content in a mixed mountain forest using Sentinel-2 data." *International Journal of Applied Earth Observation and Geoinformation* 87:102037.

Andreatta, Davide, Damiano Gianelle, Michele Scotton, and Michele Dalponte. 2022. "Estimating grassland vegetation cover with remote sensing: A comparison between Landsat-8, Sentinel-2 and PlanetScope imagery." *Ecological Indicators* 141:109102.

Atzberger, Clement, Roshanak Darvishzadeh, Markus Immitzer, Martin Schlerf, Andrew Skidmore, and Guerric Le Maire. 2015. "Comparative analysis of different retrieval methods for mapping grassland leaf area index using airborne imaging spectroscopy." *International Journal of Applied Earth Observation and Geoinformation* 43:19-31.

Bei, CUI, Qian-jun ZHAO, Wen-jiang HUANG, Xiao-yu SONG, Hui-chun YE, and Xian-feng ZHOU. 2019. "Leaf chlorophyll content retrieval of wheat by simulated RapidEye, Sentinel-2 and EnMAP data." *Journal of Integrative Agriculture* 18 (6):1230-45.

BON, GEO. 2015. "Global Biodiversity Change Indicators: Model-Based Integration of Remote-Sensing & In Situ Observations That Enables Dynamic Updates and Transparency at Low Cost." In.: GEO BON Secretariat Leipzig, Germany.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45 (1):5-32.

Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 2017. *Classification and regression trees*: Routledge.

Brown, Luke A, Richard Fernandes, Najib Djamai, Courtney Meier, Nadine Gobron, Harry Morris, Francis Canisius, Gabriele Bai, Christophe Lerebourg, and Christian Lanconelli. 2021. "Validation of baseline and modified Sentinel-2 Level 2 Prototype Processor leaf area index retrievals over the United States." *ISPRS Journal of Photogrammetry and Remote Sensing* 175:71-87.

Brown, Luke A, Booker O Ogutu, and Jadunandan Dash. 2019. "Estimating forest leaf area index and canopy chlorophyll content with Sentinel-2: An evaluation of two hybrid retrieval algorithms." *Remote Sensing* 11 (15):1752.

Bsaibes, Aline, Dominique Courault, Frédéric Baret, Marie Weiss, Albert Olioso, Frédéric Jacob, Olivier Hagolle, Olivier Marloie, Nadine Bertrand, and Véronique Desfond. 2009. "Albedo and LAI estimates from FORMOSAT-2 data for crop monitoring." *Remote Sensing of Environment* 113 (4):716-29.

CG. 1997. "1: 1 000 000 Scale Geological Map of the Republic of South Africa and the Kingdoms of Lesotho and Swaziland." In.: Council for Geosciences Pretoria.

Chai, Tianfeng, and Roland R Draxler. 2014. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature." *Geoscientific model development* 7 (3):1247-50.

Chen, Weitao, Xianju Li, Yanxin Wang, Gang Chen, and Shengwei Liu. 2014. "Forested landslide detection using LiDAR data and the random forest algorithm: A case study of the Three Gorges, China." *Remote Sensing of Environment* 152:291-301.

Cho, Moses Azong, Abel Ramoelo, and Renaud Math. 2014. Estimation of leaf area index (LAI) of South Africa from MODIS imager by inversion of PROSAIL radiative transfer model. Paper presented at the 2014 IEEE Geoscience and Remote Sensing Symposium.

Chuvieco, Emilio. 2016. *Fundamentals of satellite remote sensing: An environmental approach*: CRC press.

Clevers, Jan GPW, and Anatoly A Gitelson. 2013. "Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and-3." *International Journal of Applied Earth Observation and Geoinformation* 23:344-51.

Combal, B, Frédéric Baret, M Weiss, Alain Trubuil, D Mace, A Pragnere, R Myneni, Y Knyazikhin, and L Wang. 2003. "Retrieval of canopy biophysical variables from bidirectional reflectance: Using prior information to solve the ill-posed inverse problem." *Remote Sensing of Environment* 84 (1):1-15.

Darvishzadeh, Roshanak, Clement Atzberger, Andrew Skidmore, and Martin Schlerf. 2011. "Mapping grassland leaf area index with airborne hyperspectral imagery: A comparison study of statistical approaches and inversion of radiative transfer models." *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (6):894-906.

Darvishzadeh, Roshanak, Andrew Skidmore, Martin Schlerf, and Clement Atzberger. 2008. "Inversion of a radiative transfer model for estimating vegetation LAI and chlorophyll in a heterogeneous grassland." *Remote Sensing of Environment* 112 (5):2592-604.

Datt, Bisun. 1999. "A new reflectance index for remote sensing of chlorophyll content in higher plants: tests using Eucalyptus leaves." *Journal of Plant Physiology* 154 (1):30-6.

Daughtry, Craig ST, CL Walthall, MS Kim, E Brown De Colstoun, and JE McMurtrey Iii. 2000. "Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance." *Remote Sensing of Environment* 74 (2):229-39.

Dawson, TP, and PJ Curran. 1998. "Technical note A new technique for interpolating the reflectance red edge position."

Delegido, Jesús, Jochem Verrelst, Luis Alonso, and José Moreno. 2011. "Evaluation of sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content." *Sensors* 11 (7):7063-81.

Eberly, Lynn E. 2007. "Multiple linear regression." *Topics in Biostatistics*:165-87.

Filipponi, Federico. 2021. "Comparison LAI estimates from of high resolution satellite observations using different biophysical processors." In.: Pro-ceedings.

Gastwirth, Joseph L. 1971. "On the sign test for symmetry." *Journal of the American Statistical Association* 66 (336):821-3.

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. 2015. "VSURF: an R package for variable selection using random forests." *The R Journal* 7 (2):19-33.

Gitelson, Anatoly A, Yuri Gritz, and Mark N Merzlyak. 2003. "Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves." *Journal of Plant Physiology* 160 (3):271-82.

Gitelson, Anatoly A, Galina P Keydan, and Mark N Merzlyak. 2006. "Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves." *Geophysical Research Letters* 33 (11).

Gitelson, Anatoly A, Andrés Vina, Verónica Ciganda, Donald C Rundquist, and Timothy J Arkebauer. 2005. "Remote estimation of canopy chlorophyll content in crops." *Geophysical Research Letters* 32 (8).

Goel, Narendra S. 1987. "Inversion of canopy reflectance models for estimation of vegetation parameters." In.

Guerini Filho, Marildo, Tatiana Mora Kuplich, and Fernando LF De Quadros. 2020. "Estimating natural grassland biomass by vegetation indices using Sentinel 2 remote sensing data." *International Journal of Remote Sensing* 41 (8):2861-76.

Hebbali, Aravind, and Maintainer Aravind Hebbali. 2017. "Package 'olsrr'." *Version 0.5* 3.

Heinemann, Alexandre Bryan, Pepijn AJ Van Oort, Diogo Simões Fernandes, and Aline de Holanda Nunes Maia. 2012. "Sensitivity of APSIM/ORYZA model due to estimation errors in solar radiation." *Bragantia* 71:572-82.

Henrich, V, G Krauss, C Götze, and C Sandow. 2012. "Index DataBase-A database for remote sensing indices [WWW Document]." *IDB-Entwicklung einer Datenbank für Fernerkundungsindizes. URL www. indexdatabase. de*.

Horler, DNH, M Dockray, J Barber, and AR Barringer. 1983. "Red edge measurements for remotely sensing plant chlorophyll content." *Advances in Space Research* 3 (2):273-7.

Horler, DNH, Mo DOCKRAY, and J Barber. 1983. "The red edge of plant leaf reflectance." *International Journal of Remote Sensing* 4 (2):273-88.

Hu, Qiong, Jingya Yang, Baodong Xu, Jianxi Huang, Muhammad Sohail Memon, Gaofei Yin, Yelu Zeng, Jing Zhao, and Ke Liu. 2020. "Evaluation of global decametric-resolution LAI, FAPAR and FVC estimates derived from Sentinel-2 imagery." *Remote Sensing* 12 (6):912.

Jamieson, PD, JR Porter, and DR Wilson. 1991. "A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand." *Field Crops Research* 27 (4):337-50.

Jonckheere, Inge, Stefan Fleck, Kris Nackaerts, Bart Muys, Pol Coppin, Marie Weiss, and Frédéric Baret. 2004. "Review of methods for in situ leaf area index determination: Part I. Theories, sensors and hemispherical photography." *Agricultural and Forest Meteorology* 121 (1-2):19-35.

Jordan, Carl F. 1969. "Derivation of leaf-area index from quality of light on the forest floor." *Ecology* 50 (4):663-6.

Kganyago, Mahlatse, Paidamwoyo Mhangara, Thomas Alexandridis, Giovanni Laneve, Georgios Ovakoglou, and Nosiseko Mashiyi. 2020. "Validation of sentinel-2 leaf area index (LAI) product derived from SNAP toolbox and its comparison with global LAI products in an African semi-arid agricultural landscape." *Remote Sensing Letters* 11 (10):883-92.

Kganyago, Mahlatse, Clement Adjorlolo, and Paidamwoyo Mhangara. 2022. "Exploring Transferable Techniques to Retrieve Crop Biophysical and Biochemical Variables Using Sentinel-2 Data." *Remote Sensing* 14 (16):3968.

Kganyago, Mahlatse, Paidamwoyo Mhangara, and Clement Adjorlolo. 2021. "Estimating crop biophysical parameters using machine learning algorithms and Sentinel-2 imagery." *Remote Sensing* 13 (21):4314.

Kuhn, Max. 2008. "Building predictive models in R using the caret package." *Journal of statistical software* 28:1-26

Li, Ying, Nianpeng He, Jihua Hou, Li Xu, Congcong Liu, Jiahui Zhang, Qiufeng Wang, Ximin Zhang, and Xiuqin Wu. 2018. "Factors influencing leaf chlorophyll content in natural forests at the biome scale." *Frontiers in Ecology and Evolution* 6:64.

Liang, Liang, Zhihao Qin, Shuhe Zhao, Liping Di, Chao Zhang, Meixia Deng, Hui Lin, Lianpeng Zhang, Lijuan Wang, and Zhixiao Liu. 2016. "Estimating crop chlorophyll content with hyperspectral vegetation indices and the hybrid inversion method." *International Journal of Remote Sensing* 37 (13):2923-49.

Louis, Jérôme, Vincent Debaecker, Bringfried Pflug, Magdalena Main-Knorn, Jakub Bieniarz, Uwe Mueller-Wilm, Enrico Cadau, and Ferran Gascon. 2016. Sentinel-2 Sen2Cor: L2A processor for users. Paper presented at the Proceedings Living Planet Symposium 2016.

Masemola, Cecilia, Moses Azong Cho, and Abel Ramoelo. 2016. "Comparison of Landsat 8 OLI and Landsat 7 ETM+ for estimating grassland LAI using model inversion and spectral indices: case study of Mpumalanga, South Africa." *International Journal of Remote Sensing* 37 (18):4401-19.

Mucina, Ladislav, and Michael C Rutherford. 2006. *The vegetation of South Africa, Lesotho and Swaziland*: South African National Biodiversity Institute.

Mutanga, Onisimo, Elhadi Adam, and Moses Azong Cho. 2012. "High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm." *International Journal of Applied Earth Observation and Geoinformation* 18:399-406.

Ramoelo, Abel, and Moses Azong Cho. 2018. "Explaining leaf nitrogen distribution in a semi-arid environment predicted on Sentinel-2 imagery using a field spectroscopy derived model." *Remote Sensing* 10 (2):269.

Ramoelo, Abel, Moses Azong Cho, Renaud Mathieu, Sabelo Madonsela, Ruben Van De Kerchove, Zaneta Kaszta, and Eléonore Wolff. 2015. "Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and WorldView-2 data." *International Journal of Applied Earth Observation and Geoinformation* 43:43-54.

Ramoelo, Abel, Andrew K Skidmore, Moses Azong Cho, Martin Schlerf, Renaud Mathieu, and Ignas MA Heitkönig. 2012. "Regional estimation of savanna grass nitrogen using the red-edge band of the spaceborne RapidEye sensor." *International Journal of Applied Earth Observation and Geoinformation* 19:151-62.

Rey, Denise, and Markus Neuhäuser. 2011. "Wilcoxon-signed-rank test." In *International encyclopedia of statistical science*, 1658-9. Springer.

Richter, Katja, Tobias B Hank, Wolfram Mauser, and Clement Atzberger. 2012. "Derivation of biophysical variables from Earth observation data: validation and statistical measures." *Journal of Applied Remote Sensing* 6 (1):063557.

Rodriguez-Galiano, Victor Francisco, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. 2012. "An assessment of the effectiveness of a random forest classifier for land-cover classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 67:93-104.

Rouse Jr, John W, R Hect Haas, JA Schell, and DW Deering. 1973. "Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation." In.

Sakowska, Karolina, Radoslaw Juszczak, and Damiano Gianelle. 2016. "Remote sensing of grassland biophysical parameters in the context of the Sentinel-2 satellite mission." *Journal of Sensors* 2016.

Schwieder, Marcel, Marion Buddeberg, Katja Kowalski, Kira Pfoch, Julia Bartsch, Heike Bach, Jürgen Pickert, and Patrick Hostert. 2020. "Estimating grassland parameters from Sentinel-2: A model comparison study." *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 88 (5):379-90.

Skidmore, Andrew K, Nicholas C Coops, Elnaz Neinavaz, Abebe Ali, Michael E Schaepman, Marc Paganini, W Daniel Kissling, Petteri Vihervaara, Roshanak Darvishzadeh, and Hannes Feilhauer. 2021. "Priority list of biodiversity metrics to observe from space." *Nature ecology & evolution* 5 (7):896-906.

Snee, Ronald D. 1977. "Validation of regression models: methods and examples." *Technometrics* 19 (4):415-28.

Sun, Yuanheng, Qiming Qin, Huazhong Ren, Tianyuan Zhang, and Shanshan Chen. 2019. "Red-edge band vegetation indices for leaf area index estimation from sentinel-2/msi imagery." *IEEE Transactions on Geoscience and Remote Sensing* 58 (2):826-40.

Svinurai, Walter, Abubeker Hassen, Eyob Tesfamariam, and Abel Ramoelo. 2021. "Modelled effects of grazing strategies on native grass production, animal intake and growth in Brahman steers." *African Journal of Range & Forage Science*:1-11.

Tsele, Philemon, Abel Ramoelo, Mcebisi Qabaqaba, Madodomzi Mafanya, and George Chirima. 2022. "Validation of LAI, Chlorophyll and FVC biophysical estimates from Sentinel-2 Level 2 Prototype Processor over a heterogeneous savanna and grassland environment in South Africa." *Geocarto International* (just-accepted):1-22.

Van Engelen, VWP, and JA Dijkshoorn. 2013. "Global and national soils and terrain digital databases (SOTER)." *Report-ISRIC World Soil Information* (2013/04).

Van Staden, PJ, and GJ Bredenkamp. 2005. "Major plant communities of the Marakele National Park." *Koedoe* 48 (2):59-70.

Verrelst, Jochem, Juan Pablo Rivera, Frank Veroustraete, Jordi Muñoz-Marí, Jan GPW Clevers, Gustau Camps-Valls, and José Moreno. 2015. "Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods–A comparison." *ISPRS Journal of Photogrammetry and Remote Sensing* 108:260-72.

Wang, Fu-Min, Jing-Feng Huang, Yan-Lin Tang, and Xiu-Zhen Wang. 2007. "New vegetation index and its application in estimating leaf area index of rice." *Rice Science* 14 (3):195-203.

Weiss, Marie, Frederic Baret, and Sylvain Jay. 2020. "S2ToolBox Level 2 products LAI, FAPAR, FCOVER." EMMAH-CAPTE, INRAe Avignon.