REMOTE SENSING OF CROP BIOPHYSICAL AND BIOCHEMICAL PARAMETERS USING SENTINEL-2 AND MACHINE LEARNING ALGORITHMS

Mahlatse Lucky Kganyago

A thesis submitted to the School of Geography, Archaeology and Environmental Studies, Faculty of Science, the University of the Witwatersrand in fulfilment of the requirements for the degree of Doctor of Philosophy

Johannesburg, December 2022

DECLARATION 1

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

____
Mahlatse Lucky Kganyago

______ day of __________________in the year _______________.

DECLARATION 2

LIST OF OUTPUTS OF THE THESIS

These outputs (stand-alone original peer-reviewed manuscripts or conference proceedings) are **Under-Review**—a manuscript has been submitted to a journal and the editor has deemed it relevant and worthy of consideration to be published and is now being reviewed by experts in the field, or **Published**—a manuscript is now published in the journal issue and has a DOI.

1. **Kganyago, M.,** Adjorlolo, C., Mhangara, P., & Lesiba Tsoeleng (**Under-Review**). Remote Sensing of Crop Biophysical and Biochemical Parameters: An Overview of Advances in Sensor Technologies and Retrieval Techniques for Precision Agriculture. Computers and Electronics in Agriculture [Manuscript ID: COMPAG-D-22-03627].
2. **Kganyago, M.,** Ovakoglou, G., Mhangara, P., Alexandridis, T., Odindi, J., Adjorlolo, C., Mashiyi, N., Validation of atmospheric correction approaches for Sentinel-2 under partly cloudy conditions in an African agricultural landscape, Proc. SPIE 11531, Remote Sensing of Clouds and the Atmosphere XXV, 115310B (20 September 2020); https://doi.org/10.1117/12.2572293
3. **Kganyago, M.,** Mhangara, P., Adjorlolo, C. (**Under-Review**). Effect of Atmospheric Correction on Biophysical and Biochemical Retrieval: A Comparison of Sen2Cor and *i*Cor in Semi-arid African Agricultural landscape. *International Journal of Remote Sensing*. [Manuscript ID: TRES-PAP-2022-1462]
4. **Kganyago, M.,** Mhangara, P., Alexandridis, T., Laneve, G., Ovakoglou, G., Mashiyi, N. (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. DOI: https://doi.org/10.1080/2150704X.2020.1767823

5. **Kganyago, M.,** Mhangara, P., & Adjorlolo, C. (2021). Estimating Crop Biophysical Parameters Using Machine Learning Algorithms and Sentinel 2 Imagery. Remote Sensing, 13(21), 4314. https://doi.org/10.3390/rs13214314
6. **Kganyago, M**., Adjorlolo, C. & Mhangara, P. (2022). Exploring Transferable Techniques to Retrieve Crop Biophysical and Biochemical Variables using Sentinel-2 Data. Remote Sensing, 14(16), 3968. https://doi.org/10.3390/rs14163968
7. **Kganyago, M.,** Adjorlolo, C., Sibanda, M., Mhangara, P., Laneve, G., & Alexandridis, T. (**2022**). Testing Sentinel-2 Spectral Configurations for Estimating Relevant Crop Biophysical and Biochemical Parameters for Precision Agriculture Using Tree-based and Kernel-based Algorithms. Geocarto International. https://doi.org/10.1080/10106049.2022.2146764
8. **Kganyago, M**., G. Ovakoglou, Mhangara, P., Adjorlolo, C., Alexandridis, T., Laneve, G. & Beltran, J. S. (**2023**). Evaluating the Contribution of Sentinel-2 View and Illumination Geometry to the Accuracy of Retrieving Crop Biophysical and Biochemical Variables. GIScience and Remote Sensing. https://doi.org/10.1080/15481603.2022.2163046
9. **Kganyago, M**., Adjorlolo, C., Mhangara, P. (**Under-Review**). Novel Spectral Triads Feature Selection Algorithm for optimising Sentinel-2 Feature Space Towards Improved Crop Biophysical and Biochemical Parameter Retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. [Manuscript ID: JSTARS-2022-01690]

I declare that I had a maximum contribution (i.e., >80%) in each of the above-mentioned research outputs.

____

Mahlatse Lucky Kganyago

______ day of ________________________ in the year _______________.

ABSTRACT

Globally, achieving food security is critical to eliminating hunger, poverty, and malnutrition and reducing food-related diseases and deaths. This is enshrined in the United Nations' Sustainable Development Goals (SDGs), African Union's Agenda 2063, and national policies. The SDG 2 Target 2.4 aims to increase agricultural productivity and resilience to climate-related disasters, while also maintaining the integrity of agroecosystems. It is recommended to adopt innovative technologies and sustainable agricultural management practices to avert many challenges facing agricultural productivity and sustainability in food-insecure regions such as sub-Saharan Africa. This study sought to characterise the field variability in relevant crop biophysical and biochemical parameters for precision agriculture using the new generation sensor, Sentinel-2, and machine learning algorithms. Moreover, pressing issues regarding residual errors after atmospheric correction and their impact on the retrieval of crop parameters were addressed. The critical findings from this study are: (1) atmospheric correction (AC) techniques perform poorly under partly cloudy conditions, especially in the visible and SWIR region of the electromagnetic spectrum; (2) while it is beneficial to perform AC before estimating leaf and canopy chlorophyll content, LAI can be obtained directly from Top-of-Atmosphere (TOA) reflectance; (3) the study ascertained that there is no universal retrieval algorithm; hence, many have to be tested for specific sites, phenology, crop types, and crop parameters; (4) Sentinel-2 10 m (VNIR) and 20 m (RE-SWIR) bands can be used independently—without the need for super-resolving techniques or spatial resampling that may cause a ~7 min computational delay; (5) the incorporation of per-pixel view and illumination geometries to spectral bands and vegetation indices improves the accuracy of retrieving essential crop parameters; (6) the spatial transferability of the retrieval model can be improved by adding limited (new or unseen) samples from the target site to account for variability in that site; and (7) a novel Spectral Triad feature selection technique, developed in the current study, was capable of optimising Sentinel-2 Multi-spectral Imager (MSI) feature space (and performed better than the entire feature space and had similar performance with a well-established algorithm, i.e., Recursive Feature Elimination). The findings contribute to the existing knowledge of remotely sensed crop parameters and have practical implications for precision agriculture.

Keywords

precision agriculture; leaf area index; Sentinel-2; chlorophyll content; biophysical and biochemical parameters; remote sensing

DEDICATION

This work is dedicated to my gorgeous daughters: Oarona and Onalerena.

ACKNOWLEDGEMENTS

I would like to thank, from the bottom of my heart, my supervisors: Prof. Paidamwoyo Mhangara and Dr Clement Adjorlolo for their tutelage, encouragement, and support.

I acknowledge the University of the Witwatersrand, South African National Space Agency (SANSA), H2020-AfriCultuReS (Enhancing Food Security in African Agricultural Systems with the Support of Remote Sensing) project consortium led by GMV, and consisting of SANSA, Aristotle University of Thessaloniki, Sapienza Università di Roma, the University of Rwanda Centre for Remote Sensing and GIS, LocateIT, DRAXIS Environmental, University of Ghana, University of Mozambique, and University of Leeds, for supporting this study.

This research was conducted as part of the AfriCultuReS project which received funding from the European Union’s Horizon 2020 Research and Innovation Framework Programme under grant agreement No. 774652.

I acknowledge ESA Copernicus, United States Geological Survey (USGS), National Aeronautics and Space Administration (NASA) for providing satellite data free of charge, H2020-AfriCultuReS project Consortium for providing field data used in this study. European Space Agency (ESA) Network of Resourses (NoR) provided funded access to Planet and Sentinelhub Satellite Image Cloud-Platform by Synegise. This study used R Statistical software and R Studio for generating error metrics and graphs. We therefore appreciate the development teams for different R packages used here. We highly appreciate the implementation of Bidirectional Reflectance Distribution Function (BRDF) correction method for Sentinel-2 (Roy et al., 2017) in Google Earth Engine.

I would also like to express my appreciation for the participation of Ms Nosiseko Mashiyi, Mr Johnny Rizos, Mr Thomas Tsoeleng, Mr Morwapula Mashalane, Ms Tiisetso Kekana, and Mrs Andiswa Silinga during the fieldwork. The participating farmers at Harrismith and Bothaville are highly appreciated for allowing us to walk their fields.

Last but not least, I would like to thank my wife and best friend, Milly, who always supported and encouraged me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

AC	Atmospheric correction
ANN	Artificial Neural Networks
AOT	Aerosol Optical Thickness
ARD	Analysis Ready Data
AVHRR	Advanced Very-High-Resolution Radiometer
AVIRIS	Airborne Visible and Infrared Imaging Spectrometer
BOA	Bottom-of-Atmosphere
BRDF	Bidirectional Reflectance Distribution Function
BV	Biochemical variable
CART	Classification and Regression Trees
CASI	Compact Airborne Spectrographic Imager
CAT	Central African Time
CCC	Canopy Chlorophyll Content
CEOS	Committee on Earth Observation Satellites
CHRIS	Compact High-Resolution Imaging Spectrometry
CNN	Convolutional Neural Networks
CV	Coefficient of variation
CVI	Chlorophyll Vegetation Index
DALRRD	Department of Agriculture, Land Reform and Rural Development
DAS	Days after sowing
DDV	Dark, dense vegetation
DEM	Digital Elevation Model
DESIS	DLR Earth Sensing Imaging Spectrometer
DHP	Digital hemispherical photography
ECV	Essential climate variable
EO	Earth observation
ESA	European Space Agency
EVI	Enhanced Vegetation Index
FAO	Food and Agriculture Organisation
FOV	Field-of-view

FVC	Fractional Vegetation Cover
GBM	Gradient Boosting Machine
GBRT	Gradient Boosted Regression Trees
GCOS	Global Climate Observing System
GDP	Gross Domestic Product
GEE	Google Earth Engine
GEOGLAM	Group on Earth Observations-Global Agricultural Monitoring
GMES	Global Monitoring for Environmental and Security
GNSS	Global navigation satellite systems
GPR	Gaussian Process Regression
IFOV	Instantaneous field-of-view
IncMSE	Increase in Mean Square Error
JECAM	Joint Experiment for Crop Assessment and Monitoring
KRR	Kernel Ridge Regression
LAI	Leaf Area Index
LUT	Look-up tables
MAE	Mean absolute error
MAPE	Mean Absolute Percent Error
MERIS	MEdium Resolution Imaging Spectrometer
MLR	Machine Learning Regression
MLRA	Machine Learning Regression Algorithm
MODIS	MODerate Resolution Imaging Spectroradiometer
MSE	Mean Squared Error
MSI	Multi-spectral Imager
MTCI	MERIS terrestrial chlorophyll index
NASA	National Aeronautics and Space Administration
NBAR	Nadir BRDF-adjusted reflectance
NDI	Normalised difference index
NDVI	Normalised Difference Vegetation Index
NN	Neural networks
NOAA	National Oceanic and Atmospheric Administration
NoR	Network of Resources

OLCI	Ocean and Land Colour Instrument
OLI	Operational Land Imager
OLS	Ordinary Least Squares
OOB	Out-of-bag
PA	Precision agriculture
PCA	Principal Component Analysis
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
RAA	Relative Azimuth Angle
RBF	Radial Basis Function
RE	Red-edge
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Squared Error
RMSECV	Root Mean Squared Error of cross-validation
RRMSE	Relative root mean squared error
RTM	Radiative transfer model
SANSA	South African National Space Agency
SB	Spectral bands
SDG	Sustainable Development Goals
SDI	Spectral discrimination index
SNAP	Sentinel Application Platform
SPCA	Sparse Principal Component Analysis
SR	Surface reflectance
SRTM	Shuttle Radar Topography Mission
SSM	Site-specific farm management
STDFA	Spatial-Temporal Data Fusion Approach
SVI	Spectral vegetation indices
SVM	Support Vector Machine
SWIR	Shortwave Infrared
SZA	Sun Zenith Angle
TM	Thematic mapper

TOA	Top-of-Atmosphere
UAS	Unmanned Aerial System
UN	United Nations
USGS	United States Geological Survey
UTM	Universal Transverse Mercator
VAA	View and Sun Azimuth Angle
VHGPR	Variational Heteroscedastic Gaussian Process regression
VHR	Very-high-resolution
VIIRS	Visible Infrared Imaging Radiometer Suite
VNIR	Visible-near-infrared region
VRA	Variable Rate Application
VZA	View Zenith Angle
WDVI	Weighted Difference Vegetation Index

CHAPTER 1

INTRODUCTION

1. Background

1.1. Challenges facing global food security

One of the biggest concerns facing humanity today is ensuring global food security. To fulfil the demands of an exponentially expanding population, the Food and Agriculture Organisation (FAO) of the United Nations (UN) estimates that world food production must be boosted by 70% (UN, 2015). Food security is essential for eradicating hunger, poverty, and malnutrition and for lowering disease rates and food-related fatalities. It occurs when all persons, at all times, have access both physically and economically to adequate, safe, and nutritious food. The significance of achieving food security is enshrined in the UN Sustainable Development Goals (SDGs) and Agenda 2063 (Anderson et al. 2017; Kganyago & Mhangara 2019). For example, the UN-SDG 2 (Zero Hunger) Target 2.4 indicates that by 2030, the following sustainable food production systems and implementation of agricultural practices should be achieved (UN, 2015): (1) "increase productivity and production"; (2) "maintain the integrity of ecosystems"; (3) "strengthen the capability to adapt to climate change, extreme weather events, drought, flooding, and other disasters"; and (4) "improve land and soil quality". Generally, food security hinges on the productivity of various agricultural systems, sustainable land management practices, and resilience to extreme events through risk-averting mechanisms, particularly in sub-Saharan Africa where food shortages are rife.

Unfortunately, agricultural productivity—an indicator of an agricultural production system's effectiveness which utilises labour, land, and capital, among others—and its sustainability face a plethora of challenges owing to anthropogenic activities, environmental factors and climate change. For example, anthropogenic activities such as mining, urban expansion, and land degradation (Nellemann, 2009) threaten productive agricultural land. Moreover, climatic variability, and change (i.e., increased temperatures, erratic rainfall patterns, prolonged droughts, and heatwaves) severely threaten agricultural productivity. While the adoption and effective implementation of spatial planning policies that allow the preservation of productive land can curb competitive land uses, climate change is gradual and indiscernible until a significant loss in productivity has been experienced. Alarmingly, losses in productivity imply loss of income and jobs and slower economic growth since the agricultural industry is core to the economic development of many developing countries and the largest employer globally. Therefore, achieving farm productivity and sustainable agro-ecosystems in changing climatic conditions is a colossal task, requiring the adoption of sustainable agricultural practices and the latest technologies to increase efficiencies and productivity at reduced operational costs.

Global to national scale crop monitoring is required for informing policy formulation and their effective implementation. In contrast, local monitoring is required to optimise field-level decisions, mitigate production risks, and monitor land management practices. In essence, farmers need reliable, timeous, and accurate spatially explicit information (among others) on the occurrence, spatial distribution, and severity of climate-related phenomena such as droughts and other potential threats to productivity such as weeds, pests, and diseases (Mutanga et al., 2017; Alexandridis et al., 2020; Pritchard et al., 2022). More importantly, knowledge of the spatial and temporal variability of crop health and status can directly influence crop and land management decisions and improve the effectiveness of the remedial actions to crop stress and the productivity of agricultural systems. For example, temporal and spatial information is a critical input to precision agriculture to achieve better irrigation scheduling, variable rate fertiliser application and other farm inputs. Timely access to such information can mitigate crop failure, safeguard incomes, and jobs, and ensure food availability and environmental sustainability in the changing climatic conditions.

1.2. Remote sensing sensors for crop monitoring

The application of remote sensing for vegetation monitoring is well established. In agriculture, the utility of course (i.e., 250–1 000 m) to very high (i.e., <1 m) spatial resolution sensors and multi-spectral (i.e., broad bandwidth) to hyperspectral sensors (i.e., narrow bandwidth) have been demonstrated in various parts of the world (Main et al. 2011; Ramoelo et al. 2011; de Castro et al. 2012). For example, coarse resolution sensors, such as MODerate Resolution Spectroradiometer (MODIS), SPOT-VEGETATION (SPOT-VGT), and Advanced Very-High-Resolution Radiometer (AVHRR), have been highly exploited for large-scale agricultural monitoring and retrieval of biophysical and biochemical parameters

(Houborg et al., 2015a; Shoko & Mutanga, 2017). As a result, several operational global biophysical products exist and were extensively validated and inter-compared (Huete et al., 2010; Liu, Liu & Chen, 2012; Tong & He, 2013). At the other extreme, very-high-resolution (VHR) sensors such as Worldview and RapidEye have been applied for detailed (sub-field-level) assessments of biophysical and biochemical parameters, crop conditions, yield, and residues (Huang et al. 2017; Hively et al. 2019; Johansen et al. 2020). However, systematic assessment of crops using VHR sensors has been impeded by their high cost (Houborg et al., 2015a; Shoko & Mutanga, 2017).

Alternatively, freely available medium-resolution data, such as that from the Landsat programme, have promoted scientific advancements, development of innovative applications and decision-support systems (Wulder et al. 2019; Zhu et al. 2019). The Landsat programme has a long record of data (since 1972) at medium resolution (i.e., <30 m) and moderate temporal revisits (i.e., 16 days) at no cost to users; thus it has supported a variety of land applications (Wulder et al. 2008; Loveland & Dwyer 2012). Recently, the launch of new sensors, i.e., Operational Land Imager (OLI), OLI-2, and Multi-spectral Imager (MSI) onboard Landsat-8, Landsat 9 and Sentinel-2 satellites, respectively, with similar spectral configurations to previous Landsat sensors (i.e., Landsat Thematic Mapper [TM] and Enhanced TM [ETM+]), guarantees data continuity (Micijevic et al., 2019; Szantoi & Strobl, 2019). The latter also provides better spatial resolutions up to 10 m, temporal revisits up to 5 days, and additional spectral bands. Until now, the temporal resolution has been a major limitation for the application of medium-resolution data for time-sensitive agricultural decisions.

1.3. Drivers of vegetation reflectance properties

Generally, optical sensors provide image data in several electromagnetic spectrum regions, i.e., 350–449 nm (blue), 450–549 nm (green), 550–649 nm (red), 650–749 nm (red-edge [RE]), 750–1 299 nm (near-infrared, NIR) and 1 300–2 500 nm (shortwave infrared, SWIR), which are sensitive to various properties of the Earth's targets. Thus, various targets have varying reflective properties related to their physical and chemical compositions, which allow their distinction and quantification in the spectral domain. Accordingly, the reflective properties of vegetation are determined by their biophysical and biochemical properties (Jensen, 1983; Jensen et al., 2007). For example, the visible region (VIS) is characterised by intense absorption in the blue and red bands as a result of leaf pigments such as carotenoids and chlorophyll (Blackburn, 1998). The NIR region is characterised by high reflectance as a result of the canopy structure, water content, and spongy mesophyll cells (Schmidt & Skidmore, 2003; Jensen et al., 2007). Consequently, spectral bands in the visible-near-infrared region (VNIR) have become a standard in most satellite sensors (regardless of their spatial resolution). Hence, they are commonly used in vegetation studies to calculate yield, biomass, and nitrogen (N) content (Bajwa et al., 2010; Bolton & Friedl, 2013; Ramoelo et al., 2015a; Battude et al., 2016).

Recently, the transitional zone, known as the RE, marked by a sharp reflectance increase from the intense absorption by chlorophyll pigments to a high reflection by foliage, has been shown to enhance biophysical and biochemical prediction accuracy (Trishchenko et al., 2002; Mutanga & Skidmore, 2007; Ramoelo et al., 2012; Verrelst et al., 2012; Sibanda et al., 2019). Indeed, this capability was enabled by commercial satellites such as RapidEye (the first satellite to incorporate RE band) and subsequently, by the Worldview series of satellites at spatial resolutions of <2 m (Houborg et al. 2015a). As Houborg et al. (2015a) note, the full potential of the RE region for vegetation analysis has not yet been fully explored owing to the exorbitant costs of commercial image data. Therefore, Sentinel-2 MSI’s addition of three new RE bands, with centres at 705 nm, 740 nm, and 783 nm, presents prospects for improving biophysical and biochemical parameter estimation accuracy for crop monitoring. In contrast, the subtle biochemical traits (including sugars, cellulose, carbohydrates, protein, and water content) influence the plant canopy reflectance in the SWIR (Curran, 2001). The importance of RE and SWIR bands is demonstrated by several studies (Adjorlolo et al., 2012; Clevers & Gitelson, 2013; Kganyago et al., 2017). For example, Kganyago et al. (2017) found that SWIR and RE band were selected by feature selection techniques in discrimination of weeds. Similarly, Verrelst et al. (2012) found better prediction accuracy in Fractional Vegetation Cover, Canopy Chlorophyll Content (CCC), and Leaf Area Index (LAI) when RE bands are incorporated. Contrarily, in estimating Leaf Chlorophyll Content (LC_{ab}), LAI, and CCC, Clevers et al. (2017) showed that RE spectral vegetation indices (*SVIs*) were inferior to their VNIR equivalents owing to relatively lower (20 m) spatial resolution of the MSI's RE bands. In another study, Herman et al. (2010), claim that there is a comparable relationship between vegetation characteristics in the VNIR and SWIR regions. These contradictions are areas of active research today and are addressed in Chapter 7 of this thesis.

1.4. Remote sensing techniques for estimating crop biophysical and biochemical parameters

Remote sensing of biophysical and biochemical parameters has been topical since the dawn of satellite-based remote sensing. These parameters are among the most crucial biophysical and biochemical parameters for crop monitoring, as they characterise crop development, productivity, and stress. Generally, methods for biophysical and biochemical parameter estimations can be categorised into physically based (i.e., inversions of radiative transfer models), empirically-based (i.e., linear relationships with spectral vegetation indices), and machine learning regression algorithms. Physically based Radiative Transfer Models (RTMs) such as PROSAIL are transferable over a diversity of cover types and environments (Darvishzadeh et al., 2008; Atzberger et al., 2015; Xu et al., 2019); however, they require substantial site-specific data (Verrelst et al., 2012). In contrast, SVIs are empirically related to many biophysical and biochemical vegetation traits such as N content and leaf pigments (Haboudane et al., 2008; Lee et al., 2008). Therefore, they are important for non-destructively estimating crop parameters such as above-ground biomass, LAI, and changes in vegetation cover (Tillack et al., 2014; Shen et al., 2009; Chen & Gillieson, 2009). When locally calibrated, the empirically-based methods are more accurate; however, they are not transferable.

Machine Learning Regression Algorithms (MLRAs), in contrast, are simple, fast, and robust to nonlinearity, small training samples, and noise. Consequently, various MLRAs have been widely applied for retrieving crop parameters (Verrelst et al., 2012; Sibanda et al., 2017; Houborg & McCabe, 2018; Shah et al., 2019). Moreover, the potential of deep learning algorithms such as

Neural Networks (NN) combined with RTMs has been demonstrated (Baret et al. 2013; Weiss & Baret 2016).

Therefore, with a range of sensors and machine learning algorithms, remote sensing should be continuously explored to improve the estimation of these essential crop parameters to inform nitrogen (N) and other field variability management, crop health, and status. Hence, it can aid agricultural intensification, productivity, and optimise yield in the changing climate by optimising farm inputs and management decisions. This can improve the profitability and efficiency of the agricultural system and reduce environmental impact.

1.5. Interaction of the atmosphere and reflected signal

Changing atmospheric conditions, during image acquisition has been a challenge since the dawn of satellite-based remote sensing. For example, opaque clouds render acquired images obsolete and may cause delays of more than five days. According to Atzberger (2013), farmers require information every five days or less to promptly implement agricultural management decisions. Furthermore, clouds obscure all the optical spectral bands of the electromagnetic spectrum; thus, they reduce the frequency of valid observations necessary for crop monitoring. Aerosols (smoke, dust, and atmospheric gases such as CO_2) increase the atmospheric backscattering signal while attenuating the surface directional reflectance signal (Hilker et al. 2009). Therefore, they undermine the radiometric integrity and consistency of the measured reflectance, which are critical for phenological analysis and reliable estimates of crop parameters. In particular, the variability of the atmosphere affects spectral response to crop conditions, causing radiometric inconsistencies. Wilson, Milton, and Nield (2014) show that Aerosol Optical Thickness (AOT) may cause errors of about 1.7% in the measured reflectance and 5% change in Normalised Difference Vegetation Index (NDVI). As an integral input to precision agriculture, such errors also manifest in remotely sensed biophysical and biochemical parameters, crop type maps, and false detection of stress and other crop conditions (Vanonckelen et al., 2013; López-Serrano et al., 2016). Consequently, atmospheric conditions have adverse implications for the quality of optical satellite images and radiometric consistency, hence downstream applications that support the production and sustainability efforts in agriculture.

Therefore, to avert this challenge, cloud screening, atmospheric correction (AC) and gap-filling techniques are used (Hagolle et al., 2010; Zhu et al., 2015; Adam et al., 2018; Julien & Sobrino, 2019). By masking out cloudy pixels and considering only the clear-sky pixels in partly cloudy images, cloud screening and gap-filling techniques seek to improve the number of viable observations for quantitative analysis. In contrast, AC techniques seek to remove or minimise the atmospheric effects of atmospheric gases and aerosols, to improve the radiometric consistency of measurements with many aerosol loadings. Since the launch of Sentinel-2, many techniques for atmospheric correction have been proposed (Lonjou et al., 2016; Louis et al., 2016; de Keukelaere et al., 2018; Frantz, 2019), and validated and inter-compared (Gascon et al., 2017; Djamai & Fernandes, 2018; Doxani et al. 2018; Sola et al., 2018; Vermote et al., 2019). However, studies (Doxani et al., 2018; de Keukelaere et al., 2018; Sola et al., 2018) report inconsistent results which vary by land cover, atmospheric conditions, spectral bands, environments, and sensors. Since there is no one universal technique, these inconsistencies in previous studies highlight the necessity for additional validation and inter-comparison. The performance of various AC techniques in semi-arid agricultural landscapes under various atmospheric conditions and the effect on biophysical and biochemical properties and yield prediction is an under-studied topic in remote sensing. Chapter 4 of this thesis addresses this aspect.

2. Problem Statement

Precision Agriculture (PA) promises optimised farm inputs (i.e., fertilisers, seeds, water, and pesticides), improved efficiency and profitability of the agricultural system compared to traditional agricultural management approaches. The latter are linked to the over-use of resources such as land, water, and fertilisers. PA technologies such as remote sensing provide spatiotemporal variability related to various aspects of agricultural production, thereby aiding site-specific farm management (SSM). This averts potential losses owing to biotic and abiotic stressors and reduces environmental impact by avoiding the excessive application of inputs. The greater adoption of digital technologies and PA principles is particularly needed in sub-Saharan Africa, where climate-related phenomena such as droughts and heatwaves have become frequent, prolonged, and severe. These phenomena result in extreme food shortages that hinder progress towards achieving local to global food security and SDGs. Remote sensing technologies can provide relevant and timely information required to successfully and affordably implement PA in developing countries such as South Africa. The knowledge of the temporal and spatial variability of crop health and status directly influences crop and land management decisions, the effectiveness of remedial actions, and the productivity of agricultural systems. For example, this information is crucial for better irrigation scheduling, and management of fertiliser applications and other farm inputs using PA technologies and principles to avoid crop failure, safeguard incomes and jobs, and ensure food availability and environmental sustainability in the changing climatic conditions.

To optimise farming practices and improve crop productivity, PA relies on remotely sensed leaf and canopy parameters related to crop growth, health, and yield. Over the years, many satellite sensors and crop parameter retrieval methods with varying sophistication have been developed and tested. However, the launch of Sentinel-2 is a ground-breaking event, providing (for the first time) freely available high-resolution data with strategically positioned spectral bands (i.e., red-edge bands, critical for accurate vegetation assessments). Meanwhile, retrieval techniques have advanced from statistical parametric techniques such as *SVIs* to more sophisticated physically based techniques which use RTMs, while the MLRAs have brought prospects for integrating a variety of remotely sensed, climatic, and environmental variables in the retrieval of essential crop biochemical and biophysical parameters such as Leaf Chlorophyll *a* and *b* Content (LC_{ab}), CCC and LAI (Verrelst et al., 2013; Fernandes et al., 2014; Shah et al., 2019; Yan et al., 2019). Nonetheless, it remains challenging to accurately estimate these parameters from optical sensors such as Sentinel-2 owing to the following scientific challenges.

Challenge 1. The quality of optical images is compromised by the aerosols (i.e., aerosol optical depth and atmospheric water content) and clouds which cause atmospheric signal attenuation. Despite the availability of many approaches for correcting atmospheric effects, their effectiveness has undergone only limited exploration. Moreover, the impact of residual errors after atmospheric correction (AC) has not been extensively explored, especially as this relates to the radiometric integrity of Sentinel-2 images and subsequent influence on crop biophysical and biochemical parameters relevant for PA, i.e., LAI, CCC, and LC_{ab}, is poorly understood in semi-arid agricultural landscapes. Addressing this issue is crucial for ensuring reliable and accurate crop biophysical and biochemical parameter retrievals.

Challenge 2. Previous studies have shown that MLRAs coupled with physically based RTMs can be used to develop generic and crop-specific tools for retrieving crop biochemical and biophysical parameters over various crop types, growth stages, and climatic conditions (Fernandes et al. 2014; Shah et al. 2019). However, the spatial transferability of such models is seldom assessed, especially in semi-arid sub-Saharan study areas, which are characterised by mixed-cropping systems. Spatially transferable models are essential for practical PA applications and reducing reliance on experimental data, which come with high financial and time costs.

Challenge 3. With many MLRAs to choose from comes uncertainty about whether a single algorithm can perform equally well in various environments, sensor configurations, and farm configurations. While some MLRAs, such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM), performed better in certain conditions, they are also criticised for the associated high computational cost, complexity, and many confounding hyperparameters. In contrast, others, such as Random Forest (RF) and Gaussian Process Regression (GPR), also provide superior estimation accuracies and lower computational costs with only a few intuitive hyperparameters. Nonetheless, the literature shows that there is no universal algorithm that is suitable for all contexts. Moreover, it is essential to evaluate various MLRAs to determine their sensitivity to the different sites and sensor configurations, especially because of rapid developments in MLRAs and sensor capabilities. Therefore, there is a need to explore and evaluate state-of-the-art MLRAs that can effectively retrieve crop biochemical and biophysical parameters and identify the most influential variables towards understanding the performance of different algorithms under varying crop conditions and types, environments, and spectral and spatial configurations of sensors.

Challenge 4. The advent of quasi-hyperspectral sensors such as Sentinel-2, characterised by traditional VNIR spectral bands and strategically positioned, narrow red-edge bands at 10 m and 20 m spatial resolutions, respectively, sparks many interesting research questions such as (1) what is the contribution of the red-edge bands in the retrieval of crop biochemical and biophysical parameters? (2) Can we use the VNIR and RE-SWIR bands separately without compromising the accuracy of the retrievals? (3) Are the band positions and widths more important than the spatial resolution? Some of these questions have been explored in previous studies utilising simulated data. However, there is no consensus, and tests with the actual data have not yet been performed. In fact, some studies have found that Sentinel-2's VNIR bands are adequate for accurately retrieving crop parameters, while others argue that excluding the RE bands leads to systematic errors in retrievals. Understanding the strengths and limitations of different spectral configurations and their impact on estimating relevant crop parameters using MLRAs is necessary for effective BV retrieval. It is also questionable whether the feature selection is required to optimise the Sentinel-2 feature space for specific crop parameters. This doubt is brought by the findings of the previous studies indicating that only four to nine bands are sufficient to achieve robust BV retrievals. In this case, feature selection can reduce the computational burden, and reduce redundancy and dimensionality, especially considering many possible covariates that can be computed, such as SVIs, surface reflectance, and textural measures. Therefore, identifying the most relevant features can help streamline the BV retrieval process, reduce systematic errors, and improve overall accuracy.

Challenge 5. Apart from the atmospheric effect (highlighted in Challenge 1), measurements from remote sensing satellites are also affected by acquisition conditions (i.e., view and illumination geometry), which cause non-negligible Bidirectional Reflectance Distribution Function (BRDF) effects, introducing ambiguity in retrieved BVs. Sensors applied widely in remote sensing of biophysical and biochemical parameters of vegetation, such as Landsat and Sentinel-2, exhibit within-scene angular effects. However, the following questions have not yet been answered in the literature: (1) How much canopy reflectance variability is explained by acquisition geometry? and (2) What is the impact of incorporating per-pixel geometric variables as covariates to common radiometric variables in estimating crop BVs using MLRAs?

3. Aims and Objectives

The overarching aim of the study is to characterise the field variability in crop biophysical and biochemical parameters relevant for Precision Agriculture using Sentinel-2 data and state-of-the-art machine learning regression algorithms.

The specific objectives of the study are:

- To assess the accuracy of pre-trained hybrid models based on Radiative Transfer Modelling and Artificial Neural Networks in retrieving crop biophysical and biochemical parameters from Sentinel-2 data;
- To evaluate the performance of the atmospheric correction approaches for improving the radiometric integrity of Sentinel-2 MSI images and their effect on biophysical and biochemical parameter estimation in semi-arid agricultural landscapes;
- To evaluate the capabilities of various explainable state-of-the-art machine learning regression algorithms in retrieving essential crop biophysical and biochemical parameters for precision farming and identifying the most influential spectral bands;
- To develop spatially transferable (i.e., portable) machine learning models for the retrieval of crop biophysical and biochemical parameters relevant for precision farming;
- To examine the strength of Sentinel-2 spectral configurations (varied by their spatial resolutions) for estimating relevant crop biophysical and biochemical parameters of agricultural crops using tree-based and kernel-based machine learning algorithms;
- To evaluate the prospects of incorporating per-pixel view and illumination geometries as additional covariates to improve the accuracy of retrieving crop parameters; and
- To optimise the retrieval accuracy of biophysical and biochemical parameters of crop using filter and wrapper-based feature selection techniques.

4. Thesis Structure and Chapter Sequence

This section describes the structure of the thesis **Figure 1-1**. Each chapter addresses a specific objective towards the overall aim of the study and represents the stand-alone original peer-reviewed manuscripts or conference proceedings. The chapters are classified into broad themes, i.e.: (1) General overview and contextualisation; (2) Existing models for retrieving biophysical and biochemical parameters of agricultural crops and image radiometric quality; (3) Exploitation of machine learning algorithms for crop biophysical and biochemical retrieval; (4) Prospects of optimising Sentinel-2 covariates for improved crop parameters, and (5) Synthesis and conclusion.

4.1. Overview and contextualisation

Chapter 1. Introduction

Chapter 1 provides an overview of the study, the motivation, and the aims and objectives of the study.

Chapter 2. Remote Sensing of Crop Biophysical and Biochemical Parameters: An Overview of Advances in Sensor Technologies and Retrieval Techniques for Precision Agriculture

This chapter provides an overview of recent advances in remotely sensed retrieval of crop parameters, brought by developments in sensor technologies and robust machine learning algorithms. Moreover, technical challenges in retrieving and estimating crop parameters are identified, as well as the gaps in the literature.

4.2. Image radiometric quality and existing models for crop biophysical and biochemical retrieval

Chapter 3. Validation and Comparison of Sentinel-2 Biophysical Parameters from SNAP Toolbox in an African Semi-Arid Agricultural Landscape

This chapter seeks to understand the effect of atmospheric correction residual errors on the retrieval accuracy of LAI. In that regard, a pre-trained ANN model with a physically based RTM is used to retrieve LAI using atmospherically corrected (surface reflectance) and uncorrected (Top-of-Atmosphere) data, and the results are validated against field data and compared with well-established global products from PROBA-V and MODerate Resolution Imaging Spectroradiometer (MODIS).

Chapter 4. **Validation and Inter-comparison of Atmospheric Correction Approaches for Sentinel-2 Multi-Spectral Imager (MSI) Over a Semi-arid Agricultural Landscape**

Chapter 4 seeks to characterise errors in surface reflectance of Sentinel-2 MSI data derived from Sen2Cor and *i*Cor atmospheric correction tools and to assess their impact on the estimation of agricultural crop parameters using sparse Partial Least Squares (sPLS). Analysis is also performed on uncorrected Sentinel-2 data (Top-of-Atmosphere) for comparative purposes.

4.3. Exploitation of machine learning algorithms for crop biophysical and biochemical retrieval

Chapter 5. Estimating Crop Biophysical Parameters Using Machine Learning Algorithms and Sentinel-2 Imagery

Considering the results and recommendation from Chapters 3 and 4, this chapter selects a few state-of-the-art algorithms (i.e., Random Forest, Gradient Boosting Machines, eXtreme Gradient Boosting, and sparse Partial Least Squares) and evaluates them for their predictive capabilities and ability to find influential variables from Sentinel-2 MSI data.

Chapter 6. **Exploring Transferable Techniques to Retrieve Crop Biophysical and Biochemical Variables Using Sentinel-2 Data**

One of the critical limitations in remote sensing of crop parameters using machine learning algorithms is the cost of collecting training data and the lack of transferability. This chapter explores the prospects of spatially transferable machine learning models for retrieving relevant crop parameters for Precision Agriculture by introducing a novel framework which considers the requirement of a few training samples in the destination sites to improve spatial transferability.

4.4. Prospects of optimising Sentinel-2 covariates for improved crop biophysical and biochemical parameters

Chapter 7. Testing Sentinel-2 Configurations for Crop Biophysical and Biochemical Variables Retrieval Using State-of-the-Art Machine Learning Algorithms

This chapter considers the prospects of varying Sentinel-2 configurations (i.e., S2-10 m and S2-20 m) for retrieving crop parameters and its impact on the accuracy of three robust machine learning regression algorithms. The results are benchmarked against the common practice of using all spectral bands as covariates.

Chapter 8. Evaluating the Prospects of Incorporating Per-Pixel View and Illumination Geometry for Improving the Accuracy of Retrieving Crop Biophysical and Biochemical Parameters

Wide field-of-view sensors such as Sentinel-2 exhibit per-pixel view and illumination geometry variation, which can influence the retrieval accuracy of essential crop biochemical and biophysical variables for precision farming and crop monitoring. However, this aspect has seldom received attention in the existing literature. Therefore, this chapter evaluates the contribution of view and illumination geometries to the accuracy of retrieving LC_{ab}, CCC, and LAI using an explainable machine learning algorithm, i.e., Random Forest.

Chapter 9. Optimising Feature Space for Improved Crop Biophysical and Biochemical Variables Retrieval Using the Novel Spectral Triad Feature Selection Algorithm on Sentinel-2 Covariates

The launch of quasi-hyperspectral sensors, e.g., Sentinel-2, presents new challenges, where two or more variables may be collinear and affect the performance of machine learning algorithms. To optimise the Sentinel-2 feature space for improving the computation times and retrieval accuracy of crop parameters, this chapter presents a novel feature subset selection technique, capable of application as a filter or wrapper and was evaluated against a case where no feature selection is applied and a well-established rigorous technique, i.e., Recursive Feature Elimination.

4.5. Synthesis

Chapter 10. Synthesis and Conclusions

The final chapter synthesises the results and draws the main conclusions from each chapter, discusses critical findings, and makes necessary conclusions and recommendations for future studies.

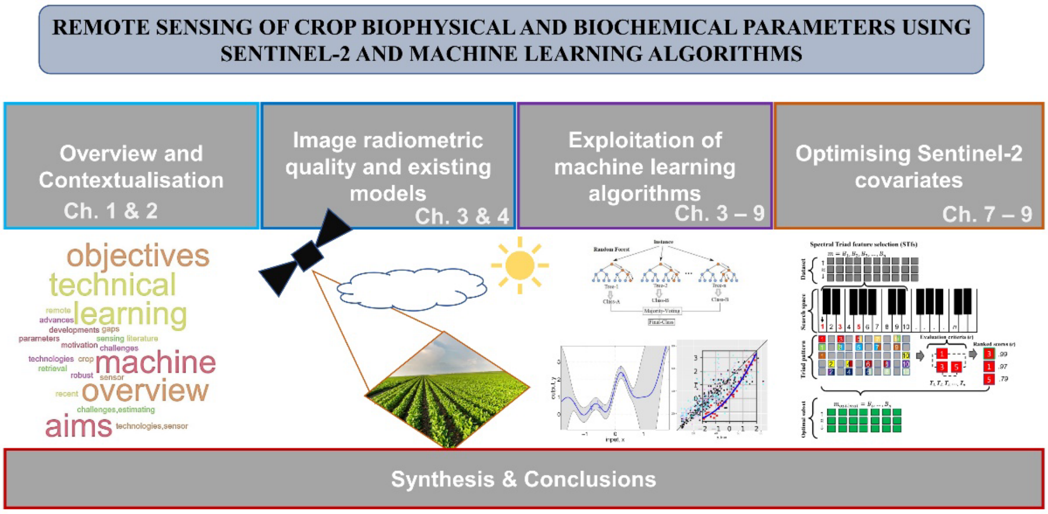


Figure 1-1. Thesis structure.

CHAPTER 2

REMOTE SENSING OF CROP BIOPHYSICAL AND BIOCHEMICAL PARAMETERS: AN OVERVIEW OF ADVANCES IN SENSOR TECHNOLOGIES AND RETRIEVAL TECHNIQUES FOR PRECISION AGRICULTURE

This chapter is based on the following manuscript:

Kganyago, M., Adjorlolo, C., Mhangara, P., & Lesiba Tsoeleng (**Under-Review**). Remote Sensing of Crop Biophysical and Biochemical Parameters: An Overview of Advances in Sensor Technologies and Retrieval Techniques for Precision Agriculture. *Computers and Electronics in Agriculture. [Manuscript ID: COMPAG-D-22-03627].*

Abstract

This paper provides an overview of the recent developments in remote sensing technology and machine learning algorithms for estimating important biophysical and biochemical parameters for precision farming. The objectives are (1) to provide an overview of recent advances in remotely sensed retrieval of crop parameters brought by the developments in sensor technologies and robust machine learning algorithms; and (2) to identify technical challenges in retrieving and estimating biophysical and biochemical parameters and implications for precision farming. The review revealed that developments in crop parameter retrieval techniques are mainly driven by announcements and the availability of new sensors. For example, the announcement of the Sentinel-2 constellation triggered the testing of its spectral configurations and upscaling of retrieval approaches using simulated data from field spectrometers and airborne hyperspectral sensors. In particular, studies showed that the red-edge bands, available in sensors such as RapidEye, Worldview-2, and Sentinel-2, and hybrid models consisting of Radiative Transfer Models and machine learning algorithms are promising for operational and accurate monitoring of stress-related crop parameters to aid time-sensitive agricultural decisions. However, many models are tested in Mediterranean climates and perform poorly in African semi-arid areas. Thus, this calls for intensified research aimed at developing new (or adapting existing) models to such areas. The lack of permanent experimental sites and systematic calibration data for many crops are some of the factors limiting the use of remote sensing technologies for precision farming in sub-Saharan Africa. Other complexities arise from farm configurations, such as small field sizes and mixed-cropping practices. These make it challenging to realise the economic benefits of drone technologies, which promise greater agility and flexibility. Nonetheless, the high cost of sensor payloads, poor rural connectivity, and strict regulations also prohibit their adoption in developing countries. On the positive side, cube satellite missions such as SuperDoves will likely reduce the cost of very high-resolution data while providing unprecedented capabilities for detailed, accurate, and frequent characterisation of field variability.

1. Introduction

Optimising food production systems while ensuring minimal environmental impacts is paramount to achieving local to global food security, climate change resilience, and sustainability goals. The significance of achieving food security is enshrined in Sustainable Development Goals (SDGs) of the United Nations (UN and Agenda 2063 of the African Union (Anderson et al., 2017; Kganyago & Mhangara, 2019). For example, SDG 2 (Zero Hunger) Target 2.4 aims to ensure sustainable food production systems and implement agricultural practices that: (1) "increase productivity and production" (2) maintain "the integrity of ecosystems" (3) "strengthen the capability to adapt to climate change, extreme weather events, drought, flooding, and other disasters" and (4) "improve land and soil quality by 2030" (UN, 2015). The Food and Agriculture Organisation (FAO) report titled "The State of Food Security and Nutrition in World 2022" reported that in 2021, 20.2% of Africa's population faced hunger, the highest percentage compared to other regions, with Asia experiencing 8%, and less than 2.5% in northern America and Europe.

Moreover, the COVID-19 pandemic has reversed the progress towards achieving food security imperatives, especially in sub-Saharan Africa, considering its linkage to the ongoing challenges of extreme poverty, high malnutrition levels, inequality, slow economic growth, and high unemployment rates. For food security, a significant challenge is that traditional agricultural practices hold low prospects for eradicating hunger in sub-Saharan Africa owing to the over-usage of resources such as land, water, and fertilisers. Also, conventional approaches to food production are susceptible to changing climatic and environmental conditions, which undermine short- and long-term productivity of agro-food systems. Therefore, addressing food security requires integrative agricultural management approaches and greater adoption of digital technologies, improving productivity, and resilience to climate change and shocks caused by pandemics (e.g., COVID-19).

Precision Agriculture (PA)—a cutting-edge management technique that makes use of many technologies and standards for managing spatiotemporal variability at field-scale to enhance agricultural production (Mulla, 2013, Peralta et al., 2015)—promises the optimisation of farm inputs (i.e., fertilisers, seeds, water, and chemicals), improved efficiency and profitability of the agricultural system by averting potential losses over stressed areas, and reduced environmental impact by avoiding the excessive application of inputs. As per the definition, PA can be accomplished by deploying various sensors and application technologies, mostly mounted on terrestrial and aerial systems. Consequently, many studies report the benefits of integrating specific PA technologies in farm operations. For example, using a real-time eXtend (RTX) automatically guided tractor for peanut sowing and digging operations, dos Santos et al. (2019) found higher precision and a reduction in overlaps, i.e., by <38 mm than when using a manually guided tractor. In another study, Bellvert et al. (2020) found cost savings of €7 090 and €9 960 over the 2016 and 2017 seasons, respectively, when using optimised precision irrigation based on an integrated vine water consumption model and remote sensing data in a commercial vineyard of 100 ha. Apart from the reported benefits, PA faces several limitations, such as information accuracy, high data volumes, complex information to be understood by farmers, and initial implementation costs. The issue of information accuracy poses a significant risk to the success of PA and controls the perception of farmers regarding the benefits of PA technologies—and thus, its adoption.

Moreover, high data volumes generated by PA technologies may be difficult to store and rapidly analyse to provide timely field variability information. Although this may be averted by adopting big data analytics on cloud computing platforms, interpretation of the complex information from PA services and their cost vis-à-vis benefits to the farmers may be another critical limitation. In sub-Saharan Africa, low education levels among the farming community, poor rural connectivity, and small field sizes threaten the feasibility of PA. Nonetheless, PA remains one of the few viable solutions in the era of climate variability and change, declining agricultural resources owing to poor land management and competition from other land uses, and imperatives to address food insecurity, among others.

Inarguably, optimising farm inputs such as water, fertilisers, and pesticides using Variable Rate Application (VRA) systems hinges on accurate, reliable, timely, and spatially explicit information on the field variability of various crop, weather, and soil parameters. While the variability of weather and soil parameters is often accomplished through in situ sensors such as weather stations and soil moisture sensors, crop parameters can be measured with proximal and remote sensors. Remotely sensed leaf and canopy parameters related to crop growth, health, and yield are appealing. Compared to traditional in situ sensors and destructive sampling, they can be acquired non-destructively and repeatedly over vast areas and in great detail. Destructive sampling methods require removing plant materials and subsequent laboratory-based analysis that may be time-consuming and costly. At the same time, in situ sensors are point-based, challenging to maintain, and must be closely or densely spaced to sufficiently cover large areas. Moreover, remote sensing systems are versatile and flexible, with the capability to mount advanced imaging and non-imaging spectroradiometers deployable on airborne (or aerial systems) and space-borne (or satellite) systems.

More specifically, the spectroradiometers deployed on unmanned spacecraft have evolved rapidly from low spatial resolution-high temporal resolution sensors. For example, Advanced Very High-Resolution Radiometer (AVHRR) and MODerate Resolution Imaging Spectroradiometer (MODIS) were available in the 1970s and 2000s, respectively, and provided coarse spatial resolution-high temporal resolution data which is relevant for characterising physiological and agroecological processes at ecosystem, biome, and biosphere scales. At the same time, medium spatial resolution-low temporal resolution sensors such as Landsat Thematic Mapper (TM) and Enhanced TM (ETM+) provided 30 m spatial resolution data at a 16-day repeat frequency, relevant for characterising plant processes at community and ecosystem scales. Recently, the Landsat Operational Land Imager (OLI) and Sentinel-2 missions have ensured the continuity of remotely sensed data while also providing improved spatial and temporal resolutions of up to ten metres and five days, respectively, which are relevant for PA. Moreover, thanks to commercial constellations such as PlanetScope and the side-looking capabilities of the Worldview series, very high-resolution data, i.e., >0.3 m and <5 m, can be acquired more frequently, i.e., <5 days.

The availability of remotely sensed data alone is insufficient, as they are complex and do not readily provide desired information for PA. Therefore, techniques of varying sophistication are required to extract relevant crop health and growth information for farm-level management and decisions. These techniques have evolved in tandem with the developments in sensor technology and access to data from such new sensors. For example, the earlier techniques for retrieving essential plant elements consisted of accentuating the sensitivity of spectrally limited remotely sensed data (i.e., with few visible: 450–690 nm and near-infrared: 760–900 nm regions) towards a specific biochemical and biophysical trait, e.g., plant vigour, using statistical parametric techniques such as spectral vegetation indices (*SVIs*). Over time, the *SVIs* have evolved to incorporate shortwave infrared (SWIR) and the narrow red-edge: 700–730 nm (RE) bands with the advent of QuickBird, IKONOS, and Worldview sensors.

Because of the limitations of SVIs, (such as saturation and lack of transferability), more sophisticated physically based techniques exist and are relatively transferable compared to SVIs. These techniques use Radiative Transfer Models (RTMs) to simulate light absorption and scattering properties of plant leaves and canopies based on various leaf biochemical content (e.g. nitrogen and chlorophyll content) and canopy structural properties (e.g., Leaf Area Index) combined with soil and background spectra and sun-view geometry (Jacquemoud et al., 2009). Meanwhile, advances in Machine Learning Regression Algorithms (MLRAs) have brought prospects for integrating a variety of remotely sensed, climatic, and environmental variables in the retrieval of important biochemical and biophysical parameters of agricultural crops such as Leaf Chlorophyll *a* and *b* Content (LC_{ab}), Canopy Chlorophyll Content (CCC) and Leaf Area Index (LAI). Coupled with RTMs, studies have shown that MLRAs can be used to develop generic and crop-specific tools for retrieving crop biochemical and biophysical parameters over various crop types, growth stages, and climatic conditions (Verrelst et al., 2013; Fernandes et al., 2014; Juutinen et al., 2017; CUI et al., 2019; Shah et al., 2019; Yan et al., 2019).

Despite the advances mentioned above in retrieving crop biochemical and biophysical parameters from remotely sensed data, several technical challenges still exist, such as residual errors from atmospheric correction, parameterisation of RTMs, and sampling and field measurement errors. However, the latest developments in sensor technology and MLRAs for biophysical and biophysical parameter retrieval in the context of PA have not been comprehensively reviewed in the last decade. To this end, we assimilate and review the literature from the past decade (2011–2020) to showcase recent developments in sensor technology and machine learning algorithms applications for biophysical and biophysical parameter retrieval in agricultural landscapes. We also assess the technical challenges and emerging trends with the potential to support PA.

The objectives of this paper were two-fold. First, to provide an overview of recent advances in remotely sensed retrieval of crop biochemical and biophysical parameters brought by the development of sensor technologies and robust machine learning algorithms. Second, to identify technical challenges in retrieving biophysical and biochemical parameters. Furthermore, we discuss the emerging trends in data acquisition and retrieval techniques. The scope of this review is limited to the retrieval of LAI and chlorophyll content using machine learning algorithms over herbaceous croplands and grasslands. The literature search was performed using relevant keywords on the Scopus Abstract and Citation Database (www.scopus.com, accessed 26 February 2021) to select "Published" and "In Press" journal articles from 2011 to 2020 (see Appendix 2-1. Literature Search and Selection for details).

2. Remote Sensing of Essential Biophysical and Biochemical Parameters for Precision Agriculture

Remote sensing of biophysical and biochemical parameters, such as LAI and chlorophyll content, has been topical since the dawn of satellite-based remote sensing. The LAI is inarguably one of the most studied parameters owing to its importance in characterising vegetation, representing the energy exchange between the plant canopies and the atmosphere, and related to the productivity of plants. LAI (measured as $m^2\ m^{-2}$) is the ratio of photosynthetically active (PV) and non-PV (NPV) leaf area to ground area. The PV component is called the green LAI (LAI_G), while the NPV component is called the brown LAI (LAI_B). Among the two, LAI_G tends to receive significant attention from researchers of various disciplines and satellite data providers based on its importance for ecological, climate, and crop yield modelling (Zaroug et al., 2013). Consequently, several operational LAI_G, based on low-resolution sensors exist, and have been extensively evaluated and validated (Fang, Wei & Liang 2012; Sun et al. 2013; Fang et al. 2019; Souza et al. 2019) and utilised for various crop monitoring (Campos-Taberner et al., 2018), drought assessments (Zhang et al., 2020), and other applications (Rasul et al., 2020). In contrast, only a handful of studies attempted to characterise LAI_B (Delegido et al., 2015; Amin et al., 2021) in the last decade.

In PA, characterising the spatiotemporal variability of LAI_G at the field level is critical for monitoring the crop physiological development and phenological status of both erect (i.e., erectophile) and non-erect (i.e., planophile) canopies during the vegetative stages of the crop. Moreover, it can be used to model crop biomass and yield, thus providing essential insights into the productivity of the fields. Despite being limitedly studied, LAI_B is critical for detecting and evaluating the impact of heatwaves and droughts, determining the extent of conservation agriculture (or soil management) practices, and modelling fire risks of senescent crops (Erisman et al., 2008; Delegido et al., 2015; Amin et al., 2021). To avoid coverage and cost limitations of destructive lab-based methods, LAI measurements can be acquired with optical field-based instruments such as DEMON, Ceptometer (Accu-Par, Decagon Devices, Pullman, WA, USA), smartphones (Pocket-LAI), digital hemispherical photography, Pastis57, multi-band-vegetation imager (MVI), TRAC (Tracing Radiation and Architecture of Canopies), and LI-COR Plant Canopy Analyzer (LI-COR, Inc., Lincoln, NE, USA). These instruments provide an effective Plant Area Index (PAIe) as they cannot distinguish leaves from other plant tissue (e.g., branches and stems). Moreover, errors based on the non-random distribution of canopy foliage, radiation interception from other plant components, and gap fraction saturation as the LAI approaches values of 5–6 m^2 m^{-2} (Gower et al., 1999; Weiss et al., 2004) are carried forward to the retrieval methods—and subsequent remotely sensed LAI retrievals. Nevertheless, these instruments are scientifically accepted owing to their ease of use, portability, affordability, and consistency.

Leaf chlorophyll *a* and *b* content (LC_{ab}) is a critical crop biochemical parameter for monitoring crop health status, stress, and gross primary productivity (Gitelson, Peng & Huemmrich 2014). Leaf chlorophyll content (LC_{ab}) is significantly correlated to nitrogen, i.e., the most limiting nutrient in plants (along with water availability), which is critical for crop growth, health, and yield (Gitelson et al., 2003). Chlorophyll, inert structural elements in cell tissue, and the carbon-fixing enzyme ribulose biphosphate carboxylase (RuBisCo) are all products of nitrogen. Direct estimation of nitrogen content through lab-based methods is destructive, laborious, and costly, thus remotely sensed proxies such as LC_{ab} are crucial for optimising nitrogen fertilisation and rapid assessments of crop nitrogen status in Precision Agriculture (Jia et al., 2013; Tian et al., 2013; Vincini et al., 2016). In contrast, CCC—obtained as a product of LAI_G and LC_{ab}—is closely related to LAI (Gitelson et al., 2005; Boegh et al., 2013), for example, with R^2 exceeding 85% for maize and barley (Ciganda et al., 2008).

Generally, there is a lack of consensus in the literature regarding the relationship between CCC and nitrogen content. For example, Baret et al. (2007) showed that CCC applies to canopy N content assessment. Contrarily, Vincini et al. (2016) argued that the factors affecting LAI, such as stand density and water stress, make CCC inadequate for distinguishing nitrogen deficiency from other crop stressors. Meanwhile, the C_{ab}-sensitive SVIs such as the red-edge green Chlorophyll Index (CI_{green}), Chlorophyll Index ($CI_{red\text{-}edge}$), and MEdium Resolution Imaging Spectrometer (MERIS) Chlorophyll Index (MTCI) tend to be the best estimators of LAI, while the LAI-sensitive ones such as NDI45 and NDVI also tend to be the best linear estimators for C_{ab} (Viña et al., 2011; Frampton et al., 2013). This non-exclusive sensitivity to various biophysical and biochemical parameters is mainly owing to the interaction of various plant traits, which influences canopy spectral properties, thus making decoupling of such interacting parameters difficult (Verrelst et al., 2016; Delloye et al., 2018). Therefore, it is recommended to use LC_{ab} for precise nitrogen status management. In contrast, CCC can be used to assess crop development, productivity, and overall stress as it combines the effects of both C_{ab} pigments, plant structure (i.e., LAI), and other effects which act on these individual parameters.

3. Developments in Satellite-based Sensor Technologies for Biophysical and Biochemical Parameters Retrieval

3.1. Advances in remote sensing platforms

The developments in sensor technology are occurring rapidly, thus enabling the extraction of accurate and reliable actionable information promptly. In the 1970s, remote sensing systems consisted of sensors mounted on aerial platforms such as fixed-wing aeroplanes. Today, the low-Earth orbiting satellites and remotely piloted aircraft systems (i.e., RPAS, also commonly referred to as unmanned aerial vehicles, UAVs) are increasingly adopted for characterising field conditions. Inarguably, UAVs are the most powerful remote sensing platforms, with the capability to fly closest to the targets (i.e., at low altitudes), thus, offering super high spatial resolutions (<1 cm) and customisable spectral coverage and flexible revisit times (Zhang & Kovacs, 2012; Colomina & Molina, 2014; López-Granados et al., 2016). Recently, UAVs are becoming more accessible and affordable and provide essential tools to acquire data at field scales and at a relatively low cost compared to other platforms.

Hence, the availability of UAVs has since driven the miniaturisation of advanced sensor technology. For example, recent studies demonstrate the capability of UAVs for carrying a range of miniaturised sensor payloads and accurately characterising crop water stress variability and status (Zhou et al., 2018), NPK fertiliser deficiency (Corti et al., 2019), weed patch detection (Zisi et al., 2018), crop diseases mapping (Abdulridha et al., 2019), above-ground biomass of crops (Zheng et al., 2019), and yield estimation (Zhou et al., 2017). Compared with manned aerial (aircraft) and space (satellite) platforms, UAVs are substantially cheaper to operate and maintain, thus holding more possibilities for Precision Agriculture. Additionally, owing to their very low altitude, images collected by UAVs do not require atmospheric correction, which limits the adoption of satellite images in Precision Agriculture as they introduce a more considerable delay than is necessary to support timely farm management decisions (Atzberger, 2013).

Regrettably, the adoption of UAVs in developing regions such as sub-Saharan Africa is still limited, attributable to a myriad of complex factors, among others, affordability, smaller field sizes—characteristic of extensive small-holder farming (i.e., usually <0.5 ha)—such that economic benefits of UAVs cannot be realised. Further, there is a lack of awareness, insufficient technical and institutional capacity, and poor rural connectivity. Besides, the legislative restrictions (such as a requirement for a piloting licence, air service licence, and RPAS Operator Certificate in South Africa) for flying UAVs essentially reserves the technology to a few individuals and companies. Moreover, although the cost of UAVs may be low, the cost of the remote sensing payload may be prohibitive. This is likely to be compounded by the repeated imaging required for crop monitoring during the season. Intrinsically, UAVs are spatially limited based on limitations posed by battery capacity; thus, they may only address a few user imaging needs at a time.

Alternatively, satellite platforms have capabilities to remotely and repeatedly acquire data over large areas, in one pass, at medium (<30 m) to very high (<1 m) spatial resolution. They are therefore a convenient compromise with the added benefit of being more affordable than manned aerial platforms. Since the launch of the first Earth Resources Satellite (Landsat 1 MSS) in the 1970s, the utilisation of space for collecting Earth observation (EO) data has increased rapidly, with numerous low-Earth orbiting satellites launched. In the last decade, various low (e.g., Sentinel-3), medium-to-high (e.g., Sentinel-2), and very-high-resolution satellites (e.g., Worldview-3) were launched into space. At the same time, ground-based systems are required to acquire calibration and validation data under various field conditions and cover types. Using proximal and handheld sensors, information on the biophysical and biochemical properties of crops, such as LAI, chlorophyll content, and spectral data can be collected by non-invasive and non-destructive means. Today, myriad data collected with ground-based, aerial (manned aircraft), low-altitude UAVs, and space-based systems (i.e., satellites) are essential for cross-calibration, training, and validation of biophysical and biochemical retrieval models, and product development.

3.2. Advances in sensors

Remotely sensed optical sensors can be classified according to their spatial resolution (i.e., low- to very-high-spatial-resolution sensors), temporal resolutions (i.e., low- to high-temporal-resolution sensors), and spectral configurations (i.e., broadband or multi-spectral sensors, narrow-band, or hyperspectral sensors).

3.2.1. Multi-spectral sensors

In the design of sensors, inevitable trade-offs between spatial and temporal resolution are well-known. Medium-resolution (20 m–30 m) sensors such as Landsat TM (1984–2012) and Satellite Pour l'Observation de la Terre (SPOT 2 to 4, 1986–2013) are only capable of acquiring data over relatively small areas and in great detail, but less frequently (i.e., 16 days for TM and ~26 days for SPOT series). In contrast, low-resolution sensors could only acquire data in low spatial resolutions (250 m–1 000 m), but more frequently (i.e., <1–2 days). Examples include MERIS (2002–2012) onboard ENVISAT; AVHRR (1979–2019) onboard NOAA satellite; SPOT-VEGETATION (SPOT-VGT, 1998–2014) onboard SPOT-4 and -5; and MODIS (2001-to-date) onboard Terra and Aqua satellites. The products from the low-resolution data such as Enhanced Vegetation Index (EVI), NDVI, the fraction of Absorbed Photosynthetically Active Radiation (fAPAR) and LAI are well-established and available operationally. They are suitable for developing Essential Climate Variables (ECVs) for Global Climate Observing System (GCOS) (GCOS, 2011), monitoring the progress towards the SDGs of the United Nations (UN) (Anderson et al. 2017; Kganyago & Mhangara 2019a), and regional agricultural monitoring (Tong & He 2013). The availability of the Visible Infrared Imaging Radiometer Suite (VIIRS) and Sentinel-3 Ocean and Land Colour Instrument (OLCI) warrants the continuity of the long-term data acquired by AVHRR and MODIS instruments. However, Precision Agriculture applications require high spatial (<20 m) and high temporal resolution (<5 days) remotely sensed data. This is needed to characterise field conditions adequately and accurately throughout the phenological stages of the crops. Solitarily, the remotely sensed data from low- and medium-resolution multi-spectral sensors are either unreasonably coarse for obtaining relevant field-scale crop condition information or are unreasonably infrequent (exacerbated by cloud cover) to meaningfully support site-specific farm management decisions. Therefore, the compromise between spatial and temporal resolution presents new research problems, and consequently the development of various data fusion algorithms.

The data fusion algorithms (consisting of the spatial and temporal adaptive reflectance fusion models and unmixing-based methods) leverage the high temporal resolution of low-resolution sensors and the spatial detail of medium- to high-resolution (<20 m) data. This generates daily or near-daily medium- or high-resolution synthetic reflectance data which are similar to the measured reflectance observations, with R^2 >0.85 (Wu et al., 2012). These methods have also been applied to increase the frequency of the *SVIs* and, thus, biophysical parameters such as LAI (Wu et al. 2015). Using the Spatial-Temporal Data Fusion Approach (STDFA), Wu et al. (2015) found that the daily LAI values generated from daily synthetic vegetation indices were better than the well-established MODIS LAI product when validated against the daily LAI measurements from the winter wheat field with $R^2 \approx 0.98$ and RMSE ≈ 0.15 m^2 m^{-2}. Therefore, one can argue that the development of data fusion algorithms reduces the shortcomings of both the low- and medium-resolution sensors while also improving their utility for Precision Agriculture by exploiting the good qualities of each—that is, higher repeat cycles (up to <1 day) and detailed coverage (<30 m). Since they have been comprehensively validated, provide long-term time series, and are based on well-established methods, products from low-resolution sensors are invaluable for the quality assessment of new estimation approaches using medium-to-high-resolution sensors. Moreover, data fusion algorithms are also useful for harmonising the spatial resolution of sensors offering multiple resolutions, such as MSI (Zhang et al., 2019) or between different sensors, i.e., OLI and MODIS (Liao et al., 2019).

Over the years, the advances in sensor technology such as tasking capabilities and satellite constellations have increased the feasibility of obtaining high spatial and high temporal resolution multi-spectral data through off-nadir viewing. The side-looking capabilities have existed since the launch of SPOT-1 in 1986, which had an oblique viewing capability of 27°, but a rather long repeat cycle (26 days). Significant improvements in both spatial and temporal resolutions were brought about with the commercial availability of the first very-high-resolution (VHR, <5 m), i.e., IKONOS, in 1999, which enhanced the state-of-the-art capabilities (at the time). It had a repeat cycle of about three days, a side-looking capability of up to 60° and four standard multi-spectral VNIR bands at a resolution of 4 m. Subsequent commercial VHR sensors such as QuickBird (launched 2001), GeoEye-1 (launched 2008), and Pleiades (launched 2009) generally provided similar spectral bands with varying spatial and temporal resolutions. In the context of biophysical and biochemical parameter retrieval, despite capturing bands (red and NIR) necessary for vegetation analysis using parametric statistical techniques, the VHR multi-spectral VNIR sensors are spectrally limited. Unfortunately, the retrieval of crop biophysical parameters using VNIR bands has been shown to suffer from saturation caused by chlorophyll absorption in the red band (see detailed discussion in Section 4).

Nonetheless, new high-resolution and VHR sensors launched in the last decade such as SPOT-6 and -7 (2013–current) and PlanetScope's Doves (2017–current) constellation continue to acquire data in the VNIR regions because the VNIR indices are widely used in various vegetation analyses. Therefore, these sensors increase the spatial and temporal resolutions; thus there is good utility of the data for site-specific farm management. Specifically, PlanetScope's large constellation of sensors (consisting of hundreds of CubeSats with 10 cm by 10 cm by 30 cm dimensions and providing imagery at 3 m spatial resolution daily) presents a significant milestone and unparalleled capabilities in EO satellite sensor technology and prospects for Precision Agriculture.

Another significant development in sensor technology is the availability of RapidEye (since 2008)—a VHR satellite constellation consisting of five identical multi-spectral sensors optimised for assessing and monitoring agricultural and natural vegetation. This marked a critical shift to using multi-spectral sensors in biophysical estimation. In addition to the conventional VNIR bands, the RapidEye sensors collected data (at a native spatial resolution of 6.5 m) in the red-edge band (690–730 nm), known to enhance the biophysical and biochemical parameter retrieval accuracy using various techniques (Ramoelo et al., 2012). These capabilities were subsequently sustained by Worldview-2, launched in 2009—the first VHR multi-spectral sensor with eight spectral bands at 2 m spatial resolution and ~1 day (on average) of revisiting time. Apart from the typical VNIR bands, Worldview-2 acquires four additional bands, namely, 400–450 nm (coastal blue), 585–625 nm (yellow), 705–745 nm (red-edge), and 860–1 040 nm (NIR-2). In the last decade, the VHR sensors such as RapidEye and Worldview-2 have been applied extensively for detailed (sub-field-level) assessments of biophysical parameters (Dong, Liu, et al., 2019).

Launched in 2014, the WorldView-3 multi-payload super-spectral sensor, provides greater detail (1.24 m) for similar multi-spectral bands to Worldview-2, as well as for eight new SWIR bands (3.7 m). WorldView-3 further has an average revisit time of <1 day. However, their prohibitive costs have impeded systematic and operational monitoring of crops using VHR sensors (Houborg et al., 2015b). Therefore, extensive archives are only available for big cities of the world, while agricultural areas have limited coverage, and data availability depends on historical customer orders over specific small areas. The third generation of PlanetScope's constellation, SuperDove, is likely to reduce the cost of VHR data, thus increasing their accessibility to researchers, product developers, governments, and farmers. Moreover, it will revolutionalise the EO satellite sensor technology for Precision Agriculture, offering five to eight spectral bands, which would include VNIR bands with two green bands (513–549 nm and 547–583 nm) and yellow (600–620 nm) and red-edge bands (697–713 nm). This will provide the capability for detailed, accurate, and frequent characterisation of field variability.

The advent in 2014 of the new generation Landsat sensor, OLI, with 30 m resolution and 16 days average revisit time and no cost to users. This ensures the continuity of the long record of the Landsat programme's data (since 1972) (Loveland & Dwyer, 2012). However, the announcement and subsequent launch in 2016 of the Sentinel-2 constellation, and Sentinel-2A and -2B (launched in 2017) carrying identical Multi-spectral Imager (MSI) cameras represented a ground-breaking event in the EO community. Sentinel-2 MSI not only guarantees data continuity and interoperability with previous missions such as Landsat, SPOT, and MERIS (Wu et al., 2019) but also provides better spatial resolutions (up to 10 m), 5-day temporal revisits, and additional spectral bands with improved spectral resolution at no cost to users. These bands, centred at 705 nm, 740 nm, and 783 nm, are located in the red-edge region and present prospects for operational and accurate monitoring of stress-related biophysical and biochemical parameters of agricultural crops to aid time-sensitive agricultural decisions and improve yields.

Consequently, several studies have exploited Sentinel-2 red-edge data for LAI, chlorophyll content, and nitrogen retrieval over various crops and environments. Before its launch, studies mainly demonstrated its potential using simulated data from leaf and canopy RTMs and hyperspectral sensors. Frampton et al. (2013) used experimental data from SEN3Exp (collected from Barrax site, Spain in 2009) and SicilyS2EVAL (collected from Sicily site, Italy in 2010) combined with synthetic Sentinel-2 MSI data generated from two airborne imaging spectrometers (Compact Airborne Spectrographic Imager [CASI], AISA Eagle, and PROSAIL model). Frampton et al. (2013) concluded that MSI red-edge bands improved correlation of the *SVIs* with LAI, LC_{ab} and CCC and averted saturation effects. Using the SPARC field data (collected from Barrax site, Spain in 2003) and Compact High-Resolution Imaging Spectrometry (CHRIS) data, Verrelst et al. (2012) evaluated the performance of the three MSI spectral configurations. These comprised S2-10 m (4 bands), S2-20 m (8 bands), and S2-60 m (10 bands), with MLRAs. Verrelst et al. (2012) concluded that LAI and Fractional Vegetation Cover (FVC) could be accurately estimated with S2-10 m configuration based on its high spatial resolution, while highest accuracies for LC_{ab} were achieved with red-edge bands. Post-launch studies such as Delloye et al. (2018) mostly confirm the findings of the previous studies, showing the contribution of red-egde bands to the accurate retrieval of LAI, LC_{ab} and CCC. These bands were found in particular to reduce uncertainties of the low and high values and provide better characterisation across various growth stages and farming practices. Meanwhile, Herrmann et al. (2011) found that in estimating LAI, Sentinel-2 MSI performs at a level equivalent to a field hyperspectral sensor, namely FieldSpec Pro FR spectrometer (Analytical Spectral Devices, USA).

3.2.2 Hyperspectral sensors

Hyperspectral data, characterised by hundreds of narrow (<10 nm) and contiguous spectral bands, are critical for accurately and reliably extracting the crop biophysical and biochemical parameters. Contrary to multi-spectral broadband sensors, hyperspectral sensors (i.e., imaging and non-imaging spectrometers) can detect minute variations in plant biochemical and biophysical traits that affect the reflected radiance. Because of this capability, hyperspectral sensors are critical for identifying significant spectral bands (or regions) where the various plant traits absorb electromagnetic energy (also called absorption features), simulating new multi-spectral sensor data, and detecting fine plant leaf traits. These absorption features are then used to design the *SVIs*, and input variables in machine learning models and other cases, informing the development of future multi-spectral sensors. Despite their recognised significance for agricultural and natural resources management applications, only a few space-based hyperspectral sensors or imaging spectrometers exist, with most being science demonstrator missions. These include sensors such as Hyperion onboard EO-1, CHRIS onboard PROBA-V, DLR Earth Sensing Imaging Spectrometer, EnMAP, and PRISMA, respectively. The capability of the technology is mainly demonstrated with airborne hyperspectral sensors such as Airborne Visible and Infrared Imaging Spectrometer (AVIRIS), HyMap, and CASI, and non-imaging field spectrometers which, while seemingly cheaper to operate relative to satellite systems, nevertheless also require sophisticated processing procedures and extensive storage.

Studies have also used physically based leaf and canopy RTMs to demonstrate the capability of upcoming optical sensors (Féret et al. 2017), design new spectral indices (Yi et al. 2014), and determine optimal spectral bands for retrieval of biophysical and biochemical parameters (Richter et al. 2012). RTMs simulate the interactions of solar radiation and plant biophysical and biochemical properties based on physical sound cause-effect relationships. Given a set of leaf and canopy vegetation traits (e.g., LAI, chlorophyll *a* and *b*, leaf water content, dead matter, and leaf angle distribution), acquisition conditions (e.g., view and illumination angles), soil background and environmental parameters, RTMs can model the full-range (i.e., 400 nm to 2 500 nm) spectral reflectance of vegetation at a leaf or canopy scales. When coupled with atmospheric RTMs, the Top-of-Atmosphere (TOA) reflectance data can be obtained.

By inverting leaf and canopy RTMs, the plant traits can be estimated based on the spectrum measured by space-borne, airborne, or ground-based sensors. In this regard, several RTMs are at our disposal but varying in complexity, the number of parameters, and physical principles. Some of these RTMs were previously reviewed by Blackburn (2007) and Ustin et al. (2009). Inarguably, PROSPECT+SAIL (also referred to as PROSAIL) is one the most utilised RTMs in vegetation studies (Jacquemoud et al., 1995; Jacquemoud et al., 2009). It was popularised by its wide availability, simplicity, fewer input parameters, comparative predictive power to complex RTMs, and relatively fast computation time. Moreover, RTMs are transferable over diverse cover types and environments (Atzberger et al., 2015). We refer interested readers to an exhaustive account by Jacquemoud et al. (2009) of the applications of this RTM in vegetation studies.

4. Machine Learning Regression Algorithms for Biophysical and Biochemical Variables Retrieval

Also called non-linear or non-parametric regression algorithms, MLRAs are robust to the non-linear functional dependence of the BVs and the spectral reflectance data. Furthermore, they are robust to small training samples and noisy data, and do not have normality assumptions when compared to linear algorithms such as Ordinary Least Squares Regression (OLS) and Partial Least Squares Regression (PLSR). MLRAs establish relationships with each predictor variable using all the available variations without making assumptions (Verrelst et al., 2012). Generally, MLRAs are typically classified into three categories, namely, tree-based (or tree ensembles), kernel-based, and deep learning (or Neural Networks), based on their architectural designs (Rivera-Caicedo et al., 2017). The MLRAs have been widely exploited for BV retrieval in recent studies (**Table 2-1**).

Tree-based MLRAs are less complicated, more intuitive, and more popular than kernel-based and deep learning algorithms. This family of algorithms have origins in the Classification and Regression Trees (CART) only rarely used in recent years following the advent of more robust and up-to-date variants such as Random Forest (RF) and Gradient Boosted Regression Trees (GBRT) (also known as Gradient Boosting Machines, [GBM]) (Friedman 2001). Chen and Guestrin (2016) proposed a more advanced, scalable, and sparsity-aware improvement on GBM, namely, eXtreme Gradient Boosting (XGBoost). The XGBoost algorithm aims to improve the implementation of GBM by: (1) handling missing values more efficiently; (2) constructing trees quickly and building vast models utilising parallel computing; and (3) utilising highly regularised formalisation and gradient boosted trees, thus avoiding overfitting. Therefore, XGBoost often outperforms other algorithms (Beltran, Valdez & Naval 2019).

However, RF still features prominently in the literature owing to good performance and relatively few required hyperparameters, with studies showing its robustness and high accuracy in predicting various BVs. For example, Li et al. (2017a) found an R^2 of 88% and an RMSE of 0.195 m^2 m^{-2} in retrieving grassland LAI Landsat TM+ and OLI. Another study (Tavakoli & Gebbers 2019) found R^2 of between 56% to 69% in N content of Wheat in Germany using images acquired from a digital camera. Overall, tree-based algorithms are appealing owing to their interpretability and comprehensibility. For example, these algorithms allow interrogation of the tree structure and variables used at each split, providing the importance (or relative influence) values for each explanatory variable. Therefore, they can be used to obtain novel insights, such as new functional relationships between the response and explanatory variables. Others (Shelestov et al. 2017) have used such variable importance measures for feature selection. The capability to interrogate the model structure is critical for troubleshooting the models to obtain robust BV retrievals from satellite images (Azodi, Tang & Shiu 2020).

In contrast, kernel-based and deep learning MLRAs are complex, computationally expensive, and opaque (or "black box"). One also has to parameterise several confounding hyperparameters, and there is no capability to directly compute variable importance measures and interrogate the inner workings of models. Nonetheless, these MLRAs have attracted significant attention in recent studies using simulated and actual data from various hyper- and multi-spectral sensors (**Table 2-1**). Support Vector Machines (SVM) is perhaps the most commonly used kernel-based algorithm in remote sensing applications and supports several kernels, with popular ones being Gaussian Radial Basis Function (RBF), Sigmoid, polynomial, and linear kernels. However, Gaussian Process Regression (GPR) and its variants, such as Variational Heteroscedastic Gaussian Processes, are popular in BV retrieval studies, offering better prospects owing to its high accuracy and a special capability of generating estimates of the uncertainty of the response variable. This enables evaluations of the reliability of BV retrieval for use in operational applications (Verrelst et al., 2013). Like tree-based algorithms, GPR provides insights into the relevant bands which indicate the relationship with BV. GPR also supports multiple kernels, the common ones being anisotropic squared exponential and scaled Gaussian kernels. The algorithm requires fewer hyperparameters, i.e., $\theta = \{v, \sigma_b, \sigma_n\}$, where (v, σ_b) refers to the signal—composed of the scaling factor v and the length-scale of the explanatory variables σ_b—and σ_n refers to the standard deviation of the estimated noise. Verrelst et al. (2012) compared the capabilities of a deep learning algorithm (ANN) and three kernel-based MLRAs (SVM, Kernel Ridge Regression [KRR] and GPR), in the retrieval of LAI, LC_{ab}, and FVC using simulated Sentinel-2 and Sentinel-3 data from CHRIS/PROBA data. Their study found that GPR had a relatively high processing speed (< 2 s) and produced the most accurate results for all the BVs considered (R^2 of 0.89–0.99) with LC_{ab} being the most accurate and within the Global Monitoring for Environmental and Security (GMES) limit of 10% accuracy. Others (Wen et al., 2018) found an accuracy of 85% and RMSE of 0.95 μg ml in estimating N in rice using hyperspectral data collected from an Unmanned Aerial System (UAS).

In contrast, ANN is the most commonly used deep learning algorithm. As its name suggests, the algorithm is inspired by neurological sciences. It consists of layered and interconnected structures of artificial neurons by weights or links. It consists of various hyperparameters, such as the number of hidden layers, weights, nodes per layer, learning rate, the shape of the nonlinearity, and regularisation parameters. The ANN structure is often optimised using a learning algorithm, such as computationally heavy Levenberg-Marquardt or a quicker optimisation algorithm, such as gradient back-propagation. Recent developments such as Recurrent Neural Networks, Convolutional Neural Networks, and Long Short-Term Memory present new prospects for improving crop parameter retrieval accuracy. These are less prone to error than other machine learning techniques and have been shown to perform optimally in various environments for a range of applications, e.g., yield prediction (Barbosa et al. 2020), LAI (Apolo-Apolo et al., 2020), chlorophyll content (Xiaoyan et al., 2020), leaf water content (Nasir et al., 2019), crop type mapping, and plant disease detection (Golhani et al., 2018).

Besides regression problems, MLRAs can also be applied to feature selection or dimensionality reduction; they are thus critical for reducing the dimensionality ($n < p$) and collinearity of predictors. Some MLRAs such as SVMs have been dubbed as “robust to dimensionality”, although recent studies have increasingly been showing that dimensionality reduction techniques help reduce uncertainties in BV retrieval. For example, Verger et al. (2011) found that feature selection benefitted ANN performance, with additional bands causing the noise. Hence, the optimal number of spectral bands for LAI retrieval was seven out of the 62 bands from CHRIS/PROBA. Meanwhile, Shelestov et al. (2017) found the same number of features were significant for LAI, fAPAR, and FCover retrieval using Landsat and SPOT-5 images collected as part of the SPOT-5 Take Five initiative. Other authors (Delegido et al., 2011; Verrelst et al., 2016) have shown that four to nine bands, selected through variable selection techniques, are sufficient to achieve robust retrievals of biophysical and biochemical parameters. The variable selection approaches are generally divided into filter-based (e.g., analysis of variance), wrapper-based (e.g., Recursive Feature Elimination-Support Vector Machine), and embedded (e.g., sparse Partial Least Squares) algorithms.

In addition to feature selection, multivariate dimensionality reduction techniques such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) have been used effectively. Rivera-Caicedo et al. (2017) showed that such techniques improved the LAI retrievals when coupled with NN, KRR, and GPR, achieving R^2_{cv} of 93% instead of all HyMap bands. Considering that the multi-spectral data from new generation sensors such as Sentinel-2 have increased the number of bands to more than seven, it remains questionable whether feature selection with these datasets can improve the accuracy and reduce uncertainties of retrieving BVs. Generally, the optimisation of multi-spectral features for specific crop BVs has received only limited study attention.

Table 2-1. Commonly used MLRAs for biophysical and biochemical retrieval with associated accuracy measures, i.e., Coefficient of determination, R^2, and Root Mean Squared Error, RMSE, where possible. RF, SVM, XGBoost, KRR, GPR, and ANN denote Random Forest (RF), Support Vector Machines (SVM), eXtreme Gradient Boosting, Kernel Ridge Regression, Gaussian Process Regression, and Artificial Neural Networks (ANN), respectively.

	Author(s)	**MLRA**	**Accuracy (R^2; RMSE)**	**Target BV(s)**	**Crop type (s)**	**Sensor**	**Explanatory Variables**	**Experimental site(s)**
Tree-based	Ramoelo et al. (2015b)	RF	R^2: 0.71–0.89 ; RMSE: 0.04–0.0.8	N	Grassland	Worldview-2	Spectral bands and Vegetation indices	Northern South Africa
	Li et al. (2017b)	RF	R^2: 0.88 ; RMSE: 0.195 $m^2\ m^{-2}$	LAI	Grassland	Landsat-7 ETM+; Landsat-8 OLI	Spectral bands and Vegetation indices	Hulunber, Inner Mongolia, China (49°20′24′′N, 119°59′44′′E)
	Tavakoli & Gebbers (2019)	RF	R^2: 0.56–0.69 ; RMSE: 0.27%–0.19%*	N	Wheat	Digital camera	Spectral bands and Vegetation indices	Marquardt experimental station, Potsdam, Germany (52°27′N, 12°57′E)
Kernel-based	Malenovský et al. (2017)	SVM	R^2: 0.52–0.54; RMSE: 242.6–238.3 nmol g^{-1} dry weight	LC_{ab}	Antarctic mosses	Hyperspectral UAS; WorldView-2	Spectral bands	Antarctic Specially Protected Area 135 (66.282°S, 110.539°E) and Robinson Ridge (66.368°S, 110.586°E)
	Verrelst et al. (2012)	GPR	R^2: 0.94–0.96; RMSE: 0.47–0.55 $m^2\ m^{-2}$ [¥]	LAI	Multi-crops[#]	Sentinel-2 MSI*; Sentinel-3 OLCI*	Spectral bands	Barrax, La Mancha region, Spain (30°3′ N, 2°6′ W)
	Verrelst et al. (2012)	GPR	R^2: 0.94–0.99; RMSE: 1.81–5.36 μg cm^{-2} [¥]	LC_{ab}	Multi-crops[#]	Sentinel-2*; Sentinel-3*	Spectral bands	Barrax, La Mancha region, Spain (30°3′ N, 2°6′ W)
	Campos-Taberner et al. (2016)	GPR	R^2: 0.88–0.89; RMSE: 0.78 $m^2\ m^{-2}$ [φ]	LAI	Rice	Landsat-8 OLI[ϕ]; SPOT-5[ϕ]	Spectral bands	Valencia, Spain; and Lomellina, rice district, Lombardy, Italy
	Campos-Taberner et al. (2016)	KRR	R^2: 0.82–0.83; RMSE: 0.94–0.97 $m^2\ m^{-2}$ [φ]	LAI	Rice	Landsat-8 OLI[ϕ]; SPOT-5[ϕ]	Spectral bands	Valencia, Spain; and Lomellina, rice district, Lombardy, Italy
	Verrelst et al. (2016)	GPR	R^2: 0.79 ; RMSE: 72.36 mg/m2	LC_{ab}	Multi-crops[#]	Field spectrometer[†]	Spectral bands	Barrax, La Mancha region, Spain (30°3′ N, 2°6′ W)
	Verrelst et al. (2016)	GPR	R^2: 0.94; RMSE: 0.40 $m^2\ m^{-2}$	LAI	Multi-crops[#]	Field spectrometer[†]	Spectral bands	Barrax, La Mancha region, Spain (30°3′ N, 2°6′ W)
	Verrelst et al. (2016)	GPR	R^2: 0.95; RMSE: 0.37 $m^2\ m^{-2}$	LAI	Multi-crops[#]	HyMap[†]	Spectral bands	Barrax, La Mancha region, Spain (30°3′ N, 2°6′ W)
	Wen et al. (2018)	GPR	R^2: 0.85; RMSE: 0.95 μg ml	N	Rice	Hyperspectral UAS	Spectral bands	Shenyang

								Agricultural University, Liaoning Province, China (41°81′63″N, 123°55′85″E)
Deep learning	Verger et al. (2011)	ANN	RMSE: 0.37 $m^2\ m^{-2}$	LAI	Multi-crops[#]	CHRIS/PROBA	Spectral bands	Barrax, Castilla-La Mancha region, Spain (30°3′ N, 2°6′ W)
	Campos-Taberner et al. (2016)	ANN	R^2: 0.83–0.84; RMSE: 0.91–0.93 $m^2\ m^{-2}$ [φ]	LAI	Rice	Landsat-8 OLI[ϕ]; SPOT-5[ϕ]	Spectral bands	Valencia, Spain; and Lomellina, rice district, Lombardy, Italy
	Delloye et al. (2018)	ANN	R^2: 0.55–0.85; RMSE: 1.00–0.70 $m^2\ m^{-2}$ [‡]	LAI	Wheat	Sentinel-2 MSI; SPOT-5	Spectral bands	Belgium
	Delloye et al. (2018)	ANN	R^2: 0.08–0.31; RMSE: 13.94–11.03 $\mu g\ cm^{-2}$ [‡]	LC_{ab}	Wheat	Sentinel-2 MSI; SPOT-5	Spectral bands	Belgium
	Delloye et al. (2018)	ANN	R^2: 0.46–0.62; RMSE: 0.51–0.35 $g\ m^{-2}$ [‡]	CCC	Wheat	Sentinel-2 MSI; SPOT-5	Spectral bands	Belgium
	Dhakar et al. (2019)	ANN	R^2: 0.56–0.75; RMSE: 1.34–0.94 $m^2\ m^{-2}$ [£]	LAI	Wheat	Sentinel-2 MSI	Spectral bands	Pataudi block of Gurugram district, Haryana, India

Notes for Table 2-1

[¥] Accuracy ranges are for various configurations tested in the study, i.e., S2-10 m (4 bands with 10 m spatial resolution), S2-20 m (8 bands, i.e., B2 to B8a, at 20 m resolution), S2-60 m (10 bands, i.e., B1 to B9, at 60 m resolution), and S3-300 m (19 bands, i.e., BO1 to BO20, at 300 m resolution)

* Simulated from Compact High-Resolution Imaging Spectrometry (CHRIS) data with a spectral range of 400 to 1 050 nm.

[#] Sunflower, maize, alfalfa, wheat, sugar beet, onion, garlic, potato, and vineyard

[ϕ] Simulated from PROSAIL Radiative Transfer Model

[φ] Accuracies are for simulated Landsat OLI and SPOT-5 data, respectively.

[†] Original data were reduced to fewer bands using Gaussian Process Regression-based band analysis tool (GPR-BAT)

[‡] Accuracies were achieved with simulated Sentinel-2 data divided into different band subsets, i.e., SPOT-5 (4), 10-bands (3), Red-edge (7) and All bands (9)

[£] Accuracies were achieved with inversion of PROSAIL Artificial Neural Networks (ANN) and Look-Up-Tables (LUT) using Sentinel-2 MSI corrected with MODTRAN and libRadtran atmospheric correction approaches.

* Accuracies are for estimated N over various phenologies, i.e., 219 and 234 days after sowing and combined dates.

5. Sources of Uncertainty in Agricultural Areas and Implications for Precision Agriculture

Various sources of uncertainties are a significant concern with the use of remote sensing approaches to retrieve biophysical and biochemical parameters for agronomic applications. These uncertainties emanate from several factors, such as field measurement errors (including instrument and sampling errors) and acquisition conditions (including atmospheric conditions and acquisition and illumination geometries). These may propagate to the BV retrievals using MLRA. On the one hand, accurate retrieval of agronomic parameters is dependent on the quality of field measurements, which in turn hinges on the appropriateness of the sampling strategy and uncertainties of the optical instruments, including LAI-2000 (commonly used for LAI measurements) and Minolta SPAD-502 chlorophyll meter (commonly used for LC_{ab} measurements). For example, LAI-2000 has 10%–20% uncertainties varying by species and environment (Rautiainen et al., 2012). On the other hand, SPAD-502 requires site-specific data (laboratory-measured chlorophyll content) for calibration over many crop types and leaf structures. Still, its values saturate at LC_{ab} > 40 µg cm^{-2} and differ according to growth stages, species, and distribution of LC_{ab}; hence, empirical calibration equations need to be determined for each situation (Uddling et al., 2007). This may be a costly exercise, primarily because sub-Saharan Africa lacks consistent field effort.

Therefore, the inversion of RTMs has been used in the operational production of biophysical and biochemical parameters of agricultural crops. However, owing to known poorly posed problems of RTMs, they often perform poorly in various environments (Fernandes et al., 2014; Kganyago et al., 2020a). In validating the Sentinel Application Platform (SNAP) Biophysical processor, based on PROSAIL-generated LUTs and Neural Networks, Kganyago et al. (2020a) found LAI uncertainties of > 2 m^2 m^{-2} using Sentinel-2 data and concluded that the method was not suitable for Precision Agriculture. Therefore, locally calibrated RTMs combined with other MLRAs may reduce uncertainties and improve the accuracy of biophysical and biochemical parameters, thus making it fit-for-purpose for Precision Agriculture.

Other uncertainties may be introduced by several confounding hyperparameters required in calibrating various MLRAs. The sensitivity of MLRAs to these hyperparameters in the context of BV retrieval has not yet been thoroughly studied. However, studies applying MLRAs often employ hyperparameter tuning strategies which evaluate all combinations of the required hyperparameters or those selected randomly. Then, hyperparameters which result in the lowest uncertainties (e.g., RMSE) are chosen for retrieval of biophysical and biochemical parameters. These hyperparameters are site-, crop condition- (phenology), and species-specific parameters, and a slight deviation in factors may result in higher uncertainties. This is especially important to consider when MLRA models are transferred to new areas (or sites) where the crop types and conditions, climatic (i.e., temperatures, humidity, and rainfall), and environmental (i.e., soil types, moisture) factors may be different from the original site where the model was trained. Therefore, a sound transferability framework, which considers the variability of crop types, physiological conditions, climate, and environmental factors is needed to improve the transferability of MLRAs and reduce the need for extensive field data collection. Poor image quality, mainly caused by varying atmospheric constituents between the acquisition dates and residual errors after atmospheric correction, also causes increased uncertainties in the BV retrieval models.

Changing atmospheric conditions during image acquisition has been a challenge since the dawn of satellite-based remote sensing. Aerosols (smoke, dust, and atmospheric gases such as CO2) increase the atmospheric backscattering signal, while attenuating the surface directional reflectance signal. (Hilker et al., 2009). Therefore, they undermine the radiometric integrity and consistency of the measured reflectance, which are critical for phenological analysis and reliable estimates of BVs. Spectral response to crop conditions is particularly affected by inconsistencies in radiometric measurements caused by the changing atmosphere. Wilson et al. (2014) show that aerosol optical thickness (AOT) may cause errors of about 1.7% in the measured reflectance and a 5% change in the NDVI. Alarmingly, as an integral input to Precision Agriculture, such errors manifest in remotely sensed biophysical and biochemical parameters and crop type maps and may lead to false detection of stress and other crop conditions (Vanonckelen et al., 2013;

López-Serrano et al., 2016). Consequently, atmospheric conditions have adverse implications for the quality of optical satellite images and radiometric consistency, hence downstream applications that support agricultural production and sustainability efforts.

To avert challenges posed by atmospheric conditions, previous studies (Hagolle et al., 2010; Hagolle et al., 2015; Adam et al., 2018) used cloud screening, atmospheric correction (AC) and gap-filling techniques. AC techniques minimise atmospheric effects, such as atmospheric gases and aerosols and improve the comparability of radiometric measurements over regions with various aerosol loadings. For Sentinel-2 MSI, various AC approaches were developed, validated, and inter-compared (Ruescas et al. 2016; Martins et al., 2017; Djamai & Fernandes, 2018). However, studies (Doxani et al., 2018; de Keukelaere et al., 2018; Sola et al., 2018) report variable results depending on spectral bands, sensors, land covers, environments, and atmospheric conditions. Therefore, there is no universal AC approach. When validated against field data, Kganyago et al. (2020b) showed that Sen2Cor (a dedicated AC algorithm for Sentinel-2) had poor accuracy under partly cloudy conditions, i.e., R^2 <20%, in the visible (VIS) and SWIR bands, while it had better accuracy, i.e., R^2 >60%, in the red-edge and NIR bands. Previous studies demonstrate that VIS and SWIR bands are crucial for retrieving the LC_{ab}, CCC, and LAI (Verrelst et al., 2013; Kganyago et al. 2021). Therefore, the accuracy assessment of various AC techniques in semi-arid agricultural landscapes, under various atmospheric conditions and the effect of AC on BV retrieval need to be prioritised in future studies.

6. Lessons and Recommendations

Overall, the critical lessons from the literature review are summarised as follows:

- Current LAI_G retrieval methods are unreliable for estimating LAI_B, which is an essential decision-support tool at the end of the season, for optimising harvest schedules and crop planting dates, and other logistics such as transportation and storage and modelling fire risks of senescent crops. Therefore, future studies should intensify research and operationalisation of high-resolution LAI_B products.

- There is a lack of consensus in the literature about the relationship between CCC and N content, where some authors such as Baret et al. (2007) show that CCC could be used in canopy N content. Others such as Vincini et al. (2016) argue that it is inadequate for distinguishing N deficiency from other crop stressors. Based on these arguments, we recommended LC_{ab} for precise N status management and CCC for assessing the general crop development, productivity, and overall stress as it combines the effects of both C_{ab} pigments, plant structure (LAI), and other effects which act on these individual parameters.
- In developing regions such as sub-Saharan Africa, the lack of adoption of UAVs can be attributed to a myriad of complex factors, such as smaller field sizes (i.e., usually <0.5 ha) such that economic benefits of UAVs cannot be realised, lack of awareness, technical, and institutional capacities, and poor rural connectivity.
- The compromise between the spatial and temporal resolutions in remote sensing reduces the utility of the datasets for Precision Agriculture. However, data fusion algorithms are promising and can reduce the shortcomings of both the low- and medium/high-resolution sensors, while also improving their utility for Precision Agriculture by exploiting good qualities of each, i.e., higher repeat cycles (up to <1 day) and detailed coverage (<30 m).
- New high-resolution and very-high-resolution sensors (e.g., SPOT-6 and -7 and PlanetScope's Doves constellation) continue to acquire data in the VNIR region, despite the reported saturation of crop biophysical parameters retrieved within this region of the electromagnetic spectrum. This is attributed to their lower cost and the fact that VNIR indices are still popular in various vegetation analyses and easily understood by users. Moreover, high spatial and temporal resolutions enabled by off-nadir viewing, tasking capabilities, and satellite constellations seem to be the most attractive factor for operational site-specific management than the spectral bands they provide.

- CubeSat constellations such as PlanetScope's SuperDove can be considered a revolution of the EO satellite sensor technology for Precision Agriculture and will likely reduce the cost of VHR data, thus increasing their accessibility to researchers, product developers, governments, and farmers. It promises five to eight spectral bands, including VNIR bands with two green bands (513–549 nm and 547–583 nm), and yellow (600–620 nm) and red-edge bands (697–713 nm) at very high temporal (i.e., daily) and spatial resolution (3 m). This will provide the capability and opportunities for detailed, accurate, and frequent characterisation of field variability.
- The launch of the Sentinel-2 constellation, i.e., Sentinel-2A (in 2016) and -2B (in 2017), carrying identical Multi-spectral Imager (MSI) cameras, was a ground-breaking event in the EO community. It has not only guaranteed data continuity of, and interoperability with, the previous missions, but also provides better spatial resolutions (up to 10 m), temporal revisits (five days), and additional spectral bands with improved spectral resolution at no cost to users. The new red-edge bands, centred at 705 nm, 740 nm, and 783 nm, have triggered new research, and present prospects for operational and accurate monitoring of stress-related agricultural crop parameters to aid time-sensitive agricultural decisions and improved yields.
- Besides the latest developments in MLRAs algorithms, this review found that RF still features prominently in the literature, with studies citing the main attractive aspects being good performance and relatively few required hyperparameters. In general, tree-based algorithms are also appealing because they are interpretable and explainable, allowing interrogation of the tree structure and variables used. They also indicate influential variables. In contrast, kernel-based and deep learning algorithms are considered complex, computationally expensive, and opaque (or "black box"). Some limitations are a requirement to tune many confounding hyperparameters, and there is no capability to directly compute variable importance and interrogate the inner workings

of models. Therefore, the effect of each hyperparameter on the accuracy of these MLRAs has not been studied yet. Nonetheless, these MLRAs have attracted significant attention in recent studies that have used simulated and accurate data from various hyper- and multi-spectral sensors and comparatively high accuracies were reported.

- The greatest interest was shown in GPR and its variants, where most studies report relatively high accuracies and a special capability to provide response uncertainty estimates of variables that enable assessment of the reliability of the crop parameter retrievals for practical applications. Like tree-based algorithms, GPR provides insights into the relevant bands, which indicate the relationship with the response variables. However, the commonly used deep learning algorithm is ANN, which is computationally expensive and complex with many hyperparameters.
- Coupling MLRAs with multivariate dimensionality reduction enhances the accuracy of crop parameter retrievals even when using powerful algorithms such as ANN, KRR, and GPR. Considering that the data from quasi-hyperspectral sensors such as Sentinel-2 have a high number of bands, it remains questionable whether feature selection with this dataset can improve the accuracy and reduce uncertainties of retrieving biophysical and biochemical parameters. Generally, the optimisation of multi-spectral features for specific crop biophysical and biochemical parameters has been the focus of only limited study. Therefore, future studies should address this gap.

7 Conclusions

This paper sought to provide a comprehensive review of recent advances in remotely sensed retrieval of biochemical and biophysical parameters of agricultural crops brought by the developments in sensor technologies and novel machine learning retrieval techniques. Moreover, sources of uncertainty in retrieving crop parameters were identified, and practical implications for precision farming were discussed. Overall, the review revealed that developments in crop parameter retrieval techniques were mainly driven by announcements and the availability of new sensors, with the launch of the Sentinel-2 constellation being a ground-breaking event. Many studies were conducted with simulated data and airborne hyperspectral sensors at specific study areas with time series of field data covering many crops. Such permanent experimental sites are missing in sub-Saharan Africa, and there is a lack of coordinated and systematic efforts targeting calibration and validation data collection.

Although some models were tested across many sites, in most cases, such sites were in Mediterranean climates. Hence, such models may not be portable to semi-arid African agricultural areas. Another consideration is that farm sizes in developed countries are significantly larger than in sub-Saharan Africa. Therefore, crop-specific models may not be relevant in most African farm configurations characterised by small field sizes (i.e., <0.5 Ha) and mixed-cropping systems. In addition to addressing gaps identified here, future research should focus on the development of generic (multi-crop), scalable (multi-sensor) and transferable models (across many sites and growing stages), especially within under-studied sub-Saharan African sites.

Appendix 2-1. Literature Search and Selection

We reviewed journal articles with the status "published" and "In Press" included by the Scopus Abstract and Citation Database (www.scopus.com, accessed 26 February 2021). The Scopus database is the most robust, reliable, and frequently used research engine, allowing the discovery of present and past research and patented materials. Other databases, such as Web of Science and Google Scholar, were not used during the search owing to irregular updates and erroneous information, as anyone can upload articles. The keywords used are outlined in **Table 1-S1**.

Table 1-S1. Keywords used to search the Titles, Abstract, and Keywords.

Category	Keywords
Biophysical parameters	"Leaf Area Index" OR "Canopy chlorophyll content" OR "Nitrogen Content" OR "n-content OR ccc OR lai"
Data	AND sentinel-2 OR landsat OR spot OR cbers OR "medium resolution" OR "remote sensing"
Methods	OR "Machine learning" OR "Deep learning" OR "Radiative transfer models"

TITLE-ABS-KEY ("Leaf Area Index" OR "Canopy chlorophyll content" OR "Nitrogen Content" OR n-content OR ccc OR lai AND sentinel-2 OR landsat OR sentinel-2 OR spot OR cbers OR "medium resolution" OR "remote sensing" OR "Machine learning" OR "Deep learning" OR "Radiative transfer models") AND PUBYEAR > 2009 AND PUBYEAR < 2021 AND (LIMIT-TO (DOCTYPE, "ar")) AND (EXCLUDE (EXACTKEYWORD , "Forestry")) AND (LIMIT-TO (SUBJAREA, "AGRI") OR LIMIT-TO (SUBJAREA , "ENVI")) AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (LANGUAGE , "English"))

We excluded conference papers (as they are often republished as journal articles), magazine articles, editorials, communications, letters to editor, book chapters and review articles, forestry and non-English articles, and journal articles published before 2011 and after 2020. The string code returned 723 journal articles. Their abstracts were scanned for relevance. The number of publications between 2011 and 2020 (**Figure 1-S1**) shows an increasing trend over time. The total number of publications increased slowly, leading to 2015 and declining in 2016. From 2017, a sharp increase can be observed, with a peak in 2020. This sharp increase from 2017 can be attributed to the availability of new sensors such as Sentinel-2 and Sentinel-3, providing the opportunity to develop new methods in various study areas. **Figure 2-S1** gives the distribution of publications by country.

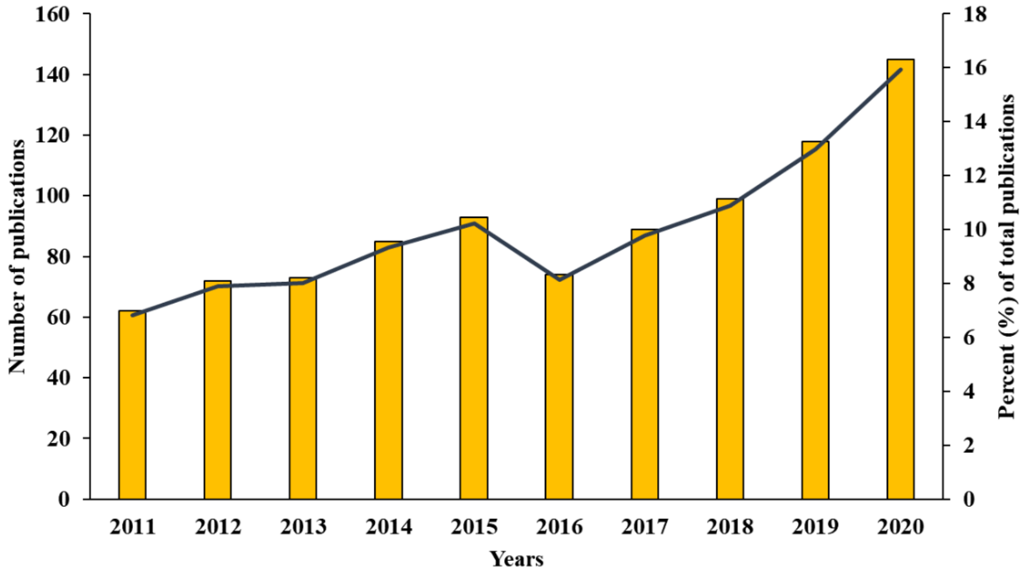


Figure 1-S1. Frequency of publications by year.

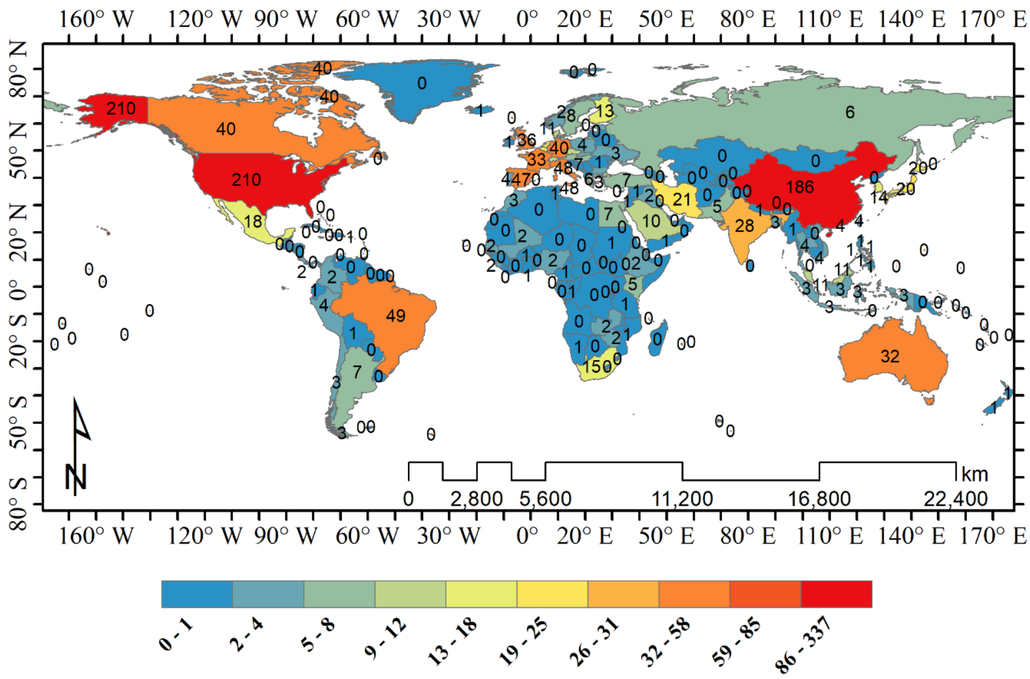


Figure 2-S2. The number of publications by country.

CHAPTER 3
VALIDATION AND COMPARISON OF SENTINEL-2 BIOPHYSICAL PARAMETERS FROM SNAP TOOLBOX IN AN AFRICAN SEMI-ARID AGRICULTURAL LANDSCAPE

This chapter is based on the following published manuscript:

Kganyago, M., Mhangara, P., Alexandridis, T., Laneve, G., Ovakoglou, G., Mashiyi, N. (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. DOI: https://doi.org/10.1080/2150704X.2020.1767823

This paper received the Taylor and Francis Remote Sensing Letters Award for best paper published in the Remote Sensing Letters in the year 2020 (see Appendix 3-1).

Abstract

This study considered the validation of SNAP-derived Leaf Area Index (LAI) from Sentinel-2 data, and its consistency with existing global LAI products. The validation and inter-comparison experiments were performed on two processing levels, i.e., Top-of-Atmosphere and Bottom-of-Atmosphere reflectances and two spatial resolutions, i.e., 10 m, and 20 m, characterising Sentinel-2 data. The processing levels and spatial resolutions were chosen to determine their effect on retrieved LAI accuracy and consistency. On the one hand, the results showed moderate R^2, i.e., ~0.6 to ~0.7 between SNAP-derived LAI and in situ LAI, but with high errors, i.e., RMSE, BIAS, and MAE >2 $m^2\ m^{-2}$ with marked differences between processing levels and insignificant differences between spatial resolutions. On the other hand, inter-comparison of SNAP-derived LAI with MODIS and Proba-V LAI products revealed moderate to high consistencies, i.e., R^2 of ~0.55 and ~0.8 respectively, and RMSE of ~0.5 $m^2\ m^{-2}$ and ~0.6 $m^2\ m^{-2}$, respectively. The results in this study have implications for future use of SNAP-derived LAI from Sentinel-2 imagery in agricultural landscapes, suggesting its global applicability which is essential for large-scale agricultural monitoring. However, owing to enormous errors in characterising field-level LAI variability, this study found that SNAP-derived LAI is not suitable for precision farming. The need for improvement in LAI retrieval arises from the study, especially to support farm-level agricultural management decisions.

1. Introduction

Globally, improving remotely sensed characterisation of biophysical properties of vegetation is of paramount importance for a variety of applications (Davi et al., 2009; Zhu, Yang, et al., 2010). In particular, the non-destructive estimation of the Leaf Area Index (LAI) from Earth observation data has been topical for decades. LAI is defined as the one-sided green leaf area per unit ground area (Myneni, 2012), and it is regarded as an Essential Climate Variable (ECV) (GCOS, 2011). This is mainly owing to its significance in characterising basic information related to vegetation growth and productivity such as foliage density, plant health, and functioning, as well as its significance for modelling water, carbon, and energy fluxes between land and atmosphere (Chen et al., 2002; Verrelst et al., 2015). In agriculture, LAI is essential for, among others, monitoring the variability in crops and rangelands productivity, crop stress and health, biomass, phenology, and yield estimation (Mulla, 2013; Cho et al., 2017; Novelli & Vuolo, 2019). Unfortunately, traditional, direct (in situ) LAI measurement methods are spatially and temporally limited, expensive, time-consuming, labour-intensive, and destructive (Alexandridis et al., 2013). Therefore, remotely sensed effective LAI (hereafter, LAI) provides a promising alternative for operational agricultural monitoring to support the implementation of global and regional food security mandates such as the Sustainable Development Goals (SDGs) of the United Nations (UN), and Agenda 2063 (Kganyago & Mhangara 2019).

The availability of free-of-cost, high-resolution space-borne satellite datasets (finer than 30 m) from sensors such as Sentinel-2, advanced open-source tools such as Sentinel Application Platform (SNAP), and the increasingly available Analysis-Ready-Data (ARD) offers prospects for accurate, consistent, and operational LAI. This should overcome the limitations of coarse spatial resolution (i.e., 300–1 000 m) LAI products, such as those derived from MODerate Resolution Imaging Spectroradiometer (MODIS) (Myneni et al., 2002), Advanced Very-High-Resolution Radiometer (AVHRR) (García-Haro et al., 2018), Satellite Pour l'Observation de la Terre Vegetation (SPOT-VGT), and Proba-V (Baret et al., 2013). Besides better resolution, Sentinel-2's key advantage over other freely available sensors is its temporal resolution (i.e., ±5 days), which is adequate (if cloud-free) for most agricultural monitoring applications. In contrast, SNAP Toolbox provides advanced functionalities for any user to perform the atmospheric correction through Sen2Cor tool (Louis et al., 2016), and estimate biophysical parameters based on a physically based Radiative Transfer Model, i.e., PROSAIL and Robust Machine Learning Algorithm, i.e., Neural Networks (Weiss & Baret, 2016). However, the validation of SNAP-derived LAI from Sentinel-2 data, and its consistency with existing LAI products, have been explored by only a few studies (Bochenek et al., 2017; Campos-Taberner et al., 2018). Such information is invaluable since accurate and reliable agricultural monitoring strongly hinges on the consistency and inter-comparability of biophysical parameters such as LAI (Alexandridis et al., 2019). Furthermore, uncertainties related to the effect of the spatial resolution and processing level of the data on derived biophysical parameters such as LAI need further assessment, especially in Africa, where such studies are rare. Moreover, the quantifying of uncertainties in remotely sensed products is useful for users and developers interested in the operational use of the product, and further development, respectively.

Therefore, the current study sought to validate and inter-compare SNAP-derived LAI from Sentinel-2 Top-of-Atmosphere (TOA) and Bottom-Of-Atmosphere (BOA) reflectances at two spatial resolutions, i.e., 10 m and 20 m, using in situ data and global LAI products, i.e., MODIS LAI (MCD15A3H) with 500 m resolution and Copernicus LAI based on Proba-V data (hereafter, Proba-V LAI) with 300 m resolution. The study was conducted in an African agricultural landscape characterised by large commercial and small-holder farming systems.

2. Study Area

The study area is located at latitudes 27°13′0″S to 28°8′0″S and longitudes 26°0′0″E to 27°05′0″E, in the vicinity of Bothaville in Free State province, South Africa. The area constitutes the primary agricultural production zone of South Africa and is characterised on the one hand by warm and wet summers with an average temperature of about 18 °C and an annual average rainfall of about 584 mm. On the other hand, winters are cold, with average temperatures of about 5 °C. Large-scale and small-holder commercial crop cultivation of mainly maize and sunflower occur on flat, undulating landscapes, with loamy to sandy-loamy soils in the summer crop calendar (i.e., from December to May or June) each year.

3. Materials and Methods

3.1. Data

3.1.1. Remotely sensed data

Sentinel-2 Multi-Spectral Imager (tile 35JMK) TOA image acquired on 15 April 2019 (i.e., Peak of Season) was retrieved from European Space Agency (ESA) Copernicus Open Access Hub (https://scihub.copernicus.eu/dhus/). The TOA image data were then converted to BOA reflectance, i.e., Level-2A using Sen2Cor version 2.8 (Louis et al. 2016). Finally, both TOA and BOA images were then resampled to the two spatial resolutions for Sentinel-2, i.e., 10 m and 20 m, using the Sentinel-2 Resampling algorithm available within SNAP Toolbox.

3.1.2. SNAP-derived Leaf Area Index

Sentinel-2 TOA and BOA images at two spatial resolutions (10 m and 20 m) were used to derive LAI, using the built-in Biophysical processor, also called Sentinel-2 Land biophysical processor (SL2P) within SNAP Toolbox. The Biophysical processor uses eight reflectance bands (B3, B4, B5, B6, B7, B8A, B11, and B12), as well as viewing zenith, solar zenith, and relative azimuth angles using Radiative Transfer Models (RTMs), i.e., PROSAIL and Neural Networks (NN) algorithm. Further details can be found in Weiss & Baret (2016). SNAP-derived LAI (i.e., 20 m) was rescaled and co-registered to match the spatial resolution of the MODIS LAI product (i.e., 500 m) and the Proba-V LAI product (i.e., 300 m), using the nearest neighbour resampling technique. The resampled SNAP LAI products (i.e., 500 m and 300 m) were used for comparison to MODIS and Proba-V LAI products.

3.1.3. Global Leaf Area Index products

Global LAI products from Proba-V with 300 m spatial resolution (Baret et al., 2013) and MODIS (MCD15A3H.006) with 500 m resolution (Myneni et al., 2015) closest to the date of Sentinel-2 acquisition, to ensure comparability, were acquired from Copernicus Global Land Service (http://land.copernicus.eu/global) and Google Earth Engine™ (data provided by NASA LP DAAC, https://lpdaac.usgs.gov/), respectively. Specifically, Proba-V LAI consisted of a temporal composite of ten days from 11 to 20 April 2019, while MODIS LAI product was a temporal composite of four days from 11 to 15 April 2019. The global LAI products were chosen based on their relatively high resolution compared to other existing global products (Camacho et al., 2013; Claverie et al., 2016) and were used for comparison with SNAP-derived LAI.

3.1.4. In situ Leaf Area Index

In situ effective Leaf Area Index (LAI) samples were collected in 40 m × 40 m plots along transects randomly selected per field using LI-COR 2200c Plant Canopy Analyzer (LI-COR, 2012). Within each plot of available LAI measurements, there was one above-canopy and an average of nine (9) below-canopy measurements taken randomly to capture the variability of each crop. These LAI measurements were collected under predominantly clear-sky conditions from 8 to 14 April 2019 (i.e., Peak of Season). **Table 3-1** indicates summary statistics for in situ LAI.

Table 3-1. Descriptive statistics of measured LAI (m^2 m^{-2}) used for validating SNAP-derived LAI.

Crop type	Number of plots	Min	Mean	Max	Standard deviation
Overall	65	1.49	3.86	5.83	1.23
Maize	31	1.54	3.62	5.42	0.98
Sunflower	34	1.49	4.07	5.83	1.40

3.2. Performance metrics

The validation of LAI and CCC with in situ data and comparison with global LAI products was performed using the coefficient of determination (R^2), Root Mean Squared Error (RMSE), mean absolute error (MAE), and BIAS (Equation 3-1 to 3-3).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \quad (3\text{-}1)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (3\text{-}2)$$

$$\mathrm{Bias} = \sum_{i=1}^{n}(x_i - y_i) \quad (3\text{-}3)$$

Where: x_i is the observed biophysical parameter (e.g., LAI), y_i is the predicted biophysical parameter (e.g., LAI), N is the sample size, and n is the number of errors.

4. Results and Discussion

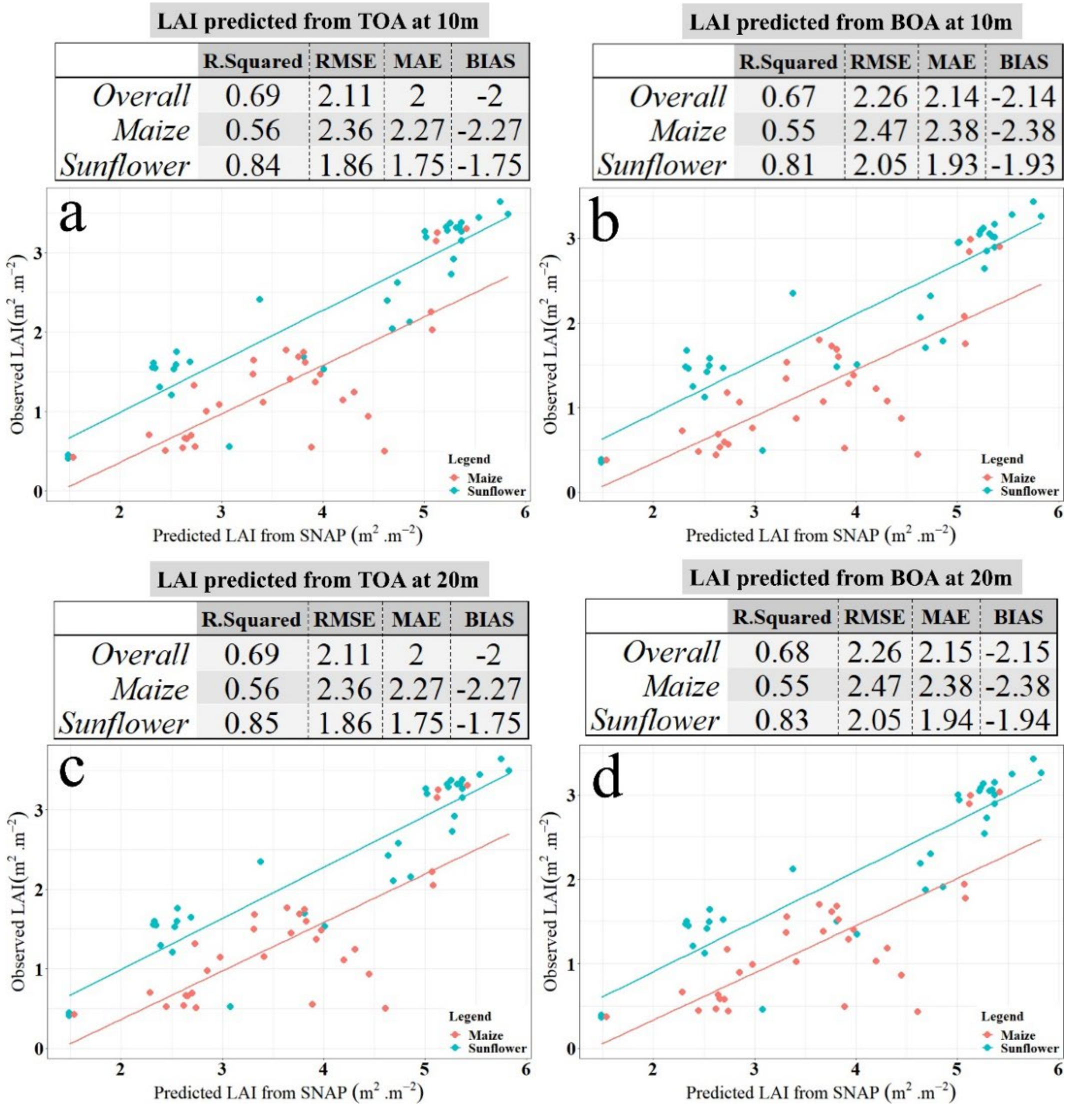
4.1. Validation of LAI derived from SNAP

SNAP-derived LAI from Sentinel-2 data at two spatial resolutions and two processing levels, i.e., TOA (L1C) and BOA (L2A), were validated using in situ LAI data. The results (**Figure 3-1**) show differences between SNAP-derived LAI at various spatial resolutions and processing levels. Specifically, SNAP-derived LAI at 10 m and 20 m spatial resolution showed similar overall agreement with observed (in situ) LAI, i.e., R^2 = 0.69 for L1C. In comparison, there were marginal differences in overall agreement at 10 m and 20 m, i.e., R^2 = 0.67 and 0.68, respectively, for L2A. Nevertheless, the error metrics were generally similar at both resolutions. Moreover, the results show that LAI derived from TOA reflectances had marginally better agreement with observed LAI, i.e., R^2 = 0.69 than that derived from BOA reflectances that achieved R^2 ~0.68 across all spatial resolutions. Generally, the performance of SNAP-derived LAI, in this study, is better than that found by Pasqualotto et al. (2019a), i.e., R^2 ~0.54, using SNAP-derived LAI from BOA reflectances at 10 m spatial resolution over two agricultural sites in Spain and Italy.

Considering LAI over individual crops, i.e., maize and sunflower, the results show remarkably strong agreement with observed LAI over sunflower, i.e., R^2 >0.8 across spatial resolutions and processing levels. In contrast, maize showed moderate R^2 values of up to 0.56 and marginal differences between processing levels. As can be expected, the error metrics, i.e., RMSE, MAE, and BIAS, were also the lowest over sunflower, with 10 m and 20 m TOA reflectances achieving equivalent RMSE (i.e., 1.86 $m^2\ m^{-2}$), and MAE (i.e., 1.75 $m^2\ m^{-2}$). In contrast, the most significant errors were evident over maize, with RMSE and MAE >2 $m^2\ m^{-2}$ across all considered spatial resolutions and processing levels. The magnitude of errors found in this study in SNAP-derived LAI from BOA data can be attributed to residual errors of atmospheric correction (AC) using the Sen2Cor procedure.

This is because studies have shown that various AC approaches perform differently for different environments, land cover types, and spectral bands (Doxani et al., 2018; Sola et al., 2018). For example, Sentinel-2 visible bands used in the Biophysical processor are known to be sensitive to Rayleigh and aerosol scattering effects (Martins et al., 2017).

Djamai and Fernandes (2018) show that such uncertainties in BOA data cause proportional uncertainties in derived biophysical parameters. However, this aspect requires further investigation in future studies, for example, by assessing the effect of various AC approaches on retrieval of LAI and other remotely sensed biophysical parameters. Furthermore, an apparent overestimation of LAI values for all crops by SNAP Biophysical processor using both TOA and BOA reflectances is evident in scatter plots (**Figure 3-1**) and BIAS. This finding is consistent with Bochenek et al. (2017) who found that SNAP-derived LAI is unreliable based on its overestimation of LAI at varying crop developmental stages. The overestimation of LAI and significant biases observed from SNAP-derived LAI may be the result of prior assumptions made based on global LAI retrieval algorithms, i.e., Artificial Neural Networks (ANN) and Look-Up-Tables (LUTs) simulated with PROSAIL RTM, i.e., PROSPECT+SAIL that may not be calibrated to local conditions. Therefore, as previous studies have demonstrated, models that are calibrated locally may be more robust and increase operational product accuracy (Djamai et al., 2019; Jiang et al., 2019).



LAI predicted from TOA at 10m

	R.Squared	RMSE	MAE	BIAS
Overall	0.69	2.11	2	-2
Maize	0.56	2.36	2.27	-2.27
Sunflower	0.84	1.86	1.75	-1.75

LAI predicted from BOA at 10m

	R.Squared	RMSE	MAE	BIAS
Overall	0.67	2.26	2.14	-2.14
Maize	0.55	2.47	2.38	-2.38
Sunflower	0.81	2.05	1.93	-1.93

LAI predicted from TOA at 20m

	R.Squared	RMSE	MAE	BIAS
Overall	0.69	2.11	2	-2
Maize	0.56	2.36	2.27	-2.27
Sunflower	0.85	1.86	1.75	-1.75

LAI predicted from BOA at 20m

	R.Squared	RMSE	MAE	BIAS
Overall	0.68	2.26	2.15	-2.15
Maize	0.55	2.47	2.38	-2.38
Sunflower	0.83	2.05	1.94	-1.94

Figure 3-1. Scatterplots and associated statistical metrics for observed (in situ) Leaf Area Index (LAI) against predicted LAI from Sentinel-2 at 10 m (a and b) and 20 m (c and d) using Biophysical processor within SNAP Toolbox. Figures 3-1a and c show observed LAI against predicted LAI from TOA data, while Figures 3-1b and d show observed LAI against predicted LAI from BOA data.

Although not considered in this study, in situ measurement errors may have a great influence on the validation of satellite-based LAI products and lead to erroneous conclusions. Therefore, future studies should consider this aspect in detail. Overall, the large uncertainties observed in this study necessitate further studies to improve and optimise LAI retrievals, especially for supporting field-level agricultural management decisions, i.e., precision farming.

4.2. Inter-comparison of SNAP LAI with global LAI products

Inter-comparison of SNAP-derived LAI to existing and well-established global LAI products, namely, MODIS LAI (MCD15A3H) and Proba-V LAI, is essential to determine its consistency and compatibility with such products. Generally, the results (**Figure 3-2**) indicate that SNAP-derived LAI from TOA data are more correlated to global LAI products than that from BOA data. As a result, the error metrics, i.e., RMSE, MAE, and BIAS, were also lower in SNAP-derived LAI from TOA than those from BOA data for both MODIS and Proba-V LAI. This may be attributed to errors in the input data, including residual errors of atmospheric correction using the Sen2Cor procedure in Sentinel-2 bands since atmospheric correction performance varies by spectral bands (Djamai & Fernandes, 2018).

Moreover, the difference in algorithmic design and input variables between SNAP-derived LAI and global LAI products may be the reason for moderate R^2 for MODIS LAI and relatively higher R^2 for Proba-V LAI. For example, the MODIS LAI algorithm is biome-specific, uses LUTs generated using 3D RTMs, and input variables include vegetation structural type, sun-sensor geometry, bidirectional reflectance functions at red (648 nm) and near-infrared (NIR, 858 nm) bands, and their uncertainties (Knyazikhin et al., 1998). Therefore, misclassification of vegetation type per pixel of interest resulting from the coarse resolution (Tian et al., 2000), might also be a source of error during the comparison with higher-resolution Sentinel-2 data. Contrarily, Biophysical processor (i.e., S2LP) is similar to the Proba-V LAI algorithm, i.e., both are based on ANN, and more input variables are considered which include the cosine of the: view zenith angle, solar zenith angle, and relative azimuth angle, as well as three Proba-V BOA bands, i.e., blue (470 nm), red (650 nm) and NIR (837 nm). Therefore, SNAP-derived LAI and Proba-V LAI are likely to be similar even though the Biophysical processor incorporates more variables such as visible (except blue band), red-edge (RE), NIR, and shortwave infrared (SWIR) bands and LUTs generated from PROSAIL RTMs. The relatively high number of errors in Proba-V may be a result of the use of the blue band. In most studies, the blue band is not used because of its high sensitivity to atmospheric contamination, i.e., Rayleigh and aerosol scattering (Martins et al., 2017). Furthermore, the higher dimensionality of inputs in the SNAP Biophysical processor may be the reason for higher errors between SNAP-derived LAI and Proba-V LAI (Li et al., 2014; Lei et al., 2018). Therefore, future studies should focus on the selection of the optimal subset of variables that improves the estimation of LAI.

The agreement between products is also evidenced by separability metrics in **Table 3–2**, which indicates poor separability (i.e., high similarity) between SNAP-derived LAI and global LAI products. Overall, the results showed that SNAP-derived LAI from TOA data had better consistencies with global LAI products than that from BOA data, owing to relatively higher R^2 with global LAI products as well as relatively lower errors, i.e., RMSE, MAE, and BIAS.

Table 3-2. Separability metrics for SNAP LAI distributions with MODIS LAI (MCD15A3H) and Proba-V LAI products at 500 m and 300 m spatial resolutions, respectively. JM-distance values closer to 2 indicate complete separability between LAI values from SNAP and MODIS/Proba-V LAI, while values < 1 indicate poor separability. M-Statistic < 1 indicates poor separability, while M > 1 indicates good separability.

	SNAP LAI from TOA		SNAP LAI from BOA	
	MODIS	Proba-V	MODIS	Proba-V
JM-distance	0.09	0.19	0.13	0.25
M-Statistic	0.29	0.36	0.35	0.41

Despite the relatively coarse resolution, SNAP-derived LAI was more consistent with MODIS LAI than Proba-V LAI, owing to relatively lower errors. This suggests that the aggregation (i.e., down-sampling) of SNAP-derived LAI from 20 m to MODIS and Proba-V spatial resolutions, i.e., 500 m and 300 m, respectively, did not have a significant effect on LAI values. This finding is consistent with Bochenek et al. (2017) who found that aggregation of information to coarser resolution has an insignificant impact on estimated LAI values. Therefore, better consistencies, i.e., lower errors with MODIS LAI, can be attributed to its shorter temporal aggregation, i.e., four days, than Proba-V, which had a relatively long temporal aggregation of ten days. Overall, downscaling of SNAP-derived LAI to global product resolutions (i.e., 500 m and 300 m) included all the errors and noises existing in the higher resolution product, which may be one of the reasons explaining some of the poorly correlating results. Hence, the application of a smoothing technique before applying the downscaling method on high-resolution products might have been beneficial for the accuracy of the relevant results.

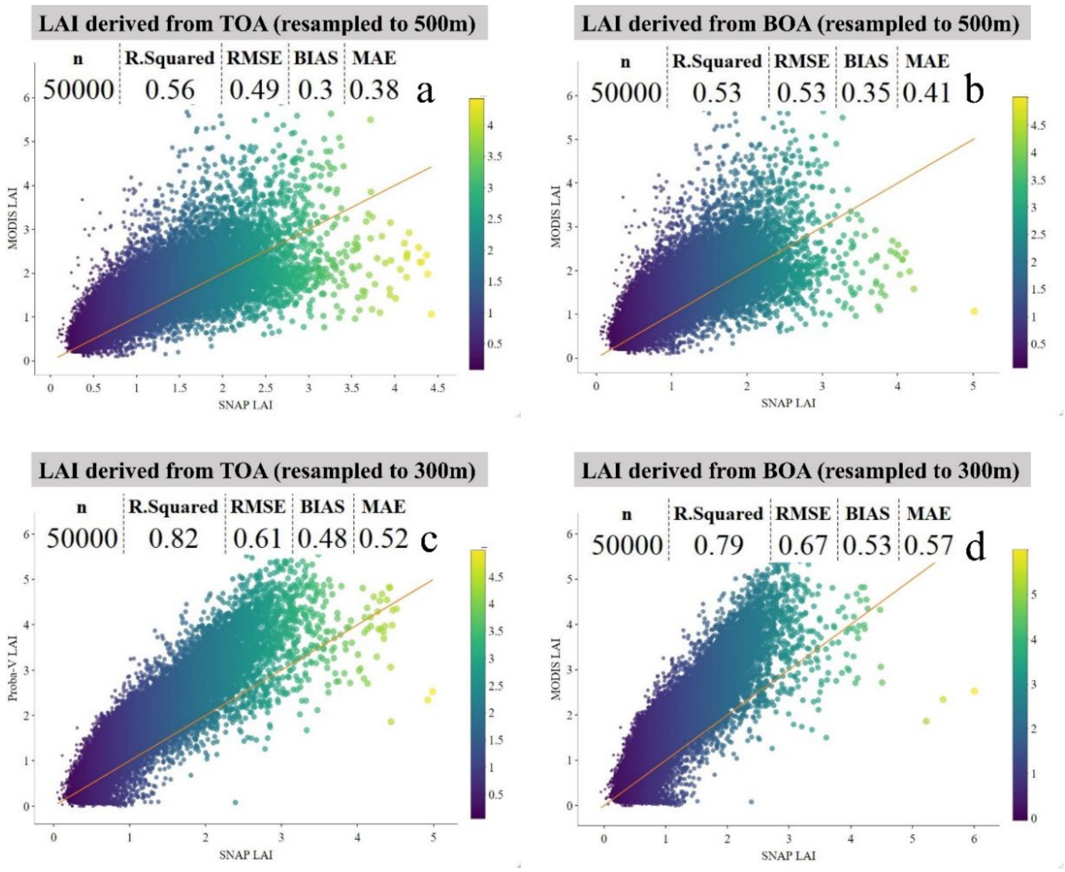


Figure 3-2. Scatterplots of MODIS LAI (MCD15A3H) against SNAP-derived LAI at 500 m (a and b), and Proba-V LAI against SNAP-derived LAI at 300 m (c and d). Figure 3-2a and c, are results for MODIS/Proba-V LAI against SNAP-derived LAI from TOA data, while Figure 3-2b and d, are results for comparison with SNAP-derived LAI from BOA data.

5. Conclusions

This study revealed interesting results:

- SNAP-derived LAI correlates moderately well in agricultural landscapes when compared with in situ LAI data; however, the errors are relatively large for field-level agricultural management. Crop-specific error assessment revealed that SNAP-derived LAI values were underestimated, and accuracy varied by crop type; therefore, future studies should investigate this finding further.
- Spatial resolution and Sentinel-2 processing levels have an impact on the

performance of SNAP-derived LAI; however, the difference in accuracy between different Sentinel-2 resolutions is negligible. The finding is consistent with previous studies (Sprintsin et al., 2007; Bochenek et al., 2017). In contrast, marked differences between BOA- and TOA-derived LAI across resolutions were observed, attributable to the residual errors after the atmospheric correction procedure in certain bands. This is the case since atmospheric correction performance varies by different spectral bands of Sentinel-2 (Djamai & Fernandes, 2018). Therefore, future studies should consider this aspect in more detail, for instance, by quantifying the effect of various atmospheric correction approaches such as Sen2Cor, FORCE, *i*Cor, and MAJA on biophysical parameters retrieval accuracy.

- SNAP-derived LAI is consistent with global LAI products; however, the errors for MODIS LAI (MCD15A3H) and Proba-V LAI are markedly different. This indicates its global applicability and potential for operational large-scale agricultural monitoring where within-field variability is not a concern. However, owing to poor field-level performance, SNAP-derived LAI is not suitable for Precision Agriculture. We hypothesise that a better fit and reduction in errors between SNAP-derived LAI and global LAI products can be achieved if noise-reduction and smoothing techniques are applied to MODIS and Proba-V.

Overall, the poor performance of SNAP-derived LAI, when validated against in situ LAI data, necessitates further improvement to support Precision Agriculture, while relatively good consistency with global LAI products suggests that it can be applicable and sufficient for monitoring large-scale agriculture. The results have implications for the future use of SNAP-derived LAI from Sentinel-2 imagery. They confirm the usefulness of Sentinel-2 data and SNAP Toolbox for supporting regional and national agricultural management decisions and policymaking, towards the achievement of global mandates such as the SDGs. This is especially important in Africa where agricultural production information is limited.

Appendix 3-1. Taylor and Francis and RSPSoc Best Paper Award

RSPSoc Office
c/o The School of Geography
The University of Nottingham
University Park
Nottingham NG7 2RD UK

Telephone: +44 (0) 115 9515435
Facsimile: +44 (0) 115 9515249
Email: rspsoc@nottingham.ac.uk
http://www.rspsoc.org.uk

PJM/rh

30th July 2021

Mahlatse Kganyago,
South African National Space Agency,
Pretoria, SA
Email: mkganyago@sansa.org.za

Dear Mahlatse

The Taylor & Francis Remote Sensing Letters Award

It gives me great pleasure to inform you that the Council of The Remote Sensing and Photogrammetry Society has awarded you its *Remote Sensing Letters Award* for your paper: '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', published in Remote Sensing Letters.

The award takes the form of a certificate and 1 year's free subscription to the *International Journal of Remote Sensing* (or £100 book token). I hope that you will accept this award, which we would like to present to you during The Society's Virtual Annual Conference (*UK EO Week 2021* https://www.eoweek2021.uk/). The Awards will be presented during RSPSoc Day on Thursday 9th September 2021 with a time to be confirmed.

We would be grateful if you would confirm whether or not you will be attending to receive this award. Please let us know if there are any time slots on that date that cause problems for you, and we will try to accommodate your needs. Additionally, if you wish to give a 10 minute talk on your award winning topic, please let us know by Monday 9th August.

As an award winner you will be eligible to attend the Conference free of charge. Please contact the RSPSoc Office at rspsoc@nottingham.ac.uk for further details of how to register.

We offer you our warmest congratulations and look forward to seeing you at the UK EO Week 2021.

Yours sincerely

Philippa J Mason
Honorary General Secretary

Dr K Smith, Chair
Dr W Xiao, Convener of the Awards & Professional Standards Committee
Mr M Williams, Conference Convenor

The Society is a charity, registered in England and Wales, (Charity No. 292647) and a company limited by guarantee.

CHAPTER 4
VALIDATION AND INTER-COMPARISON OF ATMOSPHERIC CORRECTION APPROACHES AND IMPACT ON CROP BIOPHYSICAL AND BIOCHEMICAL PARAMETERS

This chapter is based on the following manuscripts:

Kganyago, M., Ovakoglou, G., Mhangara, P., Alexandridis, T., Odindi, J., Adjorlolo, C., Mashiyi, N., "Validation of atmospheric correction approaches for Sentinel-2 under partly cloudy conditions in an African agricultural landscape", Proc. SPIE 11531, Remote Sensing of Clouds and the Atmosphere XXV, 115310B (20 September 2020); https://doi.org/10.1117/12.2572293. The conference presentation can be accessed on the DOI link.

Kganyago, M., Mhangara, P., Adjorlolo, C. (**Submitted**). Effect of Atmospheric Correction on Biophysical and Biochemical Retrieval: A Comparison of Sen2Cor and *i*Cor in Semi-arid African Agricultural landscape. *International Journal of Remote Sensing. [Manuscript ID: TRES-PAP-2022-1462]*

Abstract

This study sought to validate the surface reflectance (SR) derived by Sen2Cor and *i*Cor atmospheric correction (AC) techniques, to determine the consistency between these AC techniques and TOA reflectance, and the impact of the AC on the retrieval accuracy of crop biophysical and biochemical parameters. The results showed that both AC techniques had poor relationship with the reference SR in the VIS and SWIR bands. This is likely to negatively impact retrieval techniques—such as assessment of structural parameters (i.e., LAI), canopy water, and nitrogen content—that depend on these bands. The relationships were good (R^2 ~0.6) in the RE and NIR bands across all AC approaches, indicating that these bands can achieve reliable crop parameter estimates, especially in partly cloudy scenes (i.e., ~60% cloud cover). However, Sen2Cor and *i*Cor demonstrated a nearly 1:1 resemblance to TOA data in most bands, suggesting that the AC approach had little impact and may be disregarded when obtaining some crop parameters (such as LAI) directly from TOA. Nevertheless, the results indicated that AC may be required for chlorophyll crop parameters. The analysis in this study indicates the need for further validation in future studies to understand the magnitude of the effects of clouds on SR derived from various AC techniques, under different cloud cover conditions. Overall, this study can inform the choice of SR bands for agricultural assessment and monitoring, which is critical for developing agricultural monitoring systems based on operational Earth observation.

1 Introduction

Agricultural monitoring using remote sensing has become a viable alternative to traditional, field-intensive methods (such as surveying) that have limited spatial coverage and repeatability. In particular, medium-to-high resolution sensors (i.e., such as Landsat-8/9 Operational Land Imager [OLI] and Sentinel-2 Multi-Spectral Instrument [MSI]) provide a near-global, long-term and up-to-date record of data that has been demonstrated for crop type mapping (Xiong et al., 2017; Belgiu & Csillik, 2018), yield estimation (Battude et al., 2016; Vuolo et al. 2016), agricultural water use and non-destructive estimation of biophysical and biochemical parameters (Atzberger & Richter, 2012; Clevers & Gitelson, 2013). However, such capability does not translate to rapid extraction of agricultural information such as crop statistics, crop health, and field conditions. Rather, such information is impeded by low processing levels (i.e., mostly level 1, radiometrically calibrated and geo-rectified data) and lack of computational resources for further processing (such as atmospheric correction and extraction of crop health indicators) on which time series analysis, classification, and regression models hinges.

Correction of Top-of-Atmosphere (TOA) radiometric measurements to surface reflectance (SR) must be achieved by removal of atmospheric effects caused by atmospheric gases and aerosols such as haze, clouds, and cloud shadow contamination since these affect the quality and utility of acquired images. This is inarguably the most fundamental, yet difficult task in optical remote sensing; yet it is critical for an accurate estimation of crop parameters (de Keukelaere et al., 2018). The reduced quality of optical satellite images impedes the accuracy and reliability of derived agricultural information and is a key factor limiting the full use of satellite data for agricultural applications (Giuliani et al., 2017). Moreover, the delay introduced by AC renders subsequent information extraction steps (such as biophysical and biochemical parameter retrieval) obsolete because it is too delayed to be useful for farm management decisions (Atzberger, 2013).

Moreover, atmospheric conditions such as clouds limit the number of observations available to support timely agricultural management decisions, rendering the optical satellite images inappropriate owing to a delay of >5 days (Atzberger, 2013). For example, cloud cover in images acquired during the rainy season may exceed 90%, thus entirely obscuring the landscape during that period. To avoid this limitation, clear-sky pixels in partly cloudy images are used to increase the number of useful observations for quantitative analysis through cloud screening and removal algorithms, and subsequently, gap filling, and pixel compositing (Frantz, 2019; Griffiths et al., 2019; Qiu et al., 2019). However, the radiometric integrity of the cloud-free pixels in partly cloudy images has received little research attention. Such study is invaluable since the atmospheric gases and aerosols affect: (1) the accuracy, integrity, and consistency of optical radiometric measurements; (2) the estimation of biophysical parameters such as Leaf Area Index; (3) image classification accuracy, and (4) reliability of downstream applications (Vanonckelen et al., 2013; López-Serrano et al., 2016; Kganyago et al., 2020). Moreover, aerosols are known to enhance the atmospheric backscattering signal and cause attenuation of the surface directional reflectance signal (Hilker et al., 2009). Ohde (2013) determined the sensitivities of chlorophyll-a concentration estimation methods to atmospheric dust and clouds. Their results showed that while dusty skies caused an overestimation of >8% in chlorophyll-a concentration, atmospheric conditions consisting of mixtures of clouds and dust resulted in overestimations of 7%–14%.

To eliminate the burden of pre-processing, the Committee on Earth Observation Satellites introduced the Analysis-Ready-Data (ARD) concept, which refers to harmonised, standardised, interoperable, and radiometrically and geometrically consistent data to allow for immediate analysis (Killough, 2016; Dwyer et al., 2018; Frantz, 2019). ARD are critical for the research and development of applications that provide up-to-date, actionable information (i.e., accurate, timely, and relevant information useful for decision-making). This is necessary for supporting global food security efforts such as Sustainable Development Goals (SDGs), and national-wide planning, policy, and decision-making for enhanced agricultural productivity and sustainable land management (Anderson et al. 2017; Kganyago & Mhangara 2019a). Therefore, major satellite data providers are increasingly adopting SR products as the standard ARD (Giuliani et al. 2017). For example, at the time of writing this manuscript, the U.S. Geological Survey (USGS) provides Landsat SR products based on Land Surface Reflectance

Code, LaSRC (Claverie et al., 2018) and Landsat Ecosystem Disturbance Adaptive Processing System, LEDAPS (Gao et al., 2006) approaches for Landsat-8 and Landsat-5 to 7, respectively. Similarly, the European Space Agency (ESA) has made Sentinel-2 SR products based on Sen2Cor available (Louis et al., 2016; Djamai & Fernandes, 2018). Therefore, LaSRC, LEDAPS, and Sen2Cor have become the default and current state-of-the-art AC techniques for Landsat and Sentinel-2 satellite images, respectively.

However, since these AC techniques are globally parameterised, it is questionable if they produce accurate estimations of SR that are locally relevant. Djamai et al. (2018) found that there were differences between the Sen2Cor-derived SR product and the ESA SR product according to vegetation cover density, topography, and spectral bands, as well as proportional differences in derived biophysical parameters. A variety of AC techniques exist, offering opportunities for increased radiometric accuracy and consistency as well as image quality improvement, particularly for popular sensors like Sentinel-2. However, the effectiveness of various AC techniques under partly cloudy conditions is unknown. Previous studies validated SR derived from various AC techniques under clear-sky conditions (Djamai & Fernandes, 2018; Doxani et al., 2018; Sola et al., 2018), with only a few focusing on the SR derived from partly cloudy images (Cahalan et al. 2001). In addition, the performance of these AC techniques in semi-arid agricultural areas is still unclear, with studies only rarely focusing on inter-comparison of AC tools across different crop types. For example, recent studies focused on inter-comparison over inland and coastal waters (de Keukelaere et al., 2018; Ilori, Pahlevan & Knudby, 2019; Pereira-Sandoval et al., 2019) and broad land cover categories (Doxani et al., 2018; Sola et al., 2018).

A review of existing literature shows that there is no universal AC approach, and thus the performance of AC approaches vary according to different environments, land cover types, and spectral bands (Doxani et al., 2018; Sola et al., 2018). Achieving ARD therefore necessitates the evaluation of SR products retrieved from various AC tools to determine their accuracy, reliability, and consistency. Moreover, the effect of AC on essential crop biophysical and biochemical parameters has been explored in only a few studies (Miura et al., 2001;

Marcello et al., 2016). It is crucial to evaluate crop parameters from SR and TOA data to address uncertainty regarding the effect of AC on the retrieval of crop parameters and to inform future agronomic applications. While it is typical to perform AC before crop parameter retrieval, recent studies demonstrate that TOA reflectance may be sufficient because biophysical parameters such as LAI are dominant at-sensor in many spectral regions (Pipia et al., 2019; Bayat et al., 2020).

The objectives of this study were: (1) to validate Sentinel-2 MSI SR products retrieved from Sen2Cor and *i*Cor using in situ reference SR under partly cloudy conditions; (2) to determine consistency between SR products from Sen2Cor and *i*Cor with uncorrected TOA reflectance; and (3) to assess the impact of AC techniques on crop parameter retrieval accuracy. The study was conducted in two semi-arid African agricultural landscapes.

2. Materials and Methods

An overview of the study is provided in **Figure 4-1**.

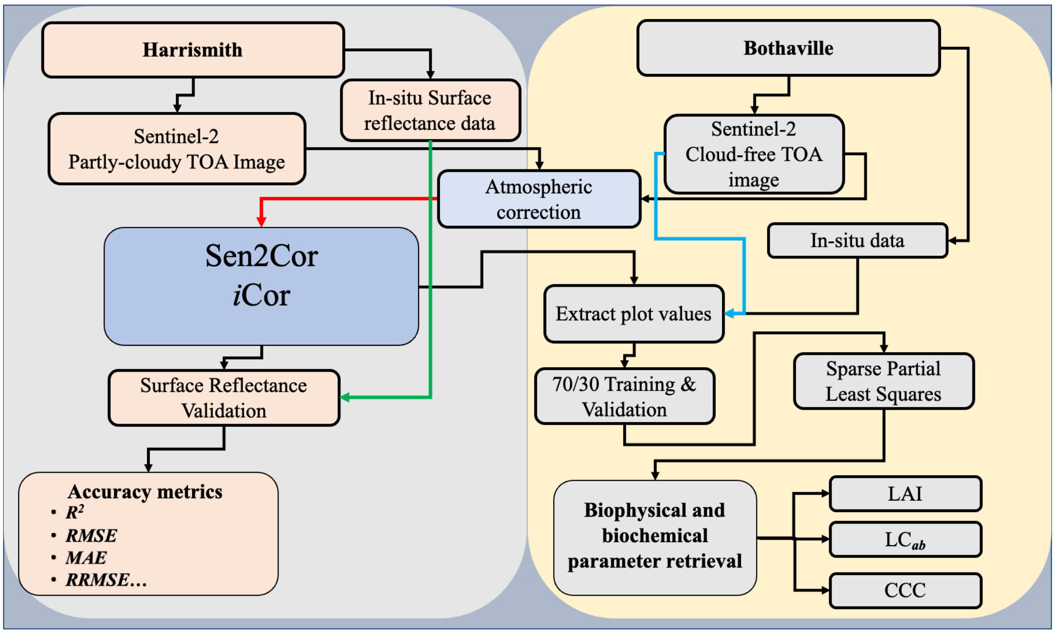


Figure 4-1. Overview of the study design. A partly cloudy image (Harrismith) was used to validate atmospheric correction (AC) techniques with in situ reflectance measurements, while a cloud-free image (Bothaville) was used for inter-comparison of AC techniques and to assess their impact on crop parameter retrieval.

2.1. Study area

Bothaville and Harrismith, located in the Free State province, South Africa constituted the study area for the current study (**Figure 4-2**). The province is one of the main agricultural production zones of South Africa, with both commercial large-scale and small-holder farming systems. The two areas are also part of the Group on Earth Observations-Global Agricultural Monitoring (GEOGLAM) initiative's Joint Experiment for Crop Assessment and Monitoring (JECAM) and recently, H2020-AfriCultuReS project (Enhancing Food Security in African AgriCultural Systems with the support of Remote Sensing, Grant Agreement No. 774652). The main characteristics of the study area are outlined in **Table 4-1**.

Table 4-1. Characteristics of the study area.

Site characteristics	Bothaville	Harrismith
Mean annual temperature	18.0 °C	19.2 °C
Mean annual rainfall	584 mm	115 mm
Soil types	Loamy to sandy-loamy	Sandy to sandy-loamy
Slope	Flat slope	Undulating slopes
Main crop types	Maize, beans, and sunflower	Maize and beans
Crop calendar	December to May/June	December to May/June

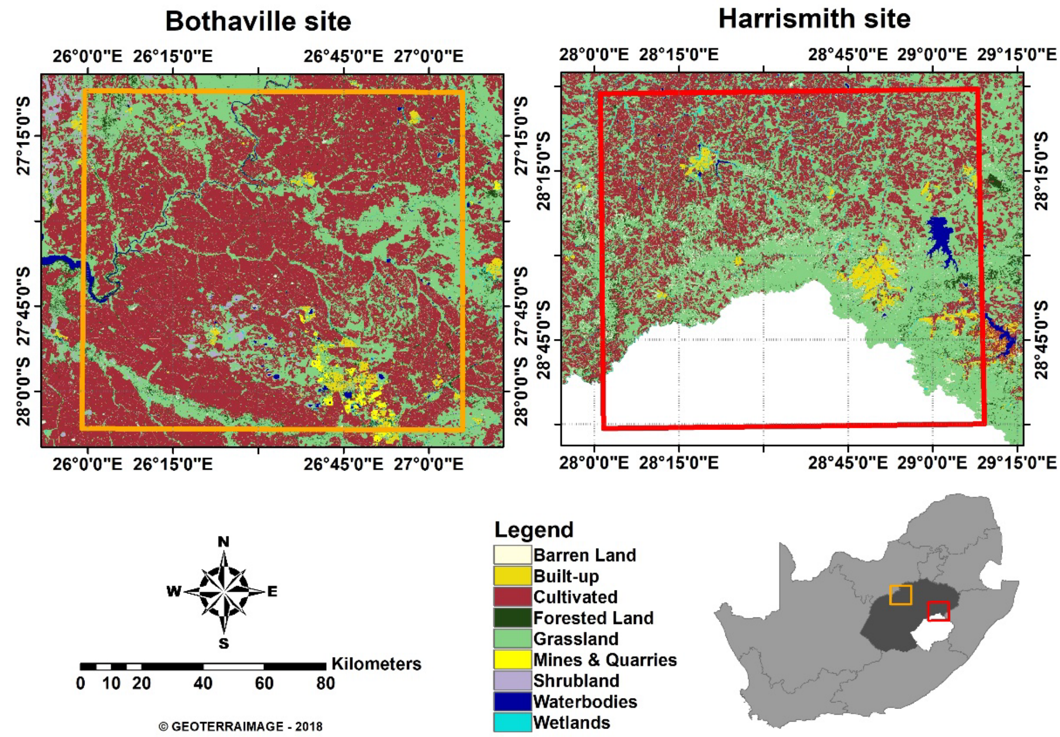


Figure 4-2. Study area maps showing the extent of; and main land cover types at Bothaville (left) and Harrismith (right).

2.2. Data

2.2.1. Satellite data

Sentinel-2 MSI (TOA, also commonly referred to as level-1C) images acquired on 18 March 2019 (Harrismith) and 14 April 2021 (Bothaville), were retrieved from Copernicus Open Access Hub (https://scihub.copernicus.eu/dhus/). The properties of the MSI images are described in **Table 4-2.** The image acquired on 18 March 2019 over Harrismith was used for validation of SR products, as it was the only usable image closest (i.e., ±3 days) to the dates of in situ spectral observations. Apart from this, cloud-contaminated images are common during the rainy season, and therefore the performance of AC approaches in such conditions should inform the utility of derived SR for agricultural monitoring. However, owing to high cloud cover dominating the image of 18 March, i.e., 61.89%, the image acquired on 14 April 2021 over Bothaville was selected for inter-comparison of AC techniques to determine the consistency in derived SR between different AC techniques and TOA reflectance, and to assess the effect of each AC technique on the retrieval of crop parameters.

Table 4-2. Sentinel-2 MSI characteristics. Bold bands are reserved for atmospheric correction (Drusch et al., 2012).

Spectral bands	Band centre	Bandwidth	Spatial resolution
Deep blue (B1)	**443 nm**	**20 nm**	**60 m**
Blue (B2)	490 nm	65 nm	10 m
Green (B3)	560 nm	35 nm	10 m
Red (B4)	665 nm	30 nm	10 m
Red-edge (B5)	705 nm	15 nm	20 m
Red-edge (B6)	740 nm	15 nm	20 m
Red-edge (B7)	783 nm	20 nm	20 m
NIR1 (B8)	842 nm	115 nm	10 m
NIR2 (B8A)	865 nm	20 nm	20 m
NIR3 (B9)	**945 nm**	**20 nm**	**60 m**
Cirrus (B10)	**1 375 nm**	**30 nm**	**60 m**
SWIR1 (B11)	1 610 nm	90 nm	20 m
SWIR2 (B12)	2 190 nm	180 nm	20 m

2.2.2. Ancillary data

Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) with approximately 90 m spatial resolution was obtained from the Consortium for Spatial Information (CGIAR-CSI, http://srtm.csi.cgiar.org /srtmdata/). The SRTM DEM is considered one of the most accurate, freely available global sources of digital elevation data. For this reason, it is also used as a pre-loaded digital elevation dataset in relevant applications by SNAP software, which is developed by the European Space Agency for processing images derived by the Sentinel satellites. To keep consistency among AC techniques, SRTM 90 m DEM (Version 4.1) was used for all AC techniques for cloud shadow detection and topographic correction. The DEM was warped to the extent and resolution of the image being processed using bilinear resampling.

2.2.3. In situ data

2.2.3.1. Reference surface reflectance data

The spectral reflectance samples were collected from 11 to 15 March 2019 from plots of 40 m × 40 m using the Spectral Evolution PSR-3500 Spectrometer (Spectral Evolution, Inc. © 2014) with a spectral range of 350–2 500 nm. The spectrometer had ~3.5 nm spectral resolution at 350–1 000 nm, 10 nm at 1 500 nm and 7 nm at

2 100 nm. The spectral bands from 350–1 000 nm, at 1 500 nm and at 2 100 nm have nominal spectral sampling intervals of 1.5 nm, 3.8 nm, and 2.5 nm, respectively. For each 40 m × 40 m plot, GPS coordinates were collected at the centre using handheld GPS receiver, i.e., Spectra Precision MobileMapper® 120 with <50 cm accuracy. Canopy reflectance measurements were then collected by consistently holding a fibre optic cable (i.e., 25° field-of-view) at nadir angle and an observation distance of 0.5 m above the canopy. This resulted in an instantaneous field-of-view (IFOV) with a radius of approximately 11.08 cm. Therefore, each plot was represented by an average of nine homogeneous spectral measurements taken at distinct random locations, separated by an approximate distance of 10 m. This ensured that the spectral measurements were a representative mixture of reflectance as determined by the proportion, physical arrangement, and reflective properties of plants (Clark & Roush, 1984).

All reflectance measurements were taken between 10:00 and 14:00 Central African Time and normalised by taking a white reference panel reflectance before every measurement (Bajwa et al., 2010; Mirzaie et al., 2014). The number of plots and spectral observations per crop type (**Table 4-3)** was limited by suitable atmospheric conditions, i.e., cloudless and short time available for spectral measurements, i.e., 4 hours. The in situ reference SR measurements were resampled to Sentinel-2 band specifications based on its spectral response functions (S2-SRF version 3.0., released December 2017) as implemented in hsdar R-package (version 1.0.0., released May 2019).

Table 4-3. The number of plots and spectral reflectance measurements per crop type in Harrismith. These measurements were used for validation of surface reflectance (SR) derived by AC techniques.

Crop types	No. of plots	No. of observations
Maize	5	45
Dry beans	5	45
Soya beans	3	27
Total	13	117

2.2.3.2. Inter-comparison and crop parameter modelling data

The data for inter-comparison and modelling of crop parameters were collected between 11 and 23 April 2021 in Bothaville over maize and beans (**Table 4-4**). In situ measurements of LAI and LC_{ab} were collected within 40 m × 40 m plots. Plots were systematically selected over a transect to capture variability, and each plot had an average of six to eight random readings of the relevant parameter. LI-COR 2200c Plant Canopy Analyzer (LI-COR, Inc., Lincoln, NE, USA) was used to record LAI, while MC-100 Chlorophyll Concentration Meter (Apogee Instruments, Inc., Logan, UT, USA) was used to capture LC_{ab} from the sunlit leaves of the upper canopy. An 180° view cap was used to cover the lens of the Plant Canopy Analyzer to reduce operator interference and the impact of uneven sky conditions on the measurements. The centroid of each plot was geo-tagged using a Trimble® TDC600 handheld Data Collector with global navigation satellite systems (GNSS) accuracy of 1.5 m. Canopy Chlorophyll Content (CCC), the third crop parameter of interest in the current study, was calculated as the product of LC_{ab} and LAI (LC_{ab} × LAI) (Jacquemoud et al. 1995). The data were divided into 70% training and 30% validation (**Table 4-4**).

Table 4-4. Training (70%) and validation (30%) data over maize and beans in Bothaville.

	Datasets	*n*	Min	Mean	Max	SD
LAI	Training	113	1.78	3.35	5.57	0.86
	Validation	48	2.02	3.60	5.75	1
LC_{ab}	Training	113	4.06	33.87	66.18	14.93
	Validation	48	3.69	32.62	70.69	19.20
CCC	Training	113	10.30	105.69	288.22	61.60
	Validation	48	7.87	116.42	339.10	90.39

The GPS coordinates were used to extract SR values from each corrected image (i.e., over Bothaville) by Sen2Cor and *i*Cor as well as the TOA reflectance data, for the purpose of inter-comparison and modelling of crop parameters.

2.3. Atmospheric correction techniques

The level-1C Sentinel-2 MSI images were subjected to Sen2Cor and *i*Cor to derive SR. The two AC techniques were chosen because they are both image-based techniques (i.e., require no additional ancillary data to approximate atmospheric conditions) and are easily accessible on a free and open graphical user-interface, i.e., SNAP (*SNAP-ESA Sentinel Application Platform v8.0,* http://step.esa.int).

2.3.1. Sen2Cor

Sen2cor (version 2.8, released February 2019) is a processor for atmospheric correction (including cirrus clouds and terrain correction) of Sentinel-2 data, available as a stand-alone command-line tool, as well as a SNAP plugin. The algorithm is based on ATCOR model and converts the Level-1C TOA image data to SR reflectance based on the libRadtran database of LUTs generated for a wide variety of atmospheric conditions, solar geometries, and ground elevations. Scene classification is also performed as part of the processing to flag clouds, snow, and cloud shadows and produces a classification map indicating areas of vegetation, soils/deserts, water, snow, clouds, and cloud shadows based on a series of thresholds from band ratios and spectral indices. Further details can be found in Louis et al. (2016).

2.3.2. iCor

*i*Cor (version 2.0, released February 2019), previously known as OPERA (Sterckx et al., 2015), is an image-based AC algorithm available as a SNAP plugin that supports multiple sensors, including Sentinel-2 MSI, Sentinel-3 and Landsat-8 OLI. *i*Cor processing includes: (1) identification of land and water pixels; (2) AOT retrieval based on spectral variability of land pixels following an adapted version of the method by Guanter Palomar (2007); (3) an adjacency correction performed using SIMEC (Sterckx et al., 2015) over water and fixed background range over land targets; and (4) the Radiative Transfer Model, MODTRAN 5 (Berk et al., 2006) LUT is used to perform the AC, which requires input information about the solar and viewing angles and a DEM. Further details can be found in De Keukelaere et al. (2018).

2.4. Bidirectional Reflectance Distribution Function correction

In order to account for variations in reflectance caused by different solar and viewing geometries between the in situ reference SR and Sentinel-2 SR, the Bidirectional Reflectance Distribution Function (BRDF) effects were corrected using c-factor approach (Roy et al. 2017). This approach is based on a semi-empirical RossThick/LiSparse-Reciprocal BRDF model (Schaaf et al. 2002) that uses a set of fixed BRDF spectral model parameters derived by averaging over 15 billion pixels of a global year of all the highest quality, snow-free MODIS 500 m BRDF product (MCD43). SR products derived using various AC approaches were normalised to nadir-view, i.e., Nadir BRDF-Adjusted surface Reflectance (NBAR) following Equation 4-1 below:

$$NBAR_{\lambda} = (\theta_v = 0, \theta_s = k) = c_{\lambda} \times \rho_{\lambda}(\theta_v = \theta_v^{Sentinel}, \theta_s = \theta_s^{Sentinel}) \tag{4-1}$$

$$c_{\lambda} = \frac{\hat{\rho}_{\lambda}(\theta_v = 0, \theta_s = k)}{\hat{\rho}_{\lambda}(\theta_v = \theta_v^{Sentinel}, \theta_s = \theta_s^{Sentinel})}$$

Where ρ_{λ} is the Sentinel-2 SR with view zenith ($\theta_v^{Sentinel}$) and solar zenith ($\theta_s^{Sentinel}$) for wavelength λ, and $NBAR_{\lambda}$ is the nadir BRDF-adjusted SR equivalent estimated for solar zenith angle ($\theta_s = k$) for a nadir-view zenith ($\theta_v = 0$). The $\hat{\rho}_{\lambda}$ values were estimated from spectrally equivalent MODIS bands (Roy et al., 2017) using Ross Thick/LiSparse-Reciprocal BRDF model (Schaaf et al., 2002), and from interpolated MODIS spectral BRDF model parameters for red-edge bands (Roy et al., 2017).

2.5. Biophysical and biochemical parameters retrieval with sparse Partial Least Squares

The retrieval of biophysical and biochemical parameters from Sen2Cor, *i*COr, and TOA reflectance data was performed with sparse Partial Least Squares (sPLS) (Lê Cao et al. 2008). The PLS-singular value decomposition (PLS-SVD) is the basis of the sPLS, which has several advantages over conventional multivariate regression algorithms. It can, for instance, be used as a variable selection algorithm and a regression algorithm. As a result, it has been widely utilised in previous studies to estimate agricultural productivity (Abdel-Rahman et al., 2014), grass biomass (Sibanda et al., 2017), and LAI (Sibanda et al., 2019).

2.6. Accuracy metrics

Approaches using AC were validated against in situ reference SR using statistical metrics (i.e., coefficient of determination, R^2, and *p*-value) based on a fitted line of Ordinary Least Squares regression (OLS) equation between in situ reference SR and derived SR from each approach. Additional statistical error metrics were used for error assessments, i.e., Root Mean Squared Error (RMSE), Mean Absolute Error

(MAE), BIAS and Mean Absolute Percent Error (MAPE). These are given in Equations 4-2 to 4-5.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{4-2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{4-3}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - y_i|}{y_i} \times 100 \tag{4-4}$$

$$BIAS = \sum_{i=1}^{n}(x_i - y_i) \tag{4-5}$$

Where x_i is the retrieved surface reflectance, and y_i is the observed surface reflectance, N is the sample size, and n is the number of errors.

The inter-comparison of Sen2Cor and *i*Cor was performed with R^2, RMSE, RRMSE (Equation 4-6), and Spectral discrimination index (SDI, Equation 4-7).

$$RRMSE = \frac{RMSE}{\bar{x}_i} \tag{4-6}$$

$$SDI = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_1} \tag{4-7}$$

Where μ_1 and μ_2 are the mean and standard deviations of SR derived by Sen2Cor and *i*Cor, respectively. SDI below 1 and above 1 indicate significant difference and no spectral difference between the two AC techniques.

3. Results and Discussion

To quantitatively analyse remotely sensed data so that crop biophysical and biochemical parameters for Precision Agriculture and crop monitoring are retrieved, it is vital first to perform AC. This study assessed the effectiveness of Sen2Cor and *i*Cor in deriving SR under partly cloudy conditions by comparing the derived SR with in situ surface reflectance and evaluated the radiometric consistency of SR products derived from the two AC approaches and the uncorrected TOA reflectance. Moreover, the impact of AC approaches on the retrieval of essential crop biophysical and biochemical parameters was assessed.

3.1. Validation of surface reflectance from Sen2Cor and *i*Cor

Generally, SR products from Sen2Cor and *i*Cor show consistently low and insignificant overall (i.e., combined) relationship with in situ reference SR in the VIS (i.e., B2–B4) and SWIR regions (i.e., B11 and B12) ($p > 0.05$). However, an analysis per crop type reveals better relationships (R^2) that were masked in the overall analysis. The results show consistently better fit, i.e., $R^2 > 0.7$ in B4 for SR over dry beans and consistently moderate fit, i.e., R^2 ranging from ~0.3 to ~0.5 in B2, B3, B5 and B12 for the two AC techniques, except for B5 SR derived from *i*Cor that showed no relationship, i.e., $R^2 < 0.2$ (**Figure 4-4**). In contrast, SR over soya beans has a consistently moderate fit, i.e., R^2 ~0.3 to ~0.5 in B2 and B12 for the two AC techniques, except for *i*Cor which had lower fit with in situ reference SR in B12. However, *i*Cor SR over soya beans show a considerably better fit, i.e., R^2 ~0.67 in B11 and a significant R^2 of 0.99 ($p = 0.03$) in B5.

The results show that the overall relationship improves in the Red-edge (i.e., B6–B7) and NIR (i.e., B8 and B8A) bands, except for B5, which exhibited lower fit with in situ reference SR. Specifically, *i*Cor SR in these bands showed a better fit, i.e. R^2 from ~0.6 to ~0.62, while Sen2Cor showed a relatively lower but significant fit, i.e. R^2 of ~0.6 to 0.61. The validation results by crop type showed consistently better relationship over dry beans, i.e., R^2 >= 0.7 for both AC techniques, with exception of B6 SR derived from *i*Cor that had moderate R^2 of ~0.35. SR over soya beans showed consistently high negative relationships with in situ reference SR, i.e., R^2 ranging from 0.75 to 0.99 in B6 for the two AC techniques. On the other hand, SR over maize had consistently moderate R^2 of between 0.5 and 0.57, i.e., highest in B6 and lowest in B8A, respectively, with the exception of B6 SR derived from *i*Cor that shows no relationship, i.e., $R^2 < 0.2$.

The results in this study indicating a lower fit with in situ reference SR in the Sentinel-2 VIS region (i.e. B2–B4) and a better fit in the RE (i.e. B6 and B7) and NIR region (i.e. B8 and B8A) are expected since NIR region has low aerosol scattering effect than the VIS region (Li et al. 2018). However, these results are inconsistent with previous studies (Sola et al. 2018) that found a better fit in the VIS bands for *i*Cor and Sen2Cor and a relatively lower fit in the RE and NIR bands. These inconsistencies and significantly higher R^2 values, i.e., R^2 >0.9 across all bands and techniques found in previous studies (Sola et al. 2018) can be attributed to the availability of almost simultaneous in situ reference SR data with Sentinel-2 cloud-free overpasses, a higher number of samples, i.e. ~200 spectra per plot, and more repetitions of in situ spectral measurements. Although the time gap between-field spectral measurements and satellite overpass, i.e., ±3 days, as well as high cloud cover, i.e., 61.89%, were a limitation in this study, the results address the knowledge gaps about the effectiveness of various AC approaches in partly cloudy atmospheric conditions. Such conditions are common during the rainy season; thus the results found here have implications for the choice of AC technique and radiometrically accurate spectral bands for agricultural monitoring applications.

Sentinel-2 MSI VIS B2–B4 exhibited lower R^2, i.e. <0.2, across all AC techniques, which can be attributed to the high presence of clouds in the image. Clouds enhance apparent path radiance, cause an overestimation of aerosol optical thickness (AOT), and as a result, an underestimation of surface reflectance in the VIS region (Wen et al., 2001). The underestimation of SR by both AC techniques can be seen in the scatterplots, i.e., **Figure 4-3** and **Figure 4-4**. Although both AC techniques can address adjacency effects, residual adjacency effects may persist especially for pixels near clouds (Marshak et al., 2008). Therefore, lower R^2 and relatively higher errors (i.e., RMSE, MAE, and BIAS), found across all bands, AC techniques, and crop types, may be a result of such residual effects. However, the effectiveness of adjacency correction in cloudy scenes as implemented in various AC techniques need to be evaluated in future studies. In addition to presenting challenges for AC techniques, clouds also affect the usability of images. After all, further validation is needed since the errors may have resulted from the sensitivity of the RE and NIR regions towards the rapid changes agricultural landscapes that may have occurred during the ±3-day difference between the date of acquisition of the Sentinel-2 image and in situ reference SR, despite stable conditions in the course of data collection for this study.

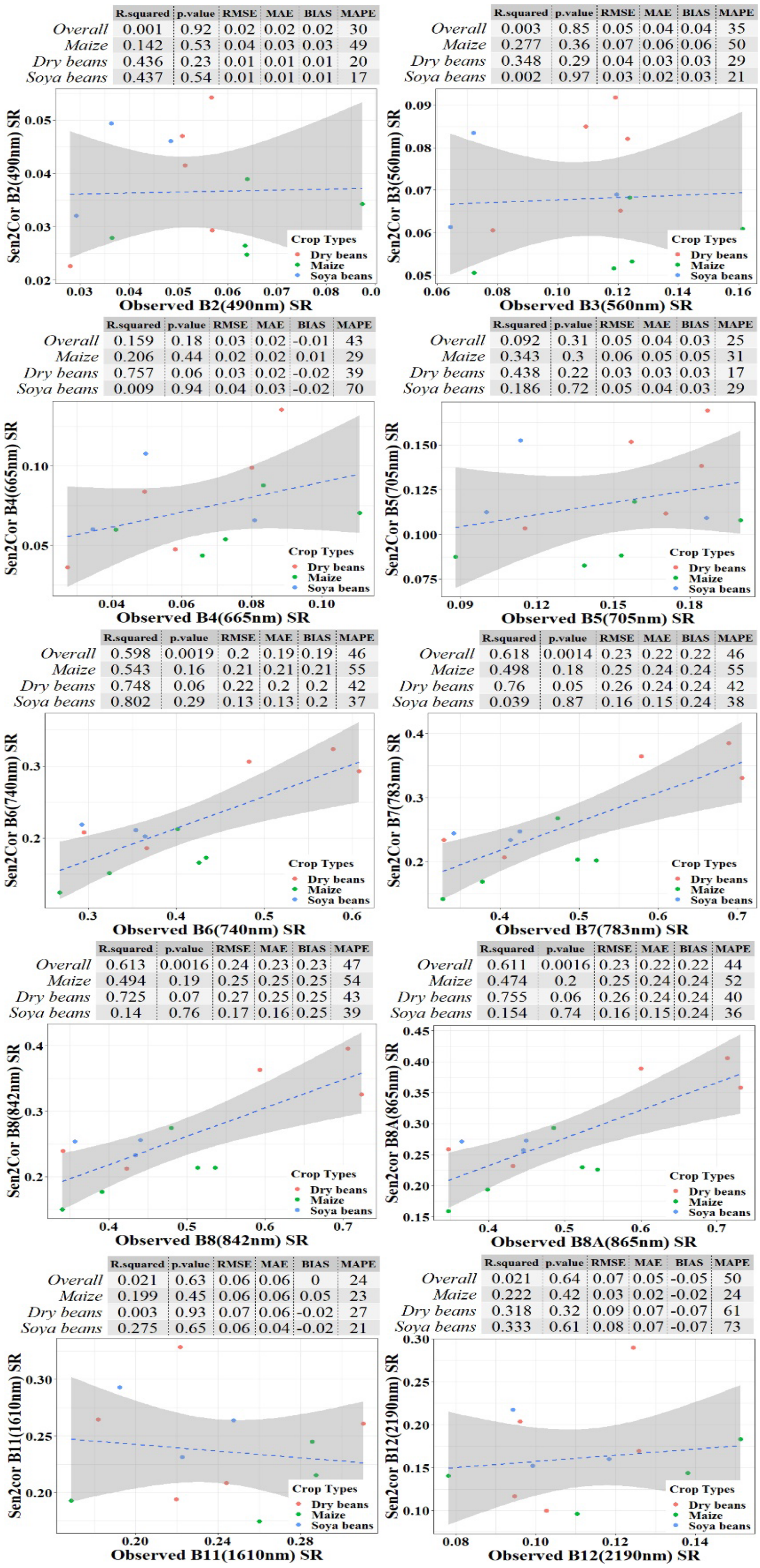
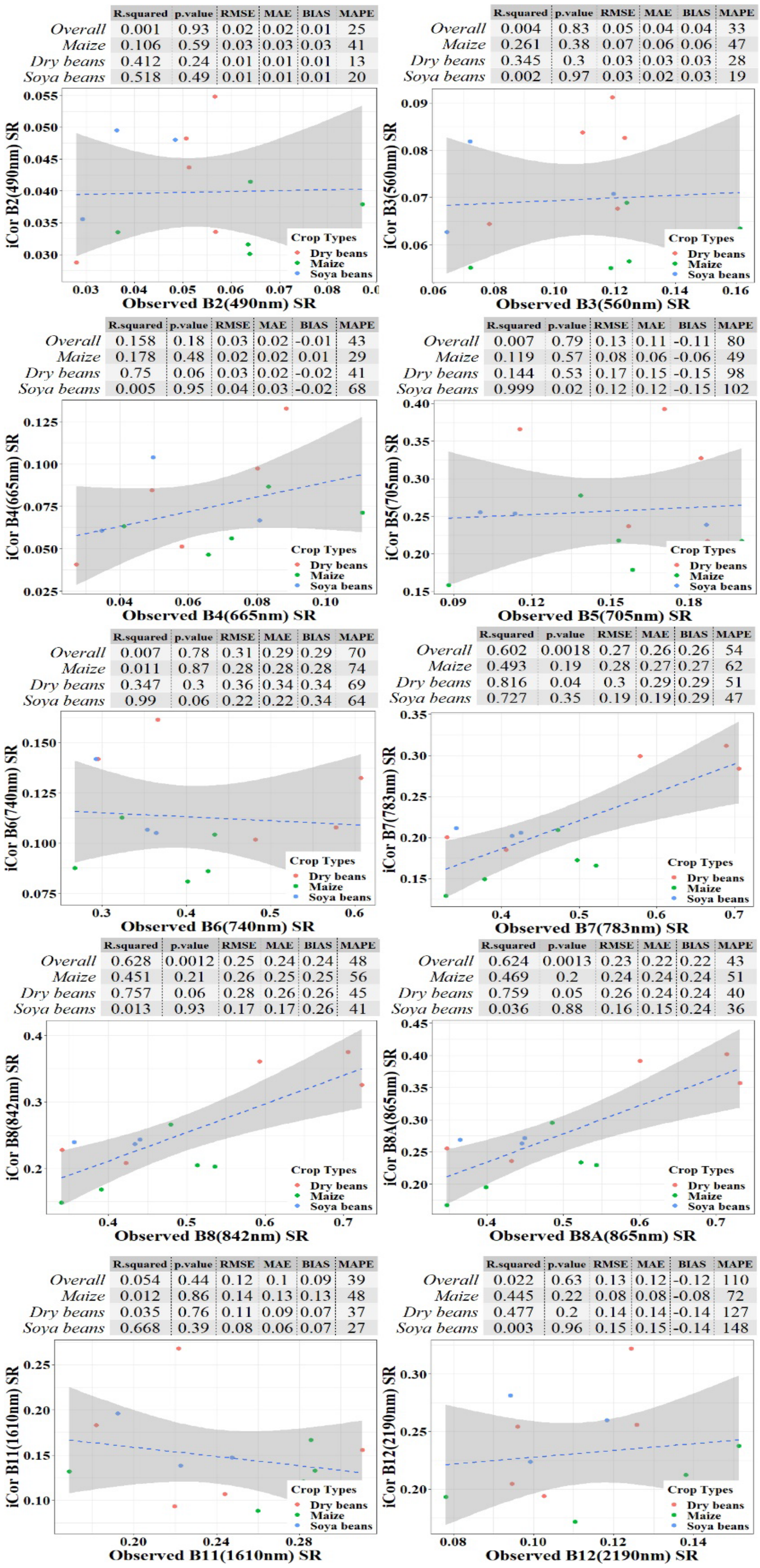


Figure 4-3. Band-wise scatterplot with associated validation metrics for SR derived from Sen2Cor. OLS regression metrics, i.e., R^2 and p-values as well as error metrics, i.e., RMSE, MAE, BIAS, and MAPE are given at the top of each plot.



	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.001	0.93	0.02	0.02	0.01	25
Maize	0.106	0.59	0.03	0.03	0.03	41
Dry beans	0.412	0.24	0.01	0.01	0.01	13
Soya beans	0.518	0.49	0.01	0.01	0.01	20

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.004	0.83	0.05	0.04	0.04	33
Maize	0.261	0.38	0.07	0.06	0.06	47
Dry beans	0.345	0.3	0.03	0.03	0.03	28
Soya beans	0.002	0.97	0.03	0.02	0.03	19

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.158	0.18	0.03	0.02	-0.01	43
Maize	0.178	0.48	0.02	0.02	0.01	29
Dry beans	0.75	0.06	0.03	0.02	-0.02	41
Soya beans	0.005	0.95	0.04	0.03	-0.02	68

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.007	0.79	0.13	0.11	-0.11	80
Maize	0.119	0.57	0.08	0.06	-0.06	49
Dry beans	0.144	0.53	0.17	0.15	-0.15	98
Soya beans	0.999	0.02	0.12	0.12	-0.15	102

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.007	0.78	0.31	0.29	0.29	70
Maize	0.011	0.87	0.28	0.28	0.28	74
Dry beans	0.347	0.3	0.36	0.34	0.34	69
Soya beans	0.99	0.06	0.22	0.22	0.34	64

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.602	0.0018	0.27	0.26	0.26	54
Maize	0.493	0.19	0.28	0.27	0.27	62
Dry beans	0.816	0.04	0.3	0.29	0.29	51
Soya beans	0.727	0.35	0.19	0.19	0.29	47

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.628	0.0012	0.25	0.24	0.24	48
Maize	0.451	0.21	0.26	0.25	0.25	56
Dry beans	0.757	0.06	0.28	0.26	0.26	45
Soya beans	0.013	0.93	0.17	0.17	0.26	41

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.624	0.0013	0.23	0.22	0.22	43
Maize	0.469	0.2	0.24	0.24	0.24	51
Dry beans	0.759	0.05	0.26	0.24	0.24	40
Soya beans	0.036	0.88	0.16	0.15	0.24	36

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.054	0.44	0.12	0.1	0.09	39
Maize	0.012	0.86	0.14	0.13	0.13	48
Dry beans	0.035	0.76	0.11	0.09	0.07	37
Soya beans	0.668	0.39	0.08	0.06	0.07	27

	R.squared	p.value	RMSE	MAE	BIAS	MAPE
Overall	0.022	0.63	0.13	0.12	-0.12	110
Maize	0.445	0.22	0.08	0.08	-0.08	72
Dry beans	0.477	0.2	0.14	0.14	-0.14	127
Soya beans	0.003	0.96	0.15	0.15	-0.14	148

Figure 4-4. Band-wise scatterplot with associated validation metrics for derived SR from *i*Cor. OLS regression metrics, i.e. R^2 and p-values as well as error metrics, i.e. RMSE, MAE, BIAS, and MAPE are given at the top of each plot.

3.2. Consistency of Sen2Cor and *i*Cor and comparison with uncorrected (TOA) data

The results (**Figure 4-5**) show 1:1 relationship (R^2 of 100%) between the SR derived from Sen2Cor and the uncorrected (i.e., TOA) data in all MSI spectral bands and crop types (i.e., maize and beans). The RRMSE were below 2% for overall and both crop types, except for maize in B2–B4 which exhibited RRMSE values around 3% to 5%. For *i*Cor (**Figure 4-6**), the same 1:1 relationship with TOA data for overall and crop-specific metrics was observed in B11 and B12 only (R^2 of 100%, RRMSE < 1%). As with results in **Figure 4-5**, maize reflectance in B2–B4 had higher RRMSE of around 3% to 6% and R^2 ranging from 64% in B2 to 94% in B4. Overall (i.e., combined) and beans R^2 values in these bands were also not perfect, ranging from 96% to 99% and 86% to 96% for the overall and beans metrics. The less than perfect relationship with TOA data in B2–B4 indicates some effectiveness of *i*Cor over Sen2Cor in correcting for high aerosol scattering effect known in the visible portion of the electromagnetic spectrum (Li et al. 2018). In the red-edge region, particularly B5 and B6, some interesting results can be observed, where significantly poor relationships can be observed for crop-specific metrics, i.e., R^2 of <1% for beans and 12% to 25% for maize. The regression line for beans showed no relationship between *i*Cor corrected reflectance and uncorrected (TOA) reflectance. In particular, TOA reflectance values were around 0.2 in B5 and scattered from 0.2 to around 0.35 in B6 while the corrected reflectance values by *i*Cor ranged from 0.3 to 0.5 in B5 and from 0.2 to just above 0.25 in B6 for beans.

Although the regression line for maize shows a positive relationship, the reflectance values seem to diverge. In B5, the TOA reflectance values are low (i.e., <0.15), while the corrected reflectance values by *i*Cor are between 0.2 to 0.3. In B6, there seems to be an opposite effect, where the range of TOA reflectance values range from approximately 0.15 to above 0.25 and the corrected values have 0.15 as the maximum. The results, therefore, indicate the effectiveness of *i*Cor in correcting for atmospheric errors in these spectral bands, which seems to increase the dynamic range of the reflectance values, particularly in B5. B5 (705 nm) is often identified as an important variable in the retrieval models of LAI and canopy storage capacity

(Sibanda et al. 2019), grass nutrient estimation (Ramoelo et al. 2012), and grass discrimination (Otunga et al. 2019).

The variations in these spectral bands, i.e., B5 and B6, can also be seen in **Figure 4-7**, where the consistency between Sen2Cor and *i*Cor SR products was evaluated. The two AC techniques were perfectly consistent (i.e., 1:1 relationship) in B2, B11, and B12 across all crop types, while the R^2 in other bands ranged from 86%–99% and varied by crop type. This consistency between Sen2Cor and *i*Cor SR products can be explained by the way AC is carried out. Both techniques retrieve AOT based on dark, dense vegetation (DDV) pixels. The consistency between Sen2Cor and *i*Cor in B11 and B12 offers great prospects for consistent crop parameter retrieval, since studies increasingly show their importance for vegetation monitoring (Adjorlolo et al. 2012; Kganyago et al. 2017; Kganyago et al. 2018). In fact, VIS/NIR and SWIR areas are thought to have a similar association to characterise vegetation conditions, according to Herrmann et al. (2010). Even though water absorption in B11 and B12 and chlorophyll absorption in B2 and B4 are the key factors controlling reflectance in these regions (Jensen 1983), both VIS/NIR and SWIR exhibit relatively low vegetation reflectance.

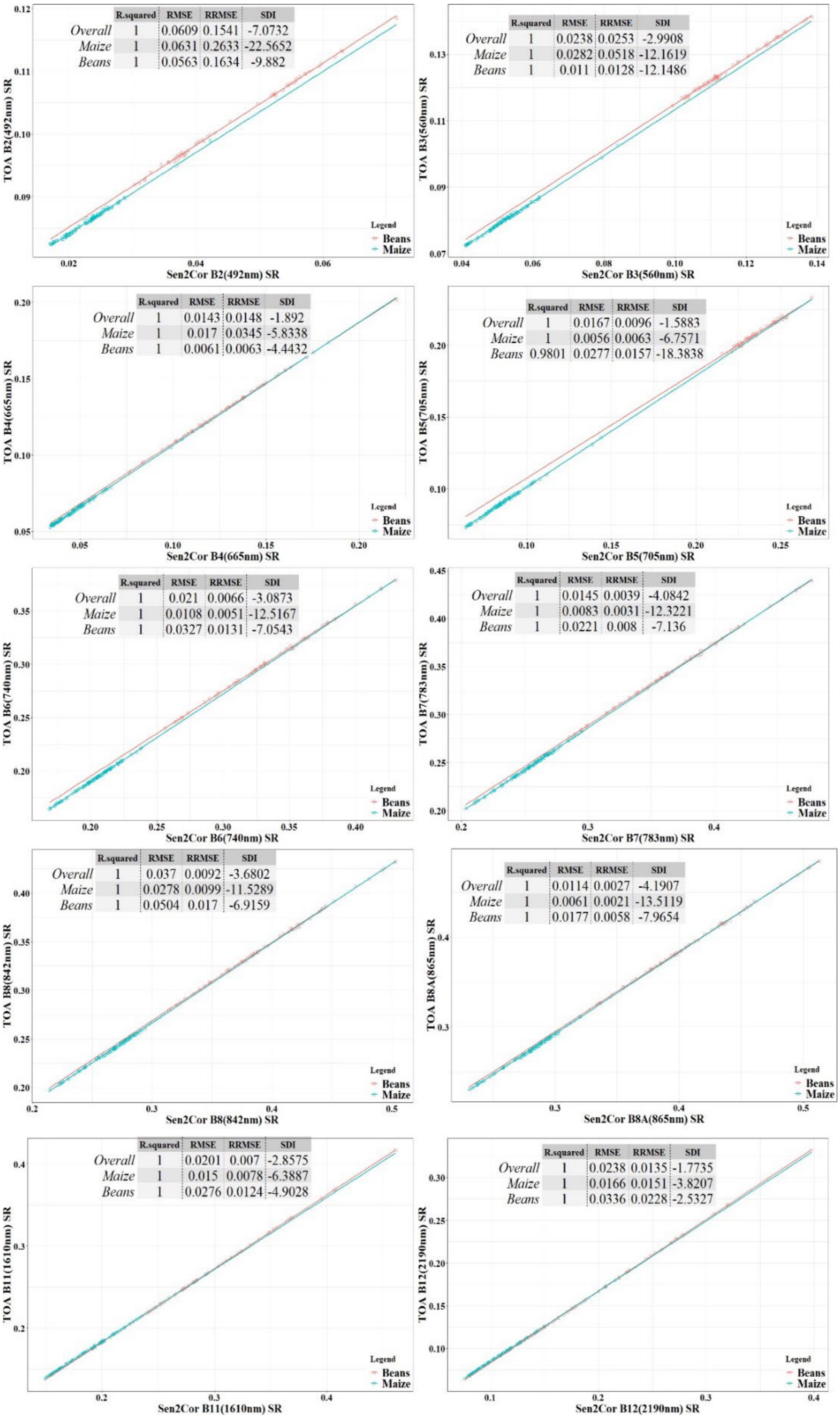


Figure 4-5. Comparison between Sen2Cor surface reflectance and Top-of-Atmosphere (TOA) reflectance by crop type.

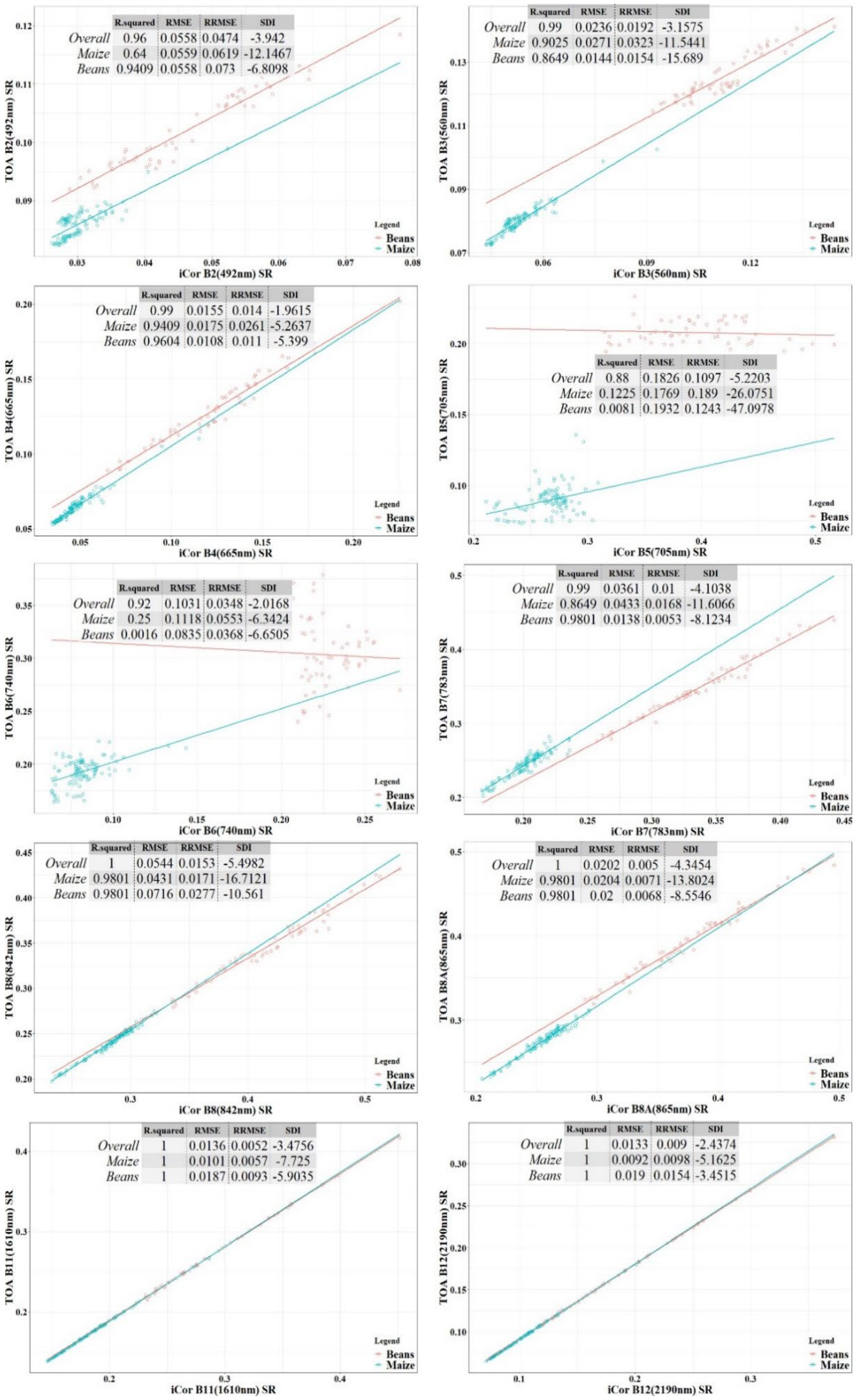


Figure 4-6. Comparison between *i*Cor surface reflectance and Top-of-Atmosphere (TOA) reflectance by crop type.

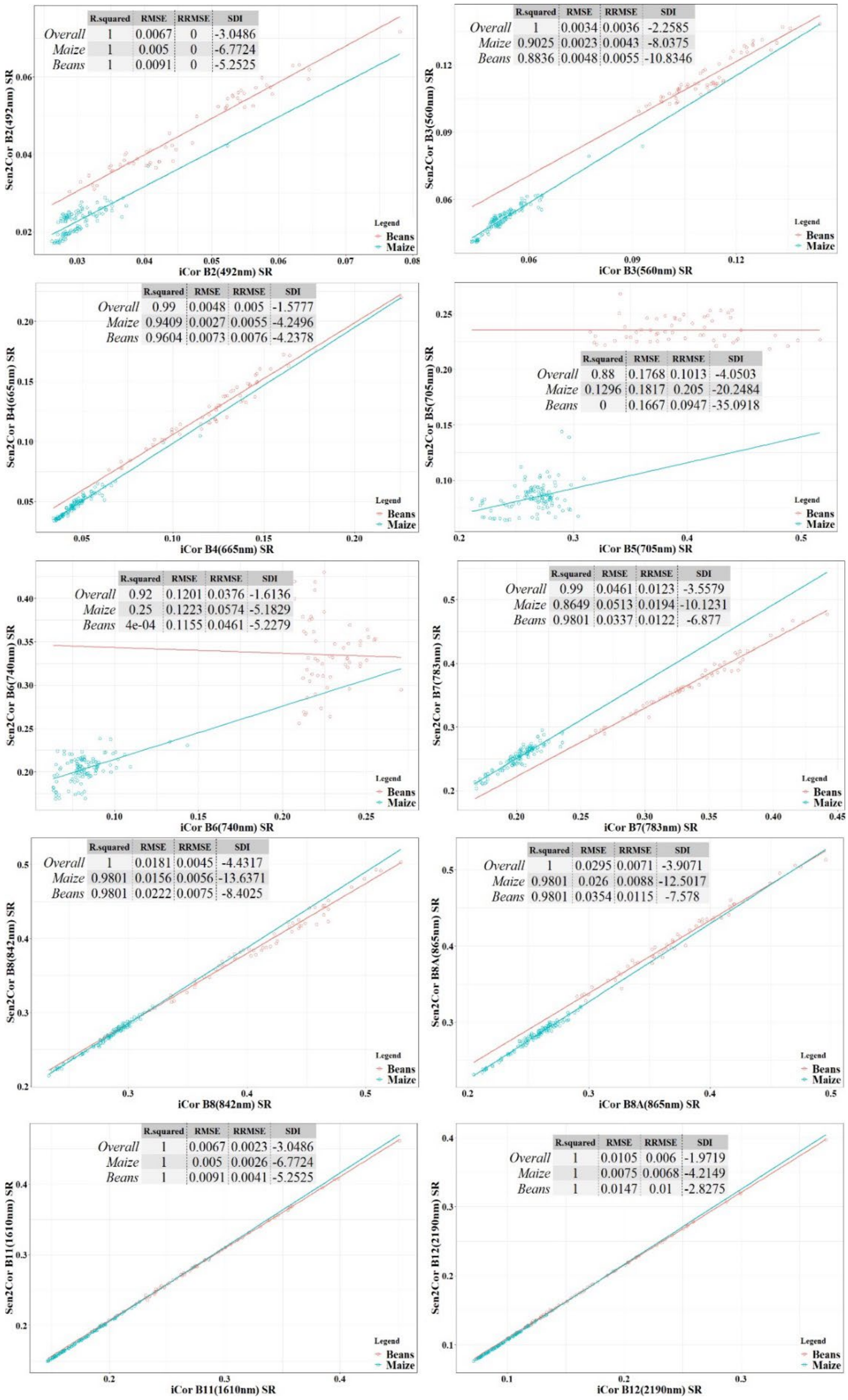


Figure 4-7. Comparison between Sen2Cor and *i*Cor surface reflectance by crop type.

3.3. The impact of AC techniques on crop biophysical and biochemical parameters

Generally, the overall R^2 for LAI were low, i.e., ranging from 28% to 40% (RMSE: 0.66–0.72 $m^2 m^{-2}$) (**Figure 4-8**a–c), when compared with LC_{ab} (R^2: ~76%, RMSE: ~8 $\mu g\ m^{-2}$) (**Figure 4-8**d–f) and CCC (R^2: 62%–67%, RMSE: ~39–41 $\mu g\ m^{-2}$) (**Figure 4-8**g–i). The crop-wise LAI retrieval accuracies also varied, where maize had better accuracies across all AC approaches, i.e., R^2: 15%–36%, RMSE: 0.73–0.84 $m^2 m^{-2}$), while the accuracies for beans were consistently lower, i.e., R^2: 12%. The same patterns can be observed for LC_{ab} and CCC (**Figure 4-8**d–i). The retrievals from Sen2Cor were consistently better than those from *i*Cor for all crop parameters, except for LC_{ab} where equivalent results were obtained, i.e., R^2: 75.69% and RMSE: ~7.95 $\mu g\ m^{-2}$. This can be attributed to differences in atmospheric RTMs used to compute Look-Up-Tables (LUTs) between the two AC techniques. Sen2Cor is based on libRadtran, while *i*Cor is based on MODTRAN 5. As an ill-posed problem, inversion of atmospheric characteristics may induce various errors in atmospheric correction, which vary depending on the AC technique. A comparison of results from Sen2Cor and TOA shows equivalent results for LAI (**Figure 4-8**a and c), while Sen2Cor resulted in slightly better retrievals for LC_{ab} and CCC (**Figure 4-8**d, g, f and i). Only CCC results from *i*Cor were slightly better than those from TOA data (considering a tradeoff between R^2 and RMSE), i.e., R^2: 64% and RMSE: 39.88 $\mu g\ m^{-2}$ (**Figure 4-8**h and i). The similarity between the LAI retrieved from atmospherically corrected data (i.e., SR) and uncorrected data (i.e., TOA) implies that TOA data can be used directly without the AC pre-processing step. This is consistent with Estévez et al. (2020) who found no significant differences between LAI from TOA and SR data using coupled leaf-canopy-atmosphere models and machine learning algorithms. However, biochemical parameters, i.e., LC_{ab} and CCC, seem to slightly benefit from AC using both Sen2Cor and *i*Cor. Therefore, AC is recommended for future studies when estimating biochemical parameters.

The spatial distribution maps from Sen2Cor showed greater discrimination of LAI values (**Figure 4-9**a), where there was great variability between individual pixels, while the results from *i*Cor and TOA data are generally smooth and similar

(**Figure 4-9**d and g). LC$_{ab}$ and CCC maps were generally similar across all AC techniques.

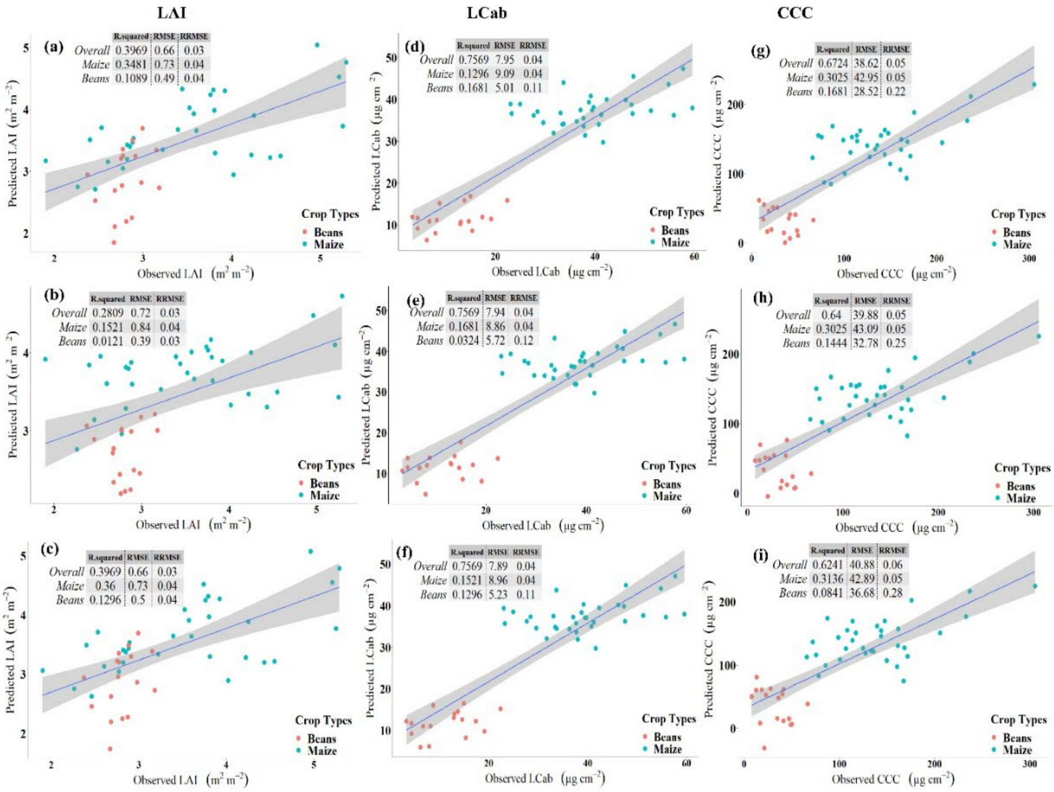


Figure 4-8. Scatter plots showing the retrieval accuracies from sPLS and Sen2Cor (a, d, and g), *i*Cor (b, c, and h), and TOA (c, f, and i).

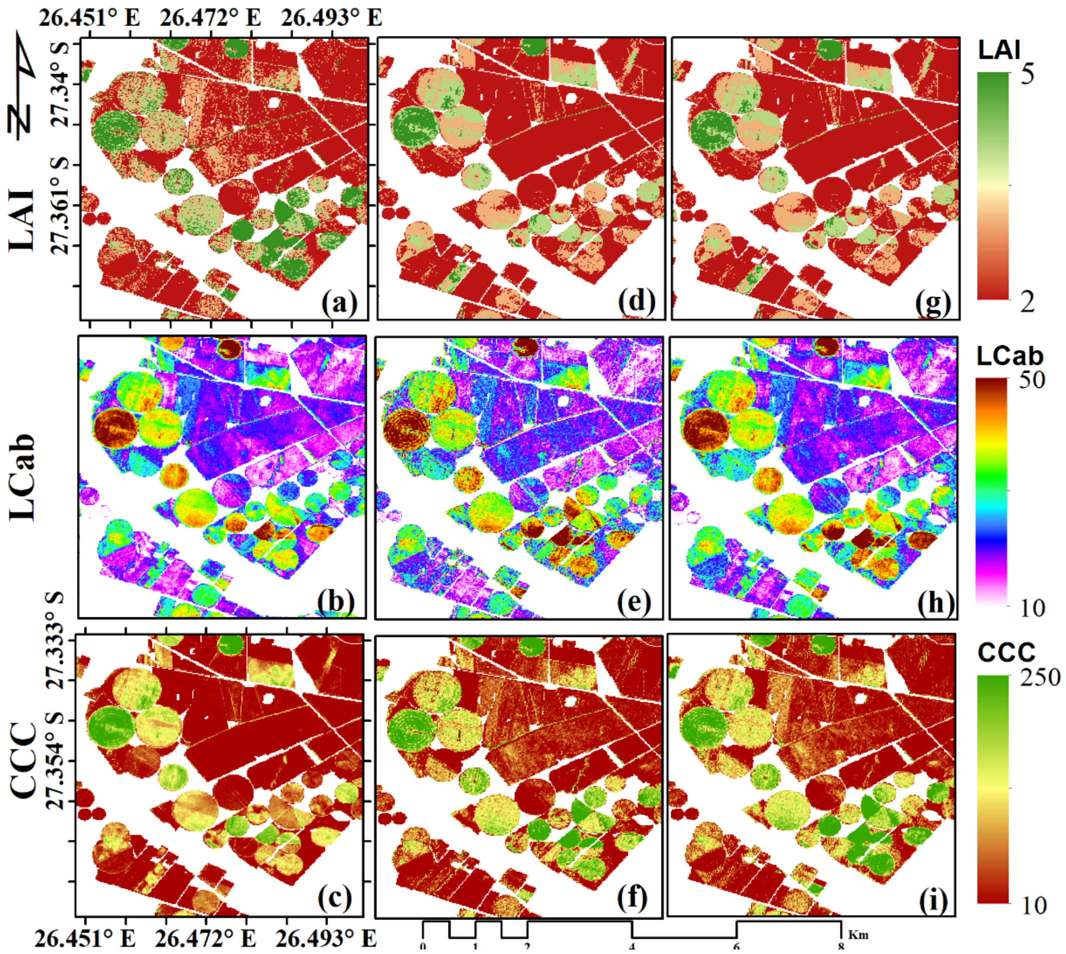


Figure 4-9. Spatial distribution of crop parameters, i.e., LAI, LC_{ab} and CCC, derived from Sentinel-2 MSI data corrected by Sen2Cor (a, b, and c), *i*Cor (d, e, and f) and uncorrected Top-of-Atmosphere data (g, h, and i).

4. Conclusions

This study assessed the SR products from various AC techniques for Sentinel-2 MSI in a semi-arid agricultural landscape. Validation using in situ surface reflectance measurements was carried out on a Sentinel-2 image with high cloud cover, i.e., 61.89%, and inter-comparison using cloud-free Sentinel-2 image. Validation of Sentinel-2 MSI SR image with >60% cloud cover showed relatively moderate consistencies with in situ surface reflectance, and negligible differences in performance between AC techniques. On the one hand, significantly low overall fit with in situ reference data were observed in the VIS and SWIR bands (R^2: ~0.1; $p > 0.05$), while the relatively good R^2 values (i.e., from ~0.6 to ~0.62) were observed in RE (i.e., B6–B7) and NIR (i.e., B8 and B8A) bands across all AC approaches. These observations have profound implications for the accuracy and reliability of crop parameters that use VIS and SWIR input bands, especially in cloudy scenes, >60%. On the other hand, the better fits in RE and NIR bands will benefit agricultural applications that depend on these bands such as assessment of structural parameters, canopy water, and nitrogen content. Further validation in different cloud cover conditions is needed in future studies to understand the magnitude of the effects of clouds on SR derived from various AC techniques.

The current study was performed with limited in situ measurements which were also not coincident with the satellite measurements. Moreover, the atmospheric conditions during the data acquisition were not fully characterised, thus limiting the conclusiveness of the results. In addition, future studies should quantitatively assess the effectiveness of adjacency correction and cloud masking procedures and their effect on derived SR. The inter-comparison of SR products from various AC techniques revealed inconsistencies in correcting for Rayleigh and aerosol scattering in the blue band, i.e., B2. The consistency of AC approaches in this study varied according to crop types and spectral bands. Interestingly, Sen2Cor and *i*Cor showed close to 1:1 similarity to TOA data in most bands, thus suggesting that the AC technique had a low effect and can be eliminated to derive some crop parameters (e.g., LAI) directly from TOA. Nonetheless, results showed that chlorophyll-related crop parameters could benefit from AC. Overall, this study has implications for agricultural applications using Sentinel-2 MSI, particularly the choice of SR bands for retrieval of crop biophysical and biochemical parameters essential for Precision Agriculture.

Both AC techniques had a poor fit with the reference SR in the VIS and SWIR bands, thus likely to negatively impact retrieval techniques that depend on these bands, such as assessment of structural parameters (i.e., LAI), canopy water and nitrogen content. The fit was better in the RE and NIR bands across all AC approaches, thus indicating that these bands can be used to achieve reliable crop parameter estimates, especially in partly cloudy scenes, ~60%.

CHAPTER 5
ESTIMATING CROP BIOPHYSICAL PARAMETERS USING MACHINE LEARNING ALGORITHMS AND SENTINEL-2 IMAGERY

This chapter is based on the following manuscript:

Kganyago, M., Mhangara, P., & Adjorlolo, C. (2021). Estimating Crop Biophysical Parameters Using Machine Learning Algorithms and Sentinel-2 Imagery. *Remote Sensing*, *13*(21), 4314. DOI: https://doi.org/10.3390/rs13214314

Abstract

Global food security is critical to eliminating hunger and malnutrition. In the changing climate, farmers in developing countries must adopt technologies and farming practices such as Precision Agriculture (PA). PA-based approaches enable farmers to cope with frequent and intensified droughts and heatwaves, optimising yields, increasing efficiencies, and reducing operational costs. Biophysical parameters such as Leaf Area Index (LAI), Leaf Chlorophyll Content (LC_{ab}), and Canopy Chlorophyll Content (CCC) are essential for characterising field-level spatial variability and thus are necessary for enabling Variable Rate Application technologies, precision irrigation, and crop monitoring. Moreover, robust machine learning algorithms offer prospects for improving the estimation of biophysical parameters owing to their capability to deal with non-linear data, small samples, and noisy variables. This study compared the predictive performance of sparse Partial Least Squares (sPLS), Random Forest (RF), and Gradient Boosting Machines (GBM) for estimating LAI, LC_{ab}, and CCC with Sentinel-2 imagery in Bothaville, South Africa. Using variable importance measures, the study further identified the most influential bands for estimating crop biophysical parameters. The results showed that RF was superior in estimating all three biophysical parameters, followed by GBM which was better in estimating LAI and CCC, but not LC_{ab}, where sPLS was better. Since all biophysical parameters could be achieved with RF, it can be considered a good contender for operationalisation. Overall, the findings in this study are significant for future biophysical product development using RF to reduce reliance on many algorithms for specific parameters, thus facilitating the rapid extraction of actionable information to support PA and crop monitoring activities.

1. Introduction

Globally, achieving food security is one of the most significant challenges facing humanity (Clark & Tilman, 2017). Addressing such a colossal challenge will require global food production to be increased by 70% to feed an estimated 9 billion people by 2050 (United Nations [UN], 2015). In addition, the capacity of agricultural systems to sustain food security is undermined by climate change and associated disasters such as droughts and heatwaves. These climatic phenomena lead to reduced productivity of agricultural systems and loss of profits and livelihoods. To cope with the climatic variability and change, farmers must adopt resilient technologies and farming practices such as Precision Agriculture (PA) to increase efficiencies and productivity and reduce operational costs. The innovative management practice of PA utilises a set of technologies and principles for site-specific management of spatiotemporal variability related to various aspects of agricultural production (Peralta et al. 2015). As a food-insecure region, Africa needs greater adoption of PA to address food security challenges and ensure the resilience and profitability of agroecological systems.

For example, Variable Rate Application (VRA) of fertilisers, herbicides, and pesticides, i.e., where and when needed, will reduce their loading in the environment and management costs relative to the traditional uniform application (Manandhar et al., 2020). For instance, better wheat grain quality (Karatay & Meyer-Aurich, 2020), reduced pesticide cost by 60%–67% and time, and fuel and labour for application by 27%–32% and 28%, respectively, have been reported (Manandhar et al., 2020). Moreover, Stamatiadis et al. (2020) show the reduced effects of excess nitrogen application such as soil acidity, electrical conductivity, and residual nitrates in the root zone of cotton.

Remote sensing satellites are critical for the rapid assessment of crop health and the development of spatially explicit site-specific information to facilitate VRA technologies and field management. Therefore, they are one of PA's most promising and sustainable technologies based on their well-established capability to acquire data over large areas remotely and repeatedly at various spatial and temporal resolutions. One of the most established approaches to assess crop spatial variability and monitor productivity of agricultural systems is by estimating key crop biophysical parameters such as Leaf Area Index (LAI), Leaf Chlorophyll Content (LC_{ab}), and Canopy Chlorophyll Content (CCC). The LAI is defined as the one-sided area (m^2) of the total developed green leaf area per unit ground surface area (m^2) in broadleaf canopies (Weiss et al., 2004). It is an essential biophysical parameter that provides valuable information on plant physical and physiological processes, and thus is critical for characterising crop growth status and health and stress.

In contrast, chlorophyll is the main biochemical parameter involved in photosynthesis and related to nitrogen content (Haboudane et al., 2008). Therefore, LC_{ab} and CCC are essential parameters for determining photosynthetic capacity, optimising N application to increase yields and profits, and reducing environmental impact from excessive fertilisation (Boegh et al., 2002). Traditionally, these parameters have been measured using direct laboratory-based methods which are destructive, spatially and temporally limited, expensive, time-consuming, and labour-intensive. Therefore, there is demand for remotely sensed approaches for biophysical parameter estimation that will rapidly and frequently characterise crop conditions at regional scales. This can then be to inform the development of policy instruments for alleviating the impacts of climate change. On a landscape scale, they facilitate land management decision-making by providing information about when, where and how much to fertilise or irrigate, thus promoting the sustainability and profitability of agricultural systems.

Remote sensing satellites measure the varying reflectance properties determined by canopy biophysical and biochemical parameters (Jensen et al., 2007). For example, reflectance properties in the visible region (VIS, 350–649 nm), characterised by intense absorption in the blue (350–449 nm) and red (550–649 nm) bands, are caused by carotenoids, xanthophyll, anthocyanin, and chlorophyll pigments (Blackburn 1998). The near-infrared (NIR, 750–1 299 nm) region is characterised by the high reflectance caused by the spongy mesophyll cells, canopy structure, and water content in leaves (Jensen et al. 2007), while the plant canopy reflectance in the SWIR region (1 300–2 500 nm) is affected by subtle biochemical properties such as sugars, cellulose, starch, protein, and water content (Curran, 2001). Studies during the past decade (Ollinger, 2011; Verrelst et al., 2016) suggest that the canopy spectral properties are influenced by the interaction of various biophysical parameters. Accordingly, the VIS-NIR and SWIR regions are commonly used for estimating crop biophysical parameters using parametric statistical approaches such as spectral vegetation indices (*SVIs*), non-parametric techniques such as Machine Learning Regression Algorithms (MLRAs), inversion of coupled leaf-canopy Radiative Transfer Models (RTMs) such as PROSAIL (Jacquemoud et al., 1995; Jacquemoud et al., 2009), and hybrid approaches such as RTMs and MLRAs (Verrelst et al., 2015). Critically, all parametric approaches, i.e., SVIs, regardless of their form—simple ratios, spectral band differences, and normalised difference ratios—are sensitive to the changes in soil background at early growth stages, leaf chlorophyll content, and saturate at lower and higher canopy cover, i.e., LAI < 2 and >5 (Fu et al., 2013; Verrelst et al., 2015).

Commercial sensors such as RapidEye and Worldview-2 have improved capabilities and prospects for estimating biophysical parameters by including narrow red-edge (RE) bands (Huang et al., 2017). The RE spectral region (650–749 nm) is characterised by a steep increase in vegetation reflectance from strong chlorophyll absorption in the red band to high foliage reflection in the NIR, i.e., LAI. This spectral region is sensitive to changes in the leaf chlorophyll content and plant stress and has been shown to improve biophysical parameters estimation accuracy (Mutanga & Skidmore, 2007; Verrelst et al., 2012). Previously, exorbitant costs of commercial image data limited the full potential and exploitation of the RE region (Houborg et al., 2015). The advent of Sentinel-2 Multi-Spectral Imager (MSI), with three new RE bands centred at 705 nm, 740 nm, and 783 nm, is particularly attractive since it is free to access and offers prospects of high spatial (i.e., <20 m) and temporal resolution (i.e., five days), as well as LAI, LC_{ab} and CCC for Precision Agriculture and crop monitoring applications (Mulla, 2013). Using the Sentinel-2 and Sentinel-3 band settings, Clevers and Gitelson (2013) observed that the RE indices, i.e., red-edge chlorophyll index (CIred-edge), the green chlorophyll index (CIgreen), and the MERIS terrestrial chlorophyll index (MTCI), were linearly related to CCC and nitrogen (N) content and hence were accurate estimators. However, Sakamoto (2020) argue that SVIs are not transferable over various canopy architectures, leaf structures, climate zones, and environmental conditions, and contain lower information content.

Alternatively, physically based methods such as PROSAIL (Jacquemoud et al., 2009) can simulate spectral properties of complex vegetation canopies and solar and acquisition angles through RTMs. Physically based methods were shown to be more robust, accurate, and transferable over various cover types and environments (Atzberger, 2004). However, they require substantial site-specific data for parameterisation, which has proven difficult, costly, and slow (Verrelst et al., 2015). Furthermore, physically based approaches are complex, computationally expensive, and ill-posed (i.e., different combinations of canopy parameters may correspond to similar spectra, thus yielding inconsistent results) (Atzberger, 2004; Houborg & Boegh, 2008). In contrast, non-parametric approaches, i.e., MLRAs, require less parameterisation, are relatively fast, and can learn non-linear relationships. Over the years, the MLRAs have evolved into reliable approaches, implementable at various spatial and temporal scales. The MLRAs such as Random Forest (RF) (Breiman, 2001), Support Vector Machines (SVM) (Vapnik, 1999), and Artificial Neural Networks (ANN) were demonstrated to be more robust regarding noisy features, small training size, and high dimensionality (when $p > n$) and collinearity (Verrelst et al., 2012; Liang et al., 2015; Houborg & McCabe, 2018).

Recent studies show that the coupling of physically based approaches and MLRAs, i.e., hybrid approaches, yields better prediction accuracies. These approaches are available operationally, such as the Sentinel-2 Level-2 Prototype Processor (SL2P) through Sentinel Application Platform (SNAP) and Sen2Agri system (Baret et al. 2013; Camacho et al. 2013) for PROBA-V, Sentinel-2 MSI, and Landsat-8 Operational Land Imager (OLI). Although some validation and inter-comparison studies (Brown et al. 2021) show their robustness and better accuracies in various agricultural environments, others (Bochenek et al., 2017; Kganyago et al., 2020) show poor accuracies with RMSE > 1 $m^2\ m^{-2}$. In contrast, simple and more robust MLRAs such as Random Forest, Gaussian Process Regression (GPR) (Rasmussen, 2003), and Kernel Ridge Regression (KRR) (Shawe-Taylor & Cristianini, 2004) have better estimation accuracy and computational cost, can learn complex relationships from input data, and require few and intuitive tuning parameters (Camps-Valls et al., 2009; Verrelst et al., 2012; Verrelst et al., 2015). Owing to continuous improvements in MLRA, optimisation techniques (e.g., gradient descent), sensors capabilities (such as increasing resolution and number), and evolving user needs(Atzberger, 2013), it is essential to evaluate various MLRAs to improve site-specific accuracy and reliability and inform future operationalisation, towards the improvement of food production in developing countries.

This paper assessed various MLRAs to estimate crop biophysical parameters using Sentinel-2 data in the Bothaville site characterised by medium to large commercial farming systems. The specific objectives were (1) to evaluate and compare the predictive performance of sparse Partial Least Squares (sPLS), RF, and Gradient Boosting Machine (GBM) in estimating LAI, LC_{ab}, and CCC using Sentinel-2 MSI data, and (2) to identify the influential spectral bands with high predictive power for estimating the crop biophysical parameters. These algorithms have been limitedly exploited in the context of crop biophysical parameters estimation from Sentinel-2 imagery for Precision Agriculture and crop monitoring. Moreover, the comparison of the three MLRAs in the context of crop biophysical parameter estimation is worthwhile to elucidate their capabilities under the same environmental and acquisition conditions and their consistency to previous performances as they may provide a promising alternative to existing operational techniques.

2. Materials and Methods

2.1. Study area

This study was conducted in Bothaville, i.e., the main agricultural production zone of South Africa (**Figure 5-1**). The area is characterised by both commercial medium- to large-scale farming of crops and livestock. Bothaville is located at latitudes: 27°13′0″S to 28°8′0″S and longitudes: 26°0′0″E to 27°05′0″E. Characterised by warm and wet summers, the average temperature is ~18 °C, with an annual average rainfall of ~584 mm. Winters are cold, with average temperatures of about 5.1 °C. The dominant summer crops are maize, sunflower, and groundnuts, while winter crops include wheat and barley grown on sandy to sandy-loamy soils on generally flat slopes. The summer cropping calendar starts in December and ends in May or June, while the winter crop calendar starts in May and ends in October or November each year (https://ipad.fas.usda.gov/rssiws/al/crop_calendar/safrica.aspx, accessed: 05 June 2021). Bothaville is one of two sites used for testing, calibration, and validation of agricultural monitoring and early warning products within the context of the AfriCultuReS project (Enhancing Food Security in African AgriCultural Systems with the support of Remote Sensing, Grant Agreement No. 774652).

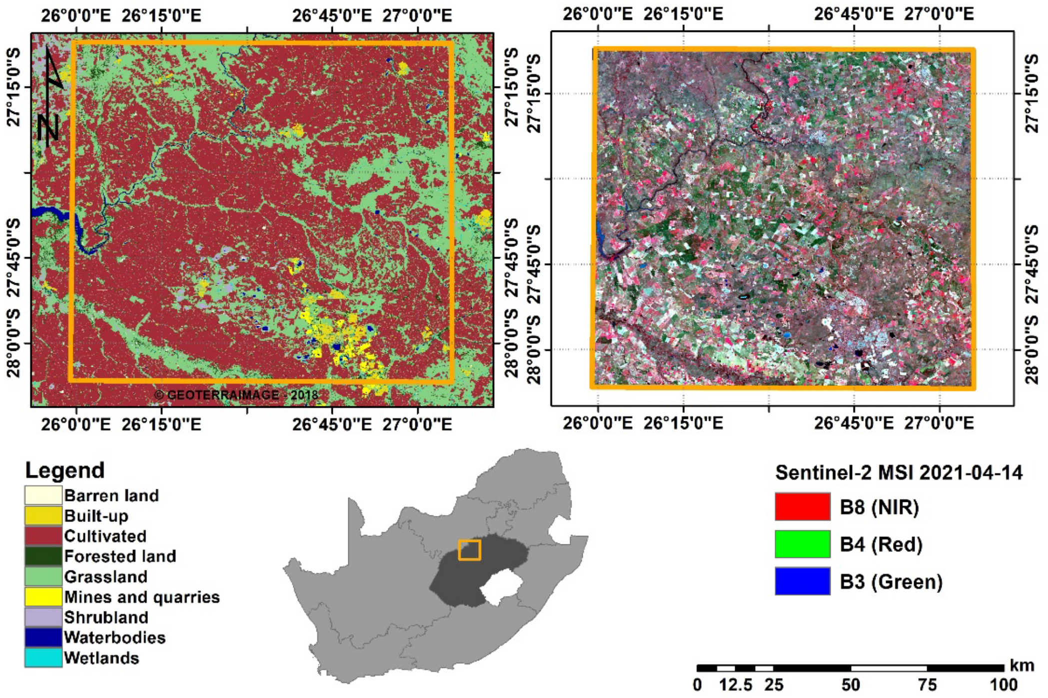


Figure 5-1. Location of Bothaville (orange) in Free State province (dark grey), South Africa. The panel on the left shows the land cover distribution in 2018 and the one on the right shows Sentinel-2 false-colour composite (red: B8, green: B4, and blue: B3) acquired on 14-04-2021.

2.2. Data

2.2.1. Remotely sensed data

Sentinel-2 MSI image acquired on 14 April 2021 covering the study area was downloaded from Sentinel Hub Cloud API for Satellite Imagery (Sinergise Laboratory for geographical information systems, Ltd., Ljubljana, Slovenia). These images are provided at Level-1C processing, i.e., Top-of-Atmosphere (TOA) reflectance in cartographic geometry. The MSI images consisted of 12 bands at three different spatial resolutions, i.e., 10 m (Band 2-Blue:490 nm, Band 3-Green:560 nm, Band 4-Red:665 nm, and Band 8-NIR1:842 nm), 20 m (Band 5-Red-edge:705 nm, Band 6-Red-edge:740 nm, Band 7-Red-edge:783 nm, Band 8A-NIR2:865 nm, Band 11-SWIR1:1 610 nm, and Band 12-SWIR2:2 190 nm), and 60 m (Band 1-Deep blue:443 nm, Band 9-NIR3:945 nm, and Band 10-Cirrus:1 375 nm). For this study, all spectral bands at 10 m, 20 m, and 60 m resolution were used for the study, with 60 m bands dedicated to atmospheric corrections and cloud screening (Drusch et al., 2012). For estimation of crop biophysical parameters, we used 10 m and 20 m spectral bands to take advantage of the entire MSI spectral range. The Level-1C data were converted to Bottom-of-Atmosphere (BOA) reflectance, i.e., Level-2A, using Sen2Cor version 2.9 (Louis et al., 2016).

Sen2Cor is a processor for atmospheric correction (including cirrus clouds and terrain correction) of Sentinel-2 data. The algorithm converts the TOA (Level-1C) image data to BOA reflectance based on the libRadtran database of Look-Up-Tables (LUTs) generated for a wide variety of atmospheric conditions, solar geometries, and ground elevations. Scene classification is also performed as part of the processing to flag clouds, snow, and cloud shadows. Further details can be found in Louis et al. (2016). The image data was corrected using parameters: atmospheric model "mid-latitude summer", aerosol type "rural", and two-band water volume retrieval (i.e., 940 nm and 1 130 nm). The image data were resampled to 20 m spatial resolution using the nearest neighbour resampling technique in Sentinel-2 Toolbox within SNAP software (*SNAP-ESA Sentinel Application Platform v8.0,* http://step.esa.int, accessed: 10 November 2020) based on its capability to preserve the fidelity of the pixel values.

2.2.2. Calibration and validation data

The data for calibrating and validating LAI, LC_{ab}, and CCC MLR models were collected in the field from 11 to 23 April 2021 over the dominant crops in the study area, i.e., maize, beans, and peanuts. LAI and LC_{ab} measurements were collected non-destructively within 40 m × 40 m plots to avoid edge effects and allow biophysical parameter mapping at 20 m resolution. The plots were selected systematically along a transect to capture variability. Each plot consisted of an average of six to eight random measurements. For LAI measurements, we used the

LI-COR 2200c Plant Canopy Analyzer (LI-COR, Inc., Lincoln, NE, USA), with a 180° view cap to shield the influence of the operator and unequal sky conditions on the measurements. Trimble® TDC600 handheld Data Collector, with global navigation satellite systems (GNSS) accuracy of 1.5 m, was used to Geo-tag the centroid of each plot and take plot pictures. In contrast, LC_{ab} measurements were collected with MC-100 Chlorophyll Concentration Meter (Apogee Instruments, Inc., Logan, UT, USA) from the sunlit upper canopy. The CCC for each plot was estimated as a product of LC_{ab} and LAI (LC_{ab} × LAI) (Jacquemoud et al., 1995). The data were divided into 70% training and 30% validation (Adelabu et al., 2015). The summary statistics for calibration and validation are given in **Table 5–1**.

Table 5-1. Descriptive statistics of the calibration (70%) and validation (30%) datasets for the measured LAI ($m^2\ m^{-2}$), LC_{ab} ($\mu g\ cm^{-2}$), and CCC ($\mu g\ cm^{-2}$).

	Datasets	*n*	Min	Mean	Max	SD
LAI	Calibration	113	1.78	3.35	5.57	0.86
	Validation	48	2.02	3.60	5.75	1
LC_{ab}	Calibration	113	4.06	33.87	66.18	14.93
	Validation	48	3.69	32.62	70.69	19.20
CCC	Calibration	113	10.30	105.69	288.22	61.60
	Validation	48	7.87	116.42	339.10	90.39

2.3. Crop and green-vegetation masking

Non-croplands were masked from the Sentinel-2 MSI data using a crop mask derived from the 2017 National Crop Boundaries Dataset (CropEstimatesConsortium, 2017) provided by the Department of Agriculture, Land Reform and Rural Development of South Africa. This dataset is generated from SPOT 5 and SPOT 6 data and consists of all the active crop fields (including small-holder and commercial fields) in a specific period, i.e., 2014–2015. Then, the fallow crop fields and non-vegetated pixels within these fields were removed from further analysis using a vegetation mask. The vegetation mask was created based on a Normalised Difference Vegetation Index (NDVI) with no vegetation threshold of <0.2, determined through trial and error (Equation 5-1). In doing so, further analysis was constrained to the active crop fields in the 2021 growing season since the boundaries did not necessarily represent the current situation in the study area, i.e., during the period of fieldwork.

$$\text{Vegetation mask} = \begin{cases} \text{NDVI} < 0.20 = 0 \\ \text{NDVI} > 0.21 = 1 \end{cases} \quad (5\text{-}1)$$

2.4. Machine Learning Regression Algorithms

The selection of MLRAs considered here was based on their simplicity, quick training, robust performance, and popularity in various remote sensing applications. On the one and, the sparse Partial Least Squares regression (sPLS) algorithm, for example, has been widely used in remote sensing applications, e.g., crop yield prediction (Abdel-Rahman et al. 2014), grass biomass, and LAI estimation (Sibanda et al. 2017; Sibanda et al. 2019), while RF has become one of the most popular algorithms for biophysical parameter estimation using a variety of remotely sensed data (Xu et al., 2017; Houborg & McCabe, 2018; Shah et al., 2019). On the other hand, GBM is computationally effective and often outperforms other algorithms (Beltran et al., 2019). The three are also regarded interpretable, (i.e., essential for obtaining novel insights, such as new causal links between explanatory and response variables) and troubleshooting the models (i.e., detecting and diagnosing biases in the input data and trained models) for improved satellite-based value-added products (Azodi et al., 2020) as opposed to kernel-based and deep learning algorithms such as KRR, SVM, and Neural Networks (NN), which are "black-boxes", complex, and computationally expensive (Moreira et al., 2021). Therefore, the comparison of these three MLRAs in the context of crop biophysical parameter estimation is worthwhile to elucidate their capabilities under the same environmental and acquisition conditions and consistency to previous performances. In addition, they may provide a promising alternative to existing operational techniques.

2.4.1. Sparse Partial Least Squares

The sparse Partial Least Squares regression (sPLS) (Chun & Keles, 2007) is the improvement of ordinary PLS to perform variable selection and prediction simultaneously, thus it is useful for reducing multicollinearity and the so-called "curse of dimensionality" phenomenon (i.e., $p > n$) (Hughes, 1968). Chun and Keleş (2010) achieved this by promoting sparsity using L_1 (the penalty imposed at the dimension reduction step of ordinary PLS), which promotes exact zero property on the surrogate of direction vector c rather than original direction vector α but keep both direction vectors close to each other. Thus, sPLS objective function, by Chun and Keleş (2010), is given by Equation 5-2.

$$\min_{\alpha,w} -\kappa\alpha^T M\alpha + (1-\kappa)(c-\alpha)^T M(c-\alpha) + \lambda_1 |c|_1 + \lambda_2 |c|_2,$$

$$\text{subject to } \alpha^T\alpha = 1. \tag{5-1}$$

where $M = X^T Y Y^T X$. The L_1 penalty encourages sparsity on the surrogate of direction vector c. In contrast, when solving for c, L_2 penalty avoids potential singularity in M. The κ parameter is used to avoid locally optimal solutions and reduce the concavity of the problem. Chun and Keleş (2010) note that the problem becomes the PLS original maximum eigenvalue problem when $\kappa = 1$, that of SCOTLASS (Jolliffe et al., 2003) when original and surrogate direction vectors are equal, i.e., $\alpha = c$, and $M = X^T X$, and that of Sparse Principal Component Analysis (SPCA) (Zou et al., 2006) when $\kappa = \frac{1}{2}$ and $M = X^T X$. Essentially, the algorithm determines relevant (or active) variables by solving the minimisation problem in Equation (2). Then, using ordinary PLS regression and selected variables, it updates all direction vectors to form a Krylov subsequence—critical for algorithm convergence—on the subspace of the relevant variables (Chun & Keleş, 2010). Previously, this algorithm has been used for crop yield prediction (Abdel-Rahman et al., 2014), estimation of grass biomass (Sibanda et al., 2017), and LAI (Sibanda et al., 2019).

The sPLS regression tuning parameters, i.e., λ_1 (eta) and *K*, were tuned using the 10-fold cross-validation (*cv*) strategy for all the direction vectors. The λ_1 (or eta) is a sparsity or thresholding parameter and must be between 0 and 1, while *K* is the number of latent (hidden) components and can be between 1 and $min\left\{p, \frac{(v-1)n}{v}\right\}$, where p is the number of predictors and *n* is the sample size. The optimal parameters minimised the Mean Squared Error of *cv* (MSE*cv*, Equation 5-3). The sPLS analysis was accomplished in R-statistics software (Tierney, 2012) with "spls" package (Chung, Chun, & Keles, 2012). The MSE*cv* was converted to Root Mean Squared Error of CV (RMSE*cv*) to compare with other MLRAs.

$$MSECV\,(\lambda_1, K) = \frac{1}{nq}\sum_{s=1}^{10} ||Y_{[s]} - X_{[s]}\hat{\beta}_{\lambda,K}^{(-s)}||_F^2 \tag{5-2}$$

where $\hat{\beta}^{(-s)}_{\lambda,K}$ is the coefficient estimates with K latent components from the sth training datasets, i.e., the entire data without sth validation dataset, $Y_{[s]}$ and $X_{[s]}$ are the sth validation datasets, and $||A||^2_F = trace\ (A^T A)$ is the square of the Frobenius norm.

2.4.2. Random Forest

The RF regression (Breiman, 2019) is an ensemble tree-based machine learning algorithm and an improvement of Classification and Regression Trees (CART) (Breiman et al., 1984). In contrast to CART, RF uses bagging (bootstrap aggregating) to iteratively and independently build many decision trees (*ntree*) based on a random subset of training samples created by resampling with replacement from the original sample. Bagging is renowned for its robustness against model overfitting, thus it is critical for obtaining a consistent model (Fawagreh et al., 2014). Out of all the training data, about 64% are regarded as in-bag samples, and the remaining 36% are out-of-bag (OOB) samples. These OOB samples are used for evaluating model performance and variable importance (Gislason et al., 2006). At each node in the tree, a small set of randomly selected features (*mtry*) is used to find the optimal split to grow individual trees. Therefore, the trees grown from different and random subsets ensure increased diversity of decision trees and reduce the bias in the regression model (Rodriguez-Galiano et al., 2012). The final prediction is obtained by aggregating all trees (Gislason et al., 2006).

Variable importance for each explanatory variable, calculated as the Percentage Increase in Mean Squared Error (%IncMSE) when OOB samples are permuted, and while all others are constant, is used to rank the variables according to their predictive power in the model. Essentially, the higher the importance of a specific explanatory variable, the stronger the relationship with the response variable (Mutanga et al., 2012). The optimal *mtry* and *ntree* parameters for estimating LAI, LC$_{ab}$, and CCC with Sentinel-2 MSI were determined by the grid-search strategy (Okun & Priisalu, 2007) using values ranging from 1 to p with a single interval, where p is the total number of input explanatory variables, and from 100 to 500 with an interval of 10, respectively. Grid-search strategy is a commonly used approach for hyperparameter tuning as it considers, exhaustively, all possible parameter combinations and chooses the pair of parameters that yields minimum OOB error. The RF analysis was performed in R-statistics software (Tierney, 2012) using “randomForest” library (Liaw & Wiener, 2002).

2.4.3. Gradient Boosting Machines

GBM, also known as Gradient Boosted Regression Trees (GBRT) (Friedman, 2001) uses gradient boosted decision trees and a more regularised formalisation to avoid overfitting, handles missing values (or sparse data) more efficiently, employs parallel and distributed computing for rapid tree construction and building of large models, respectively, and can fit new data added to the trained model. Like RF, it uses CART as a base regressor and the trees are fitted sequentially to minimise the loss function, with new trees added to the model iteratively.

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i) \tag{5-3}$$

where $\hat{y}$ is the prediction output of sample i, x_i is the corresponding input of sample i, and f_k is the predictive function of the decision tree *K*. The objective function of the ensemble model is shown in Equation 5-5.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5-5}$$

where $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ is the loss function which measures the difference between the predicted $\hat{y}$ and the target y_i; $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega^2\|$ is the regularisation term which represents the complexity of the model; γ, *T*, λ, and ω represent the complexity, the number of leaf nodes in the tree, the fixed coefficient, and the quantised weight vector of leaf nodes, respectively. *K* is the number of trees to be generated and n is the number of instances in the training set. A gradient addition strategy is presented by Friedman (2001), where new trees are added to the trained model one at a time, and the final prediction is achieved through iterative computation.

$$
\begin{aligned}
\hat{y}_i^{(0)} &= 0, \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i), \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i), \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k\,(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}
\tag{5-4}
$$

where $\hat{y}_i^{(t)}$ and $\hat{y}_i^{(t-1)}$ are the predicted values in the t^{th} and $t-1^{th}$ rounds, respectively; and $f_t(x_i)$ is the new predictive function of the added tree in the t^{th} round. This predictive function and the structure of the new tree are determined through an optimisation algorithm, i.e., greedy algorithm, to minimise the objective function and perform prediction. In this study, GBM was performed using "gbm" R-package, which required optimisation of four parameters, i.e., the number of trees (n.trees), learning rate (shrinkage), tree depth (interaction.depth), and subsample (bag.fraction) (**Table 5-2**). These were tuned using the grid-search strategy, and the best combination of parameters was selected based on the model fit with the lowest root-mean-square error of cross-validation (RMSE*cv*). The GBM algorithm also calculates the relative influence of each variable on the model, by averaging the relative influence of variables across all trees.

Table 5-2. Gradient Boosting Machine (GBM) parameters required for optimisation in "gbm" R-package and their descriptions.

Parameters	Description
Number of trees (*T*)	This is the total number of trees to fit or iterations.
Tree depth (*K*)	The depth of a tree determines the number of splits in each tree to control the complexity of the boosted ensemble.
Learning rate (ϡ)	The learning rate controls the speed of the algorithm down the gradient descent. The smaller values improve the performance and reduce the chance of overfitting.
Subsample (*p*)	The subsample ratio of the training instance controls the randomly collected data instance to grow trees. For example, a value of 0.5 causes GBM to randomly collect half of the data instances and prevent overfitting through implementing stochastic gradient descent. The values for this parameter should be between 0 and 1.

2.5. Prediction accuracy assessment

The prediction accuracies of each Machine Learning Regression (MLR) model were assessed with the Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Relative Root Mean Squared Error (RRMSE) (Equations 5-7 to 5-

11). These measures were recommended by Richter et al. (2012) and are frequently used in literature; thus, they facilitate comparison between studies of biophysical parameters. The R^2 is a correlation-based, dimensionless measure that reflects spatial (temporal) patterns, with values ≥ 0.9 interpreted as Excellent and $0.5 \leq R^2 \leq 0.8$ as Good. In contrast, the RMSE indicates the magnitude of error in the units of the biophysical parameter, i.e., $m^2\ m^{-2}$ and $\mu g\ cm^{-2}$ for LAI, and LC_{ab} and CCC, respectively. For LAI, the RMSE values < 0.5 $m^2\ m^{-2}$ can be interpreted as Excellent and 0.5 $m^2\ m^{-2}$ ≤ RMSE < 1.0 $m^2\ m^{-2}$ as Good. Lastly, RRMSE is a dimensionless index suitable for comparisons between different variables or ranges, where the values ≤ 10% are regarded as Excellent and 10% < RRMSE ≤ 20% as Good (Richter et al., 2012). Finally, percentage Bias (%Bias) is a measure of the tendency of a model to underestimate or overestimate a biophysical parameter, where the ideal value is 0% and values close to 0% indicate an accurate model (Gara et al. 2019).

$$R^2 = \frac{\sum(y_i^N - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{5-5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i \times y_i)^2} \tag{5-8}$$

$$RRMSE = \frac{RMSE}{\bar{x}_i} \tag{5-9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{5-10}$$

$$\%BIAS = \sum_{i=1}^{n}(x_i - y_i) / \sum_{i=1}^{n}(x_i) \tag{5-11}$$

where x_i is the observed biophysical parameter (e.g., LAI), and y_i is the predicted biophysical parameter (e.g., LAI), $\bar{x}_i$, and $\bar{y}_i$ are the mean of observed and predicted biophysical parameters, respectively; n is the sample size, and N is the number of errors.

All analysis, including training and validation of the models, as well as the mapping of biophysical parameters, was performed in R-statistics software (Tierney, 2012), while the biophysical parameter maps were generated in ArcMap 10.1 (Environmental Systems Research Institute, Redlands, CA, USA).

3. Results

3.1. Optimal Tuning Parameters

Table 5-3 shows the optimal tuning parameters used for the sPLS, RF, and GBM models selected through a grid-search strategy and 10-fold cross-validation. The results are presented for each MLRA and biophysical parameter. sPLS used all variables, p = 10, for LAI and CCC, relatively fewer variables, i.e., p = 7, are selected for LC_{ab}. The excluded variables for sPLS-LC_{ab} are B6, B8, and B8A.

Table 5-3. Optimal tuning parameters used to train Leaf Area Index (LAI, $m^2\ m^{-2}$), Leaf Chlorophyll Content (LC_{ab}, µg cm^{-2}), and Canopy Chlorophyll Content (CCC, µg cm^{-2}) models based on grid-search parameterisation strategy and k-fold cross-validation.

	sPLS	RF	GBM
LAI	eta = 0.7; K = 5; p = 10; $RMSE_{CV}$ = 0.8	m_{try} = 5; OOB error = 0.34	n_{trees} = 146; interaction depth = 5; shrinkage = 0.1; n.minobsinnode =15; $RMSE_{CV}$ =0.65
LC_{ab}	eta = 0.9; K = 5; p = 7 *; $RMSE_{CV}$ = 7.58	m_{try} = 3; OOB error = 7.21	n_{trees} = 28; interaction depth = 3; shrinkage = 0.1; n.minobsinnode =15; $RMSE_{CV}$ = 7.56
CCC	eta = 0.8; K = 5; p = 10; $RMSE_{CV}$ = 41.22	m_{try} = 2; OOB error = 34.56	n_{trees} = 94; interaction depth = 5; shrinkage = 0.1; n.minobsinnode = 15; ; $RMSE_{CV}$ = 37.21

Variables selected by sPLS: * B2, B3, B4, B5, B7, B11, B12.

3.2. Model performance

Sentinel-2 MSI data were used for estimating LAI, LC_{ab}, and CCC using sPLS, RF, and GBM. The results for LAI (**Figure 5-2**a–c) show that RF yields better predictive performance, i.e., RMSE 0.5 $m^2\ m^{-2}$, explaining 76% of LAI variability, followed by GBM with RMSE of 0.63 $m^2\ m^{-2}$ and R^2 of 0.61, while sPLS offers relatively poor performance, i.e., RMSE: 0.77 $m^2\ m^{-2}$, explaining only 49% of the variability in LAI. Consistently, sPLS has the highest %Bias, i.e., 5%, while GBM has the lowest %Bias, i.e., 1%, among the compared MLRAs. Similarly, in predicting LC_{ab} (**Figure 5-2**d–f), the results show that the RF algorithm is superior, i.e., RMSE: 7.57 µg cm^{-2}, followed by sPLS with RMSE of 7.90 µg cm^{-2}, while GBM is relatively poor with 8.25 µg cm^{-2}. The variability explained by the sPLS is the lowest (i.e., 81%) when compared to that achieved by RF and GBM models, i.e., 83%, respectively. Moreover, the %Bias is highest in the sPLS model, i.e., 5%, and lowest in the GBM model, i.e., 1%, while RF has a %Bias of 3%. For CCC estimation (**Figure 5-2**g–i), RF results show a relatively good predictive performance with RMSE of 39.49 µg cm^{-2}, when compared to sPLS and GBM. The GBM model has the next best predictive performance, achieving an RMSE of 44.19 µg cm^{-2}, while sPLS is relatively poor with an RMSE of 52.76 µg cm^{-2}. Both the GBM-CCC and sPLS-CCC models have markedly high %Bias, i.e., 12% and 13%, as compared to only 2% achieved by the RF-CCC model. As with LAI and LC$_{ab}$ models, the RF-CCC model explains the greatest variability, i.e., 83%, followed by GBM explaining 77% of CCC variability, while sPLS explains the least variability, i.e., 74%, among the compared MLRAs.

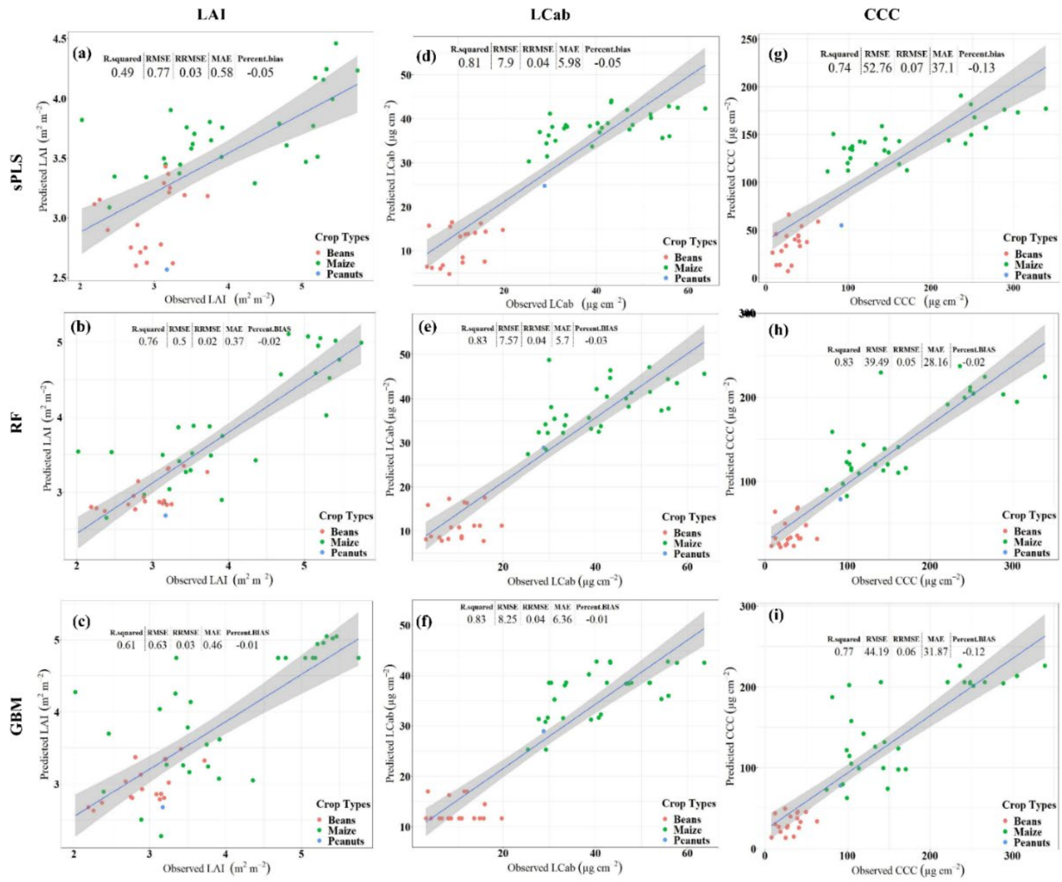


Figure 5-2. Scatterplots for Leaf Area Index (LAI, m^2 m^{-2}, a–c), Leaf Chlorophyll Content (LC$_{ab}$, µg cm^{-2}, d–f), and Canopy Chlorophyll Content (CCC, µg cm^{-2}, g–i) showing the performance of sparse Partial Least Squares (sPLS, a,d,g), Random Forest (RF, b,e,h), and Gradient Boosting Machine (GBM, c,f,i) with Sentinel-2 data.

3.3. Biophysical parameter mapping

The three MLRAs, i.e., sPLS, RF, and GBM, were applied to Sentinel-2 MSI image over the study area to characterise the spatial variation of the crop biophysical parameters, i.e., LAI, LC_{ab}, and CCC, within and between crop fields. The results are presented in **Figure 5-3**. As shown in **Figure 5-3**a,d and g, the spatial variation of LAI within the field is discernible with some differences between MLRAs. The sPLS results (**Figure 5-3**a) show a wide distribution of low LAI values (i.e., ~2 m^2 m^{-2}), particularly over rainfed (regular) fields, while the lowest LAI values estimated by RF (**Figure 5-3**d) and GBM (**Figure 5-3**g) are mostly above 2 m^2 m^{-2}. However, GBM show relatively lesser values within some fields, indicative of senescing leaves. For LC_{ab} (**Figure 5-3**b,e,h), a similar spatial variation can be observed, where sPLS (**Figure 5-3**b) results in widely distributed low LC_{ab} values over rainfed fields and maximum values, i.e., ~50 µg cm^{-2}, over irrigated fields. In contrast, the LC_{ab} values around 50 µg cm^{-2} are not widely distributed in the LC_{ab} maps estimated from RF (**Figure 5-3**e) and GBM (**Figure 5-3**h). At the canopy level, the chlorophyll content (i.e., CCC) maps for RF (**Figure 5-3**f) and GBM (**Figure 5-3**i) are more similar, while sPLS (**Figure 5-3**c) contained lower values (i.e., red areas, ~10 µg cm^{-2}). The results observed here are consistent with previous studies that found that PLS is well adapted to estimating lower values than other algorithms (Rivera-Caicedo, Verrelst, Muñoz-Marí, Camps-Valls, & Moreno, 2017).

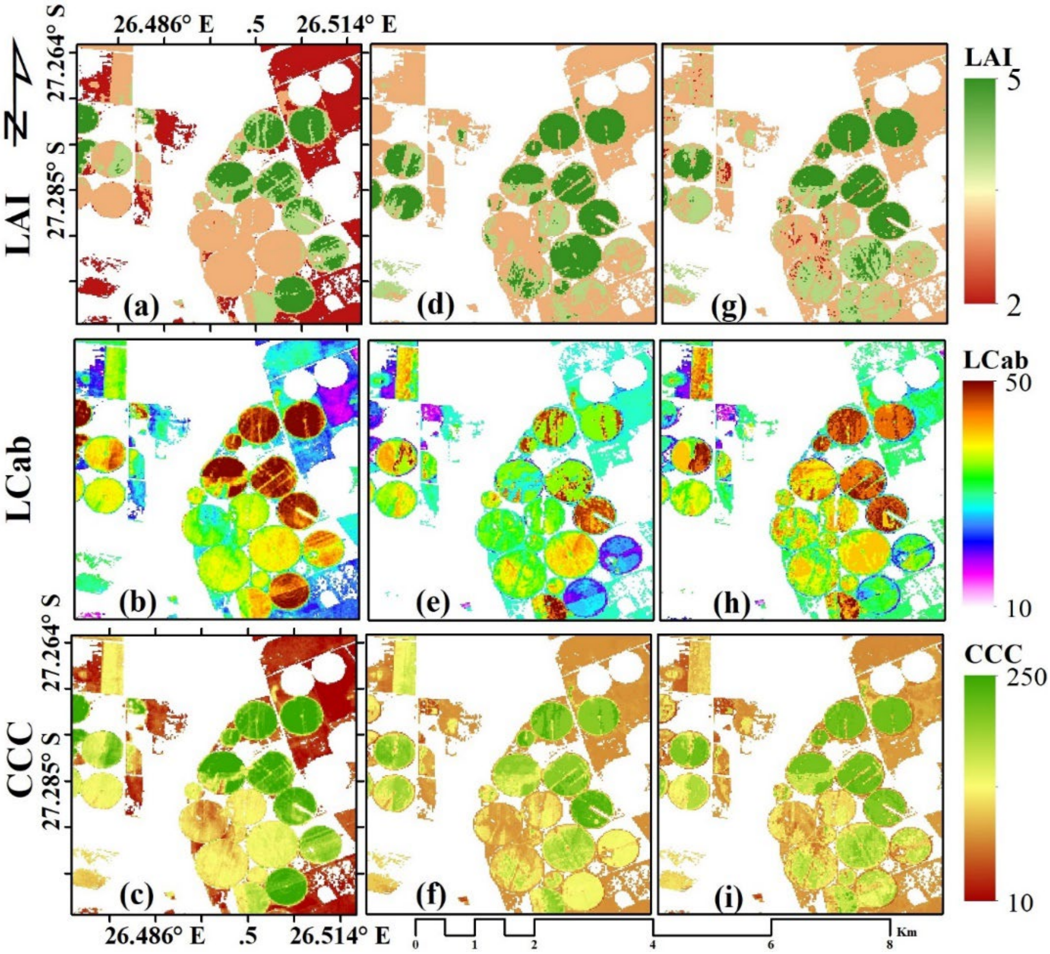


Figure 5-3. Spatial variation of LAI (m^2 m^{-2}, a,d,g), Leaf Chlorophyll Content (LC_{ab}, µg cm^{-2}, b, e, h), and Canopy Chlorophyll Content (CCC, µg cm^{-2}, c,f,i), estimated with sparse Partial Least Squares (sPLS, a–c); Random Forest (RF, d–f); Gradient Boosting Machine (GBM, g–i) using Sentinel-2 data, respectively.

3.4. Variable Importance

RF and GBM algorithms have built-in variable importance measures, measured by %IncMSE and Relative influence (%), respectively. For ease of interpreting the results in this study, the spectral bands with importance values, <10%, >10% <20%, and >20%, are regarded as having a low, moderate, and high, predictive power, respectively. Generally, the results (**Figure 5-4**) indicate the varying influence of Sentinel-2 spectral bands on the model accuracy for each MLRA and processing level. For the RF-LAI model (**Figure 5-4**a), RE bands, i.e., B5:705 nm, have the highest contribution to the model performance, i.e., >20%, followed by SWIR bands, i.e., B11:1 610 nm and B12:2 190 nm, and VIS bands, i.e., B3:560 nm, and B4:665 nm, and RE band, B6:740 nm, which have moderate importance, i.e., >10%

<20%, while NIR bands, i.e., B8A:865 nm and B8:842 nm, and RE band, i.e., B7:783 nm, have relatively low importance (i.e., <10%). In the GBM model, the VIS band, B4:665 nm, has the highest influence on the model performance, while B11:1 610 nm, B5:705 nm, and B3:560 nm, have a moderate influence on the model performance, i.e., >10% <20%. Other bands, i.e., B12:2 190 nm, B8A:865 nm, B2:490 nm, B6:740 nm, B7:783 nm, and B8:842 nm, have the lowest contribution to the model performance, i.e., <10%.

In estimating LC_{ab}, the RF importance results (**Figure 5-4**b) show that all Sentinel-2 spectral bands have a high contribution to the model performance, i.e., >20%, while the GBM model shows that only one variable out of ten variables (i.e., spectral bands) has a similar high contribution to the model performance, as depicted by **Figure 5-4**e. Specifically, the SWIR bands (B11:1 610 nm and B12:2 190 nm), RE band (B5:705 nm), and VIS bands (B3:560 nm and B4:665 nm), are the most influential to the RF-LC_{ab} model performance, with %IncMSE of >30%. The NIR bands (B8:842 nm and B8A:865 nm) have relatively low importance alongside VIS band (B2:490 nm), and RE bands (B6:740 nm and B7:783 nm). In contrast, the GBM results show the greatest influence of the VIS band (B3:560 nm), i.e., >20%, while the VIS bands (B2:490 nm and B4:665 nm), and RE band (B6:740 nm) have a moderate influence (i.e., >10% <20%) on the GBM-LC_{ab} model. As with RF-LC_{ab} models, the variable importance for the RF-CCC model (**Figure 5-4**c) is also high, i.e., %IncMSE >20%, for all Sentinel-2 spectral bands, except for B12:2 190 nm, which has a moderate influence on the model, i.e., >10% <20%. The most important variables, i.e., the VIS bands (B3:560 nm, B2:490 nm, and B4:665 nm), RE bands (B5:705 nm, B6:740 nm, and B7:783 nm), SWIR band (B11:1 610 nm), NIR band (B8:842 nm), and narrow NIR band (B8A:865), have a %IncMSE of >30%. In contrast, the GBM-CCC model (**Figure 5-4**f) shows that the SWIR band (B11:1 610 nm), has the highest influence on the model performance, i.e., >20%, while VIS bands (B4:665 nm, B3:560 nm, and B2:490 nm), have a moderate influence on the model performance, i.e., >10% <20%. The rest of the other bands (i.e., B8A:865 nm, B5:705 nm, B7:783 nm, B6:740 nm, B8:842 nm, and B12:2 190 nm) have the least influence on the model, i.e., <10%.

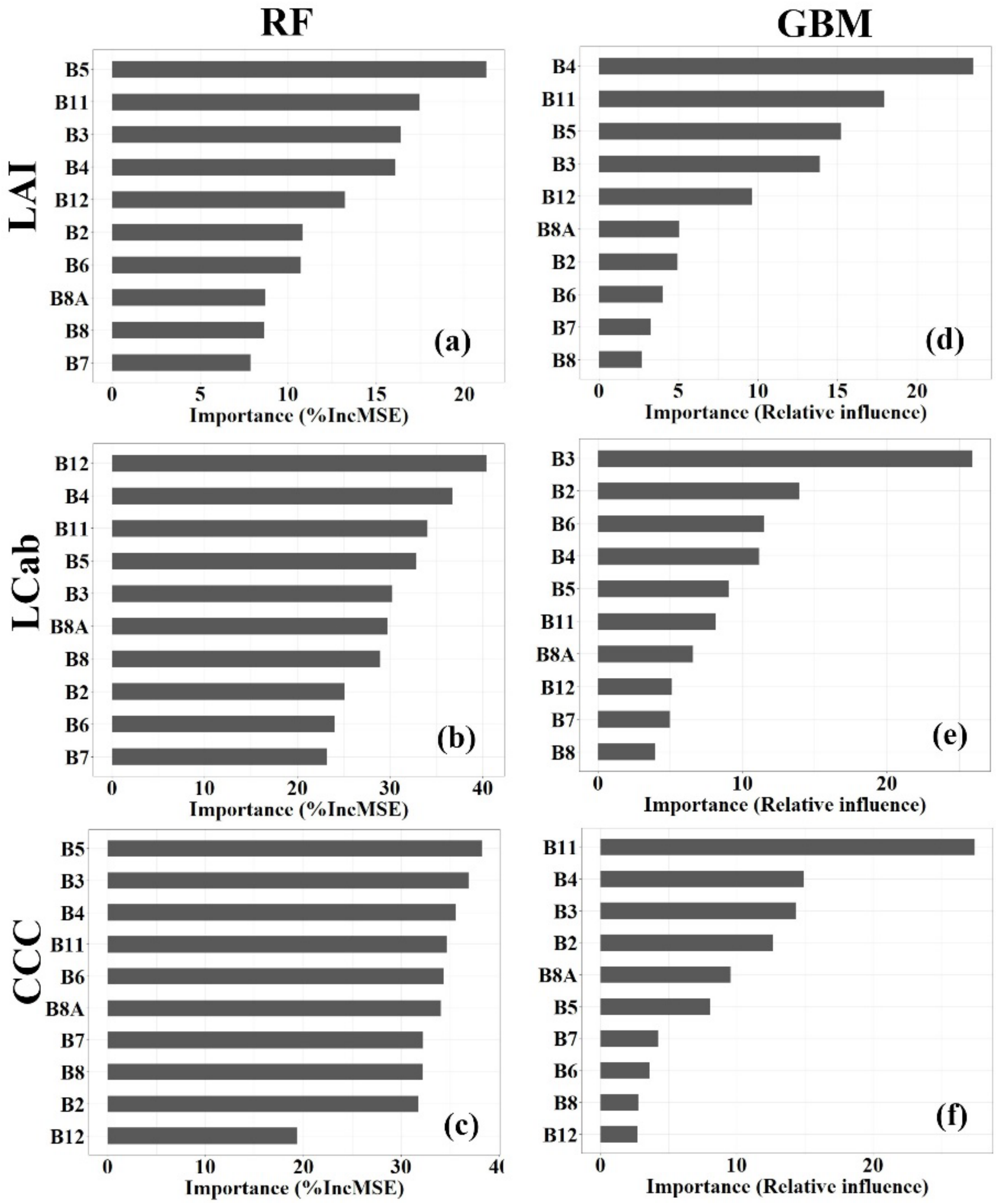


Figure 5-4. Important Sentinel-2 spectral bands (ranked from high to low) for estimating Leaf Area Index (LAI, a,d), Leaf Chlorophyll Content (LC_{ab}, b,e), and Canopy Chlorophyll Content (CCC, c,f), from Random Forest (RF, a–c) and Gradient Boosting Machine (GBM, d–f).

4. Discussion

4.1. Predictive performance of Machine Learning Regression Algorithms

Accurate estimation of crop biophysical parameters from satellite imagery is an ongoing endeavour in various climatic regions across the world, enabled by advances in estimation approaches and satellite data, and in response to changing user needs. To optimise farm inputs such as fertilisers, irrigation, and labour to improve yields and profits, spatially explicit, site-specific, and accurate biophysical parameters are required at 5-day intervals. Moreover, international agencies such as the Food and Agriculture Organisation (FAO) of the United Nations regional agencies such as the African Union Development Agency (AUDA-NEPAD), and national governments aim to reduce hunger by ensuring adequate and affordable food supply en route to achieving Sustainable Development Goals (SDGs) by 2030. Therefore, there is demand for continuous information on the productivity of agricultural systems that can inform food-related policy- and decision-making. The high-resolution crop biophysical parameters from Sentinel-2 MSI data such as LAI and chlorophyll content at the leaf (i.e., LC_{ab}) and canopy levels (i.e., CCC) are essential for both Precision Agriculture and crop monitoring needs. This paper evaluated the performance of three MLRAs in estimating these foliar biophysical parameters in an African semi-arid agricultural area in Bothaville, South Africa.

MLRAs are among the most promising biophysical retrieval approaches and have the greatest potential for improving the accuracy and reliability of biophysical parameters from satellite data. One of the attractive features of MLRAs is their capability to determine non-linear relationships between biophysical parameters of interest and satellite-based covariates such as spectral reflectance. The results in this study (**Figure 5-2**) demonstrate that the RF regression algorithm yields superior predictive performance in estimating LAI, LC_{ab}, and CCC across compared MLRAs. These results present the RF as a good contender for the operationalisation of biophysical parameters to support Precision Agriculture and crop monitoring since all relevant parameters for detecting crop health could be achieved in a single algorithm. Previous MLRA comparison studies found that different algorithms were optimal for characterising different and individual parameters (Verrelst et al., 2015; Upreti et al., 2019). Therefore, the results here are significant for product developers aiming to reduce reliance on multiple algorithms for multiple parameter estimation. The inner workings of the algorithm are one of its greatest strengths as these are relatively transparent and require only a few intuitive parameters, i.e., m_{try} and n_{tree}, when compared to complex algorithms such as NN. For example, since it is based on the traditional CART algorithm, one can interrogate the variables used to split the nodes in the individual trees and thus explain the predictions. The strength of ensemble learning methods is also shown by the better predictive performance of GBM in estimating LAI and CCC. Although RF and GBM use CART as a base regressor, GBM makes predictions by sequentially and iteratively combining weak regression trees to improve the predictive performance, rather than independently constructing several decision trees (Konstantinov & Utkin, 2021). However, as shown by the results of this study, the RF formulation is superior in terms of predictive performance.

The weakness of the GBM formulation is also shown by its relatively poor performance when estimating LC_{ab}, compared to sPLS which showed better prediction accuracy. The relatively poor performance of sPLS in estimating other biophysical parameters, i.e., LAI and CCC, can be attributed to its inability to deal with non-linear relationships between explanatory variables, i.e., spectral bands, or between these variables and the biophysical parameters (Rivera-Caicedo et al., 2017). Since Sentinel-2 contains optimised and few spectral bands for land monitoring, the variable selection capability of sPLS was not fully exploited. Thus, the observed results can be attributed to ordinary PLS used in the algorithm for projecting the variables into orthogonal space and estimation of biophysical parameters, with the exception of the sPLS-LC_{ab} model. Concerning the sPLS-LC_{ab} model, the capability of reducing dimensionality was observed, which assisted its better prediction accuracy against GBM. Despite its better performance in terms of RMSE and R^2, it showed greater underestimations, i.e., %Bias ~5%, when compared to GBM. Delloye et al. (2018) found that extreme LC_{ab} values, i.e., <30 and >70 µg cm^{-2}, tend to be underestimated. Our results show this underestimation at >50 µg cm^{-2}, possibly owing to the difference in the range of training and validation data and the limitations of MLRAs in estimating beyond the range of training data. In fact, GBM resulted in the lowest %Bias for LAI and LC_{ab} compared to both RF and sPLS. Therefore, its potential should be explored further in future studies, particularly its improved version, i.e., eXtreme Gradient Boosting (XGB), which has been reported to be highly efficient and more accurate.

The predictive performance of RF in estimating LAI found in this study (RMSE: 0.5 m^2 m^{-2}, R^2: 0.76) is slightly lower than that found by Verrelst et al.

(2015), i.e., $RMSE_{cv}$:0.44 $m^2 m^{-2}$ (R^2: 0.90), using variational heteroscedastic Gaussian Process Regression (VHGPR) and simulated Sentinel-2 bands in Barrax, Spain, though their data had a higher spatial resolution, i.e., 5 m. Over the same site, Verrelst et al. (2012) found an RMSE of 0.49 $m^2 m^{-2}$ (R^2: 0.92) using GPR and simulated Sentinel-2 data at 20 m resolution. This finding is comparable to the RF-LAI results in this study; however, their RRMSE is considerably poorer, i.e., 23%, thus failing to reach the prescribed minimum accuracy of 10% for LAI product (ESA, 2019). Our LC_{ab} prediction performance was slightly worse than that reported by Delloye et al. (2018) using NN and Sentinel-2 data, i.e., RMSE: 7.26 $\mu g\ m^{-2}$ and better than that reported by Upreti et al. (2019), i.e., RMSE: 8.88 $\mu g\ cm^{-2}$ using Sentinel-2 data and RF. Considering that all the MLRAs evaluated in this study resulted in considerably better RRMSE, i.e., 2%–3%, the study supports the recommendation by Verrelst et al. (2015) that slower and non-optimal MLRAs such as NN should be replaced by simpler but more powerful ones.

4.2. Influential variables for biophysical parameter estimation

The measures of variable importance have been used extensively in the literature to explain the performance of the modelled relationships with biophysical parameters, as well as variable selection (Mutanga et al., 2012; Ramoelo et al., 2015a; Shah et al., 2019). Moreover, they probe the structure and parameters learned by a trained MLR model to understand important variables or combinations of important variables for deriving predictions. Fortunately, the MLRAs used here, i.e., RF and GBM, had built-in variable importance measures, while the selected variables by sPLS can be considered to contain high information content. The variable importance results (**Figure 5-4**) showed varying importance for each biophysical parameter and MLRA. This is not surprising since the different biophysical parameters affect various regions of the electromagnetic spectrum differently. However, their influence is not exclusive to particular regions. For example, chlorophyll is known to have strong absorption in the 400–500 nm region (blue) and 650–700 nm region (red-edge), while carotenoids also absorb the blue radiation. Moreover, the water content in leaves causes strong absorption features in the SWIR region (1 450 nm and 1 950 nm) and relatively weak absorption in the NIR region (980 nm and 1 150 nm) (Ollinger, 2011). Both regions have been associated with an accurate estimation of biophysical parameters such as LAI since most plant functions depend on water (Verrelst et al., 2015). In agreement, our results show that the high influence of SWIR, VIS, and RE bands is evident over NIR bands, which are less important for LAI using RF and all biophysical parameters using GBM.

Among the three Sentinel-2 MSI RE bands, B5 centred at 705 nm is more important in estimating all the biophysical parameters considered here, irrespective of the algorithm used. Wu et al. (2008) showed that spectral vegetation indices using this band (i.e., 705 nm) are better estimators of chlorophyll content than other RE bands. This finding is consistent with Verrelst et al. (2011) who found that wavelengths above 730 nm result in poor LC_{ab} results owing to low absorption of solar radiation by chlorophyll pigments. Thus, LC_{ab} induces the largest variation in reflectance in the RE region below 730 nm (Delloye et al., 2018). The other Sentinel-2 RE bands are centred above 730 nm, i.e., B6:740 nm and B7:783 nm. Interestingly, in the current study, both LAI and CCC models were influenced by similar spectral bands. For example, in addition to B5, B11:1 610 nm, B3:560 nm, and B4:665 nm, showed the greatest influence on the MLR model performance. Therefore, it is clear that these bands have high information content, explaining the greatest variability of LAI, LC_{ab}, and CCC in the study area. The similarity in variable importance between LAI and CCC reflects the co-variation of LAI and LC_{ab} at the canopy level. The VIS bands (B3:560 nm and B4:665 nm), and SWIR band (B11:1 610 nm), featured prominently in LC_{ab} models, while the contribution of NIR was relatively low. Consistently, the sPLS-LC_{ab} model excluded NIR bands (B8 and B8A). In few instances, i.e., LC_{ab} and CCC with RF, NIR bands (B8:842 nm and B8A:865 nm) were among the variables highly influential (i.e., %IncMSE >20%) to the model performance.

Nonetheless, it should be noted that all the Sentinel-2 MSI covariates contributed towards the MLRA performance; hence they should not be discarded. This is especially true for LAI, which has been found to drive spectral variation in all bands owing to co-variation of biophysical parameters in various spectral regions (Ollinger, 2011; Verrelst et al., 2016). However, it may be interesting to test the different subsets consisting of four to ten bands based on these variable importance measures. This suggestion is informed by the sPLS-LC$_{ab}$ model which used fewer (i.e., seven) variables, resulting in better performance than GBM. Verrelst et al. (2016) consistently found that fewer HyMap bands (i.e., 9 and 7) were optimal (NRMSE$_{cv}$: <10%) for estimating LC$_{ab}$ and LAI, respectively. In another study, Verrelst et al. (2011) found that only four Compact High-Resolution Imaging Spectrometer (CHRIS) bands, i.e., 674 nm, 605 nm, 942 nm, and 978 nm as well as 725 nm, 471 nm, 997 nm, and 511 nm, are sufficient to obtain accurate LAI and LC$_{ab}$ estimations, respectively. Therefore, the RF and GBM importance measures can be explored in future studies to optimise the number of spectral bands for estimating biophysical parameters with MLRAs based on some threshold.

As shown by the results, RF seems to utilise the information content of all Sentinel-2 bands more efficiently than the GBM, where all variables had relatively high or close importance values, especially for LC$_{ab}$ and CCC (**Figure 5-4**b,c). In contrast, GBM's most important variable showed a markedly high relative influence compared to the next or other important variables. This observation can be attributed to the differences in the architecture of these MLRAs (i.e., RF and GBM). RF selects a subset of random variables to split the nodes of each tree; thus, all Sentinel-2 bands have an equal chance of contributing to the model. This may work against or in favour of prediction performance. For example, in datasets containing highly correlated variables, such as hyperspectral data, collinear variables may be selected for all or most of the trees in the forest, resulting in overfitting. In such a case, feature selection and dimensionality reduction become essential (Verrelst et al., 2016; Rivera-Caicedo et al., 2017). Fortunately, in our case the spectral bands were discrete and any redundant bands, (such as B12:2 190 nm, B6:740 nm, B7:783 nm, B8:842 nm, and B8A:865 nm), ranked relatively low in variable importance. The influential variables found here are consistent with the known absorption features and relationships with plant biophysical parameters (Blackburn, 1998; Curran, 2001; Ollinger, 2011).

4.3. Limitations of the study

Despite the superior performance of RF in this study, the model was not tested in a different study area to confirm its prediction consistency and transferability. Owing to the high cost of collecting calibration data, transferability is one of the key issues in biophysical parameter estimation using MLRAs. The coupling of RTM-MLRA seems to be a promising solution for operationalisation, despite the ill-posed nature of physically based approaches. Unfortunately, such efforts have focused mainly on ANN (Fernandes et al., 2014; Delloye et al., 2018), which has been proven inconsistent in semi-arid African environments (Kganyago et al., 2020a). Moreover, the variable importance used to explain the model performance here is inadequate as it provides global interpretations (i.e., based on the entire model architecture and dataset) and explains overall relationships between the explanatory and response variables. However, it is desirable to understand the local explanations of the model, i.e., why a specific prediction was made for a particular observation. In the future, we will explore the transferability and local explanations of MLRAs.

Finally, the MLRAs predictions were limited to the dynamic range of input field data. Nonetheless, the data represented the crop conditions at a specific time. Moreover, the possible errors in the field data measured using field instruments were propagated to the MLRAs (Weiss et al., 2004). Other hyperparameter tuning strategies should be explored in the future, in addition to grid-search strategy to evaluate their consistency and effect on prediction accuracy. It is further recommended to validate (with-field reflectance data) the surface reflectance data derived from various atmospheric correction approaches (including Sen2Cor used here) as well as TOA reflectance data. This is necessary, not only to inform the selection of the most accurate input bands for estimating crop biophysical parameters but also, as recommended by Djamai and Fernandes (2018), to further our understanding of the effect of various AC approaches on such parameters. It is anticipated that improved atmospheric correction algorithms may yield better results and future studies should explore this.

5. Conclusions

This paper evaluated the performance of sPLS, RF, and GBM in estimating LAI, LC_{ab}, and CCC using Sentinel-2 data. Moreover, the spectral bands that had the greatest influence on the model accuracy were identified using RF and GBM variable importance measures. The results showed that RF was superior in estimating all three biophysical parameters, followed by GBM which was better than sPLS in estimating LAI and CCC, but not LC_{ab}, where sPLS showed relatively better prediction accuracy. Nevertheless, all MLRAs resulted in acceptable accuracy by GCOS/GMES, i.e., RRMSE of ≤10%, for all the biophysical parameters. This result is comparable to (and in some cases outperforms) studies using simulated and hyperspectral data, RTMs, and advanced MLRAs such as NN and GPR (Verrelst et al., 2012; Fernandes et al., 2014; Verrelst et al., 2015; Delloye et al., 2018). Based on the better predictive performance of sPLS when using only seven variables over GBM in estimating LC_{ab}, it is recommended that fewer (i.e., four to eight Sentinel-2 spectral subsets) should be evaluated in the future. This recommendation is consistent with previous studies (Delegido et al., 2011; Verrelst et al., 2016) that found that four to nine bands were sufficient to achieve robust estimates of biophysical parameters. Among all Sentinel-2 MSI bands, B5 (705 nm) was consistently more important in estimating all the biophysical parameters considered here. This band was previously shown to be a better estimator of chlorophyll content (Verrelst et al., 2011; Wu et al., 2008) than other RE regions.

Interestingly, the results also showed some level of dependency between canopy biophysical parameters, i.e., LAI and CCC, as the variables with the greatest influence on the model performance were somewhat similar, i.e., B5, B11, B3, and B4. This finding shows that these bands have high information content, hence high predictive power, and are highly related to both LAI and CCC; this is consistent with previous studies that found that structural and leaf parameters have a co-dependent effect on canopy spectral variations which is difficult to decouple (Ollinger, 2011; Delloye et al., 2018). Overall, our results demonstrated that the RF regression algorithm provides the most robust predictive performance for all the crop biophysical parameters considered here by efficiently utilising all Sentinel-2 MSI bands, thus it is a good contender for operationalisation. However, this assertion should be tested further in different growth stages, crops, and climatic environments, in future studies. This is essential for future satellite-based biophysical product development using a single MLRA, thus improving the prospects and effectiveness of developing accurate prescription maps to support VRA Precision Agriculture techniques and regional crop monitoring activities.

CHAPTER 6

EXPLORING TRANSFERABLE TECHNIQUES TO RETRIEVE CROP BIOPHYSICAL AND BIOCHEMICAL VARIABLES USING SENTINEL-2 DATA

This chapter is based on the following manuscript:

Kganyago, M., Adjorlolo, C. & Mhangara, P. (2022). Exploring Transferable Techniques to Retrieve Crop Biophysical and Biochemical Variables using Sentinel-2 Data. *Remote Sensing*, 14(16), 3968. DOI: https://doi.org/10.3390/rs14163968

Abstract

The current study aimed to determine the spatial transferability of eXtreme Gradient Boosting (XGBoost) models for estimating biophysical and biochemical variables (BVs), using Sentinel-2 data. The specific objectives were to: (1) assess the effect of different proportions of training samples (i.e., 25%, 50%, and 75%) available at the Target site ($\mathcal{D}_T$) on the spatial transferability of the XGBoost models and (2) evaluate the effect of the Source site ($\mathcal{D}_S$) (i.e., trained) model accuracy on the Target site (i.e., unseen) retrieval uncertainty. The results showed that the Bothaville ($\mathcal{D}_S$) → Harrismith ($\mathcal{D}_T$) Leaf Area Index (LAI) models required lower proportions, respectively, i.e., 25% or 50%, of the training samples to make optimal retrievals in the $\mathcal{D}_T$ (i.e., RMSE: 0.61 m^2 m^{-2}; R^2: 59%), while Harrismith ($\mathcal{D}_S$) →Bothaville ($\mathcal{D}_T$) LAI models required up to 75% of training samples in the $\mathcal{D}_T$ to obtain optimal LAI retrievals (i.e., RMSE = 0.63 m^2 m^{-2}; R^2 = 67%). In contrast, the chlorophyll content models for Bothaville ($\mathcal{D}_S$) → Harrismith ($\mathcal{D}_T$) required large proportions of samples (i.e., 75%) from the $\mathcal{D}_T$ to make optimal retrievals of Leaf Chlorophyll Content (LC_{ab}) (i.e., RMSE: 7.09 µg cm^{-2}; R^2: 58%) and Canopy Chlorophyll Content (CCC) (i.e., RMSE: 36.3 µg cm^{-2}; R^2: 61%), while Harrismith ($\mathcal{D}_S$) →Bothaville ($\mathcal{D}_T$) models required only 25% of the samples to achieve RMSEs of 8.16 µg cm^{-2} (R^2: 83%) and 40.25 µg cm^{-2} (R^2: 77%), for LC_{ab} and CCC, respectively. The results also showed that the source site model accuracy led to better transferability for LAI retrievals. However, the accuracy of LC_{ab} and CCC source site models did not necessarily improve their transferability. Overall, the results elucidate the potential of transferable Machine Learning Regression Algorithms and are significant for the rapid retrieval of important crop BVs in data-scarce areas, thus facilitating spatially explicit information for site-specific farm management.

1. Introduction

Accurate and reliable estimation of foliar crop biophysical and biochemical variables (BVs), such as Leaf Chlorophyll Content (LC_{ab}), Canopy Chlorophyll Content (CCC) and Leaf Area Index (LAI), is critical for facilitating Variable Rate Application (VRA) of fertilisers and the spraying of disease and fungi preventative (Maine et al., 2010; Boyer et al., 2011; Stamatiadis et al., 2020), precision irrigation (Meron et al., 2010; Monaghan et al., 2013; Chen et al., 2020), and for supporting other site-specific farm management effort and crop monitoring (Campos-Taberner et al., 2018; Dimitrov et al., 2019). Green LAI (i.e., half of the total photosynthetically active plant area [m^2] per unit ground surface area [m^2]) and chlorophyll *a* and *b* (C_{ab}) (i.e., a critical parameter related to plant photosynthetic processes), are essential for indicating the structural and physiological changes of the plant throughout the season (Wilhelm et al., 2000; Gitelson et al., 2014). However, the direct methods for determining these crop BVs are destructive, spatially limited, and have high financial implications (Bréda, 2003), despite their high accuracy. Therefore, optical handheld instruments, such as an LI-COR LAI-2200 Plant Canopy Analyzer, commonly used for LAI measurements, and chlorophyll meters such as Minolta SPAD-502 and MC-100 Chlorophyll Concentration Meter, provide rapid and non-destructive assessments of effective Green LAI and LC_{ab}. Combined with satellite imagery, these BVs can be estimated in detail over large areas and at frequent intervals that are suitable for Precision Agriculture applications and crop monitoring.

Over the years, the capabilities of satellite sensor technology have provided unparalleled opportunities for research and application development. Specifically, heritage missions such as Landsat and SPOT provided consistent data for agricultural monitoring at various scales, allowing the advancement of various approaches for the estimation of crop BVs (Myneni et al., 2002; Baret et al., 2013; García-Haro et al., 2018). Unfortunately, earlier parametric approaches, such as spectral vegetation indices (Huete et al., 2002; Gitelson et al., 2003), suffer from saturation and interferences from background signals (Mutanga & Skidmore, 2004; Jiang et al., 2008). Robust red-edge indices emerged with the advent of quasi-hyperspectral sensors such as Sentinel-2 Multi-Spectral Instrument (MSI) and Worldview-2 and its successors, which provide an increased number of multi-spectral and strategically positioned narrow bands (Fitzgerald et al., 2010; Delegido et al., 2013; Dimitrov et al., 2019). These bands were shown to improve relationships with and the accuracy of crop BVs (Mutanga & Skidmore, 2004; Ramoelo et al., 2012; Clevers & Gitelson, 2013; Yi et al. 2014; Sibanda et al., 2019).

However, vegetation indices are not transferable and may correlate with many BVs (Viña et al., 2011). Previous studies have undertaken only limited exploration of transferability techniques for estimating BVs. Although physically based approaches such as inversion of Radiative Transfer Models are considered transferable, they require local parameterisation to simulate various canopies accurately (Combal et al., 2003; Bsaibes et al., 2009). Moreover, they are complex, yield inconsistent results, and are computationally expensive (Combal et al., 2003; Atzberger 2004). In contrast, Machine Learning Regression Algorithms (MLRAs) can learn non-linear relationships between multiple variables and are robust to small and unbalanced sample sizes. Recent studies increasingly show the potential of MLRAs in accurately and reliably predicting crop BVs (Verrelst et al., 2013; Estévez et al., 2020; Amin et al., 2021). To date, the application of supervised MLRAs in remote sensing of crop BVs has been isolated and limited to specific geographic areas. Moreover, calibration data for BV models mainly depends on the crop types and phenology, geographical areas of origin, and prevailing local environmental and meteorological conditions. Unfortunately, field surveys to keep up with the Precision Agriculture and crop monitoring needs would be an expensive, impractical, and labour-intensive exercise. Therefore, the transferability of the models is essential to reduce the need for extensive training sets, reduce computational costs of calibrating MLRAs and improve prediction accuracy in data-scarce areas by utilising the knowledge learned from data-rich areas.

Transfer learning (or domain adaptation) aims to adapt existing models in the Source domain $\mathcal{D}_S$, trained to solve a specific source task $\mathcal{T}_S$, to a new but related target task $\mathcal{T}_T$ in a Target domain $\mathcal{D}_T$ (Pan & Yang 2010; Li et al. 2017). Each domain $\mathcal{D}_S \ / \ \mathcal{T}_T = \{\boldsymbol{\mathcal{X}}, P(\boldsymbol{\mathcal{X}})\}$ has a feature space $\boldsymbol{\mathcal{X}}$ and a marginal probability distribution

$P(\boldsymbol{X} = \boldsymbol{x_1}, \dots, \boldsymbol{x_n} \in \boldsymbol{\mathcal{X}})$ over the feature space. In contrast, a task $\mathcal{T}_S$ / $\mathcal{T}_T$ has a label space $\boldsymbol{\mathcal{Y}}$ and a conditional probability $P(\boldsymbol{Y}|\boldsymbol{X})$ that is learned from the training data with pairs of $\boldsymbol{x_i} \in \boldsymbol{X}$ and $\boldsymbol{\mathcal{Y}_i} \in \boldsymbol{Y}$ (Pan & Yang, 2010; Ruder, 2017). The type of transfer learning is determined by whether the labelled training set is available for both the $\mathcal{D}_S$ and $\mathcal{D}_T$ (i.e., Supervised), only a few labelled samples in the $\mathcal{D}_T$ (i.e., Semi-supervised) and no labelled samples in both $\mathcal{D}_S$ and $\mathcal{D}_T$ (i.e., Unsupervised) (Rumelhart et al., 1986; Tong et al., 2020; Zhu et al., 2020). In particular, semi-supervised transfer learning is more attractive for rapid estimation of crop BVs, as it requires only a few labelled samples in the feature space of the Target domain $\mathcal{D}_T$ to effectively model the distribution of the response variable (Persello & Bruzzone, 2012, 2014; Zhao et al., 2017; Matasci et al., 2018). Moreover, it does not require strictly matching features ($\boldsymbol{\mathcal{X}_S} \neq \boldsymbol{\mathcal{X}_T}$) and label spaces ($\boldsymbol{\mathcal{Y}_S} \neq \boldsymbol{\mathcal{Y}_T}$) between the Source $\mathcal{D}_S$ and Target $\mathcal{D}_T$ domains, and the marginal and conditional probability distributions may also differ between domains and tasks, i.e., $P(\boldsymbol{X_S}) \neq P(\boldsymbol{X_T})$ and $P(\boldsymbol{Y_S}|\boldsymbol{X_S}) \neq P(\boldsymbol{Y_T}|\boldsymbol{X_T})$, respectively. Previous studies have mostly explored the transferability of complex, black box, and computationally expensive deep learning models, e.g., Convolutional Neural Networks, in classification problems (Persello& Bruzzone, 2012, 2014; Matasci et al., 2018; Zhu et al., 2020). Only a few studies (Vuolo et al. 2013) were found that focused on the retrieval of BVs. Therefore, there is demand for accurate and transferable models based on intuitive and interpretable MLRAs that will allow retrieval of BVs for operationalisation across various spatial and temporal scales. This will facilitate Precision Agriculture and crop monitoring applications in data-scarce areas such as semi-arid areas in sub-Saharan Africa.

Against this background, this study sought to determine the spatial transferability of machine learning BV retrieval models based on the eXtreme Gradient Boosting (XGBoost) algorithm and Sentinel-2 Analysis-Ready-Data (ARD), where good spatial transferability is achieved when a model can reliably predict BVs in the Target site $\mathcal{D}_T$ (i.e., unseen) with minimal uncertainty compared to the fully trained model in that $\mathcal{D}_T$. The specific objectives were: (1) to assess the effect of different proportions of randomly selected training samples (i.e., 25%, 50%, and 75%) available at the Target site on the spatial transferability of the

XGBoost models, and (2) to evaluate the effect of the Source site (i.e., trained) model accuracy on the Target site (i.e., unseen) retrieval uncertainty. To the best of our knowledge, no study has assessed the spatial transferability of the MLRA models under various transfer scenarios that represent different proportions of training samples available in the Target site $\mathcal{D}_T$ and the effect of source model accuracy on its transferability. The $\mathcal{D}_T$ consisted of the two experimental sites, i.e., Bothaville and Harrismith.

2. Materials and Methods

2.1. Characteristics of the experimental sites

Two experimental sites, i.e., Bothaville (27°13′0″S to 28°8′0″S; 26°0′0″E to 27°05′0″E), and Harrismith (28°0′0″S to 29°0′0″S; 28°0′0″E to 29°8′0″E), located in the Free State province, South Africa were used in this study (**Figure 6-1**). The experimental sites, each with 10 000 km^2 area, are situated in the main agricultural production zone of the country. The two experience wet and warm summers, with mean temperatures of approximately 18 °C and 19.2 °C, and annual mean rainfall of approximately 584 mm and 115 mm, respectively. The summer season represents the main cropping season (i.e., December to May/June), where primarily commercial farming of main crops such as maize (*Zeal mays* L.), soybeans (*Glycine max*), dry beans (*Phaseolus vulgaris*), sunflower (*Helianthus annus* L.) and peanuts (*Arachis hypogaea* L.) in Bothaville and maize, soybeans and dry beans in Harrismith, occurs. The soil in Bothaville is characterised by sandy-loamy to sandy soils on visibly flat slopes, while Harrismith soil is dominantly clay-loamy with higher water-retention capacities on undulating slopes.

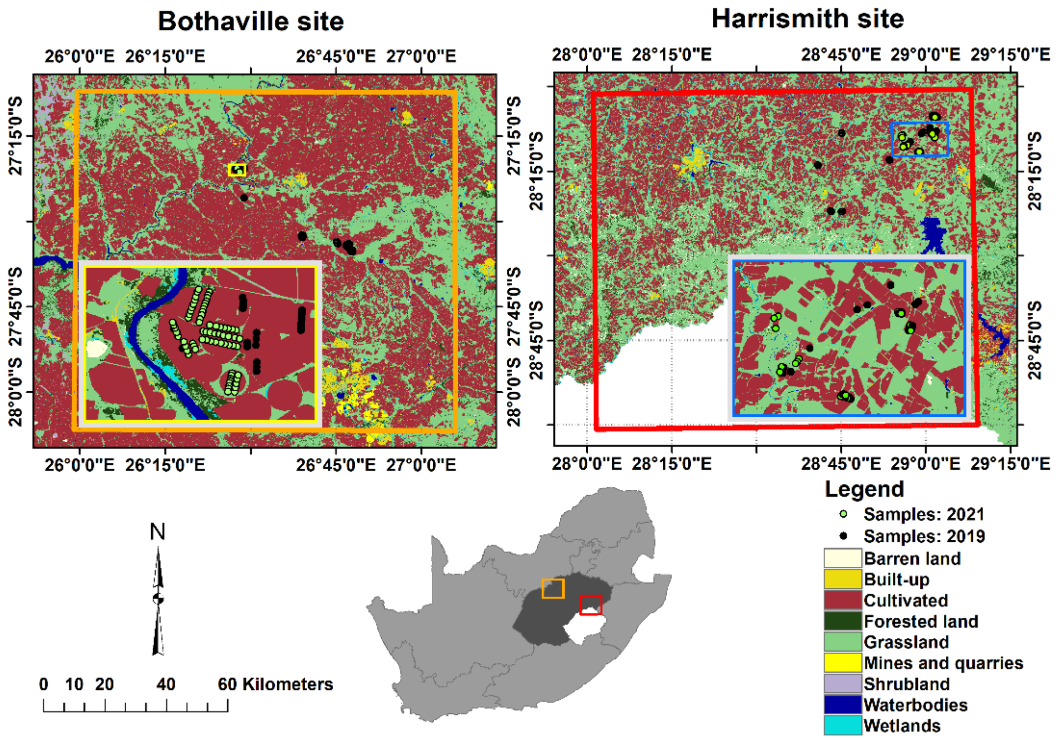


Figure 6-1. The location of Bothaville (i.e., orange square) and Harrismith (i.e., red square**)** in Free State province (dark grey), South Africa, as well as the broad land cover types and the locations of the sampling plots. The insert maps for each site (in yellow and blue) are the zoomed-in regions.

2.2 Data

2.2.1. In situ data

The multi-temporal in situ LAI and LC_{ab} data were collected in the two respective experimental sites in 2019 and 2021 at the peak of the season when crops were in their physiological maturity. In Bothaville, the data collection occurred from 8 to 14 April 2019, and 11 to 23 April 2021. In Harrismith, data was collected from 11 to 15 March 2019, and 15 to 26 March 2021. No field campaigns were conducted in 2020 owing to the national lockdown imposed to reduce the spread of the COVID-19 pandemic, which coincided with the summer growing season. At each site, LC_{ab} and LAI measurements were collected from plots of 40 m × 40 m, which were selected along a randomly selected transect. The centroids of each plot were Geo-tagged with a coordinate using a handheld Data Collector (i.e., Trimble® TDC600), which has a positional accuracy of 1.5 m. LAI measurements were collected with a LI-COR 2200c Plant Canopy Analyzer (LI-COR, Inc., Lincoln, NE, USA) in both field campaigns. To avoid the influence of the operator and unequal sky conditions on the measurements, we used a 180° view cap. Further, LC_{ab} measurements of sun-exposed leaves were taken with a Minolta SPAD-502 chlorophyll meter (Minolta, Osaka Co., Ltd., Osaka, Japan) in 2019 and MC-100 Chlorophyll Concentration Meter (Apogee Instruments, Inc., Logan, UT, USA) in 2021, respectively. MC-100 measures chlorophyll content in absolute units, i.e., $\mu mol\ m^{-2}$, realised through crop-specific and generic calibration coefficients applied to the measured transmission ratio at 931 nm to 653 nm (Parry et al., 2014). To be consistent with previous studies, we converted the chlorophyll concentration values from $\mu mol\ m^{-2}$ to $\mu g\ cm^{-2}$. In contrast, SPAD-502 provides unitless values (i.e., a relative value proportional to C_{ab} based on the ratio of transmitted and incident radiation at 940 nm and 650 nm). We used relationships established previously for the same crop types (Markwell et al., 1995; Uddling et al., 2007; Houborg & Boegh, 2008) to convert the SPAD values to LC_{ab} ($\mu g\ cm^{-2}$; Equation 6-1).

$$LC_{ab} = 10^{SPAD^{0.264}-4} \times M \quad (6\text{-}1)$$

where *SPAD* is the value read from SPAD-502Plus and M means the chlorophyll molar mass set at 907 g/mol (Zhang et al., 2019).

For each plot, the CCC was calculated by multiplying LC_{ab} by LAI (LC_{ab} × LAI) (Ustin et al., 2009). Since intercropping and mixed crop management practices dominate African agriculture, our interest was to develop generic BV retrieval models (i.e., with a potential for wide application, especially in semi-arid African contexts). Therefore, the multi-temporal in situ data for all crops found at both sites, i.e., maize (*Zea mays* L.), soybeans (*Glycine max*), dry beans (*Phaseolus vulgaris*) in Harrismith, and maize, soybeans, dry beans, sunflower (*Helianthus annuus* L.) and peanuts (*Arachis hypogaea* L.) in Bothaville, were combined and used to extract intersecting surface reflectance values from respective image dates (see Section 2.2.2) within 40 m × 40 m pixel-blocks. Since they have contrasting physiological pathways, leaf and canopy structures and architectures, these crop types represent anatomical and physiological traits of many other crop types (Nguy-Robertson et al. 2012). Therefore, we hypothesise that the BV models calibrated using these crops will likely be transferable and widely applicable.

2.2.2. Remotely sensing data

Sentinel-2 MSI ARD (Level-2A) close to the dates of the fieldwork (see Section 2.2.1) were retrieved from Sentinel Hub (Sinergise Laboratory for geographical information systems, Ltd., Ljubljana, Slovenia) using the Request-Builder for Process API (https://apps.sentinel-hub.com/requests-builder/, accessed on 15 June 2021). The ARD refers to harmonised, standardised, interoperable, and radiometrically and geometrically consistent data that eliminates the pre-processing burden to allow immediate analysis (Dwyer et al., 2018). The retrieved ARD images were acquired on 15 April 2019 (cloud cover: 0.64%) and 14 April 2021 (cloud cover: 0.29%) for Bothaville (35JMK) and 18 March 2019 (cloud cover: 51.47%) and 22 March 2021 (cloud cover: 6.11%) for Harrismith (35JPJ). We used spectral bands centred at B2:490 nm, B3:560 nm, B4:665 nm, and B8:842 nm acquired at 10 m resolution, and B5:705 nm, B6:740 nm, B7:783 nm, B8A:865 nm, B11:1 610 nm, and B12:2 190 nm at 20 m resolution. To match the resolutions between all bands, we resampled the bands with 10 m resolution to 20 m using the nearest neighbour resampling in Sentinel Application Platform (SNAP) v8.0 (http://step.esa.int). Nearest neighbour resampling was chosen based on its ability to retain the spectral fidelity of the data. The systematic ARD data processing is achieved with Sen2Cor by the European Space Agency (ESA) ground segment (Drusch et al. 2012; Louis et al. 2016; Li et al. 2018).

Further analyses were restricted to areas covered by a crop mask obtained from the National Crop Boundaries Dataset (CropEstimatesConsortium, 2017); a vegetation mask created from NDVI was also used to mask values < 0.2 following Kganyago (2021).

2.3. eXtreme Gradient Boosting

Extreme Gradient Boosting (XGBoost), proposed by (Chen & Guestrin, 2016), is an improved implementation of Gradient Boosted Regression Trees (GBRT). Also known as Gradient Boosting Machines (GBM) (Friedman 2001), these bring several additional features and advantages. As with GBM, the XGBoost is an ensemble of classification and regression trees (CART) where regression trees are computed sequentially, and new weak learners are included iteratively in the model using additive functions to correct errors from the previous learner. The final model is an ensemble of several weak learners, and final predictions are the sum of the predictions at each iteration. Generally, given the training dataset with predictor and response variables, the XGBoost algorithm first sorts the predictors and searches for the optimal splits. Then, an optimal split is chosen from the predictor that optimises the objective function, i.e., Squared Error. The above steps are repeated until the most extreme tree depth is achieved. Then, the prediction scores are assigned to the leaves, and any negative nodes are pruned using a bottom-up approach.

Until the predetermined number of rounds is reached, the above steps are repeated in a value-adding manner. Mathematical descriptions of XGBoost can be found in Ayumi (2017) and Gupta et al. (2016). The algorithm has several advantages over GBM, which include: (1) it can efficiently handle sparse data or missing values; (2) it uses distributed and parallel computing to construct trees and build large models rapidly; (3) it avoids overfitting by using a more regularised formalisation and gradient boosted decision trees; and (4) the trained model can be used to predict the unseen (i.e., new) data. Thus, XGBoost often outperforms other algorithms, and it is scalable, sparsity-aware, and computationally efficient (Chen & Guestrin 2016; Beltran et al. 2019). XGBoost has several hyperparameters which can be tuned. In this study, the hyperparameters were tuned using the random search strategy (Bergstra& Bengio, 2012), over a predefined search space consisting of general, booster-specific, and learning task parameters in "xgboost" R-package (Chen et al. 2019). The optimal hyperparameters were selected using the lowest Root Mean Squared Error of cross-validation ($RMSE_{cv}$) based on the 10-fold cross-validation (cv) resampling strategy. The open-source XGBoost library for different programming languages and documentation can be found online (https://xgboost.readthedocs.io/en/stable/index.html, accessed on 24 July 2022).

2.4. Spatial transfer scenarios

The Source site, i.e., $\mathcal{D}_S = \{(\mathcal{X}_i^s, \mathcal{Y}_i^s)\}_{i=1}^{n_s}$, refers to the study site where modelling was performed with all training samples n_s, and the Target site $D_T = \{(\mathcal{X}_i^l, \mathcal{Y}_i^l)\}_{i=1}^{n_l}$ is the destination study site where the trained models were transferred with limited training samples n_l, where $n_l \ll n_s$. In this regard, three spatial transfer scenarios were assessed, i.e., Base $\mathcal{D}_S + 25\%$, Base $\mathcal{D}_S + 50\%$, and Base $\mathcal{D}_S + 75\%$, where Base $\mathcal{D}_S$ is the pre-trained model with multi-temporal data for two spatial transferability cases: Bothaville ($\mathcal{D}_S$) → Harrismith (D_T) and Harrismith ($\mathcal{D}_S$) → Bothaville (D_T), i.e., when either Bothaville or Harrismith is the Source site. The percentages represent the proportion of available training samples used for re-training the Base $\mathcal{D}_S$ model at the Target site D_T. For example, trained BV models in the Source site $\mathcal{D}_S$ (e.g., Harrismith) were transferred to the Target site D_T (e.g., Bothaville) and vice versa. Alternating the two sites as $\mathcal{D}_S$ and D_T provided an opportunity to assess the impact of site characteristics and acquisition conditions on model transferability. The flow chart summarising the methodology is shown in **Figure 6-2**.

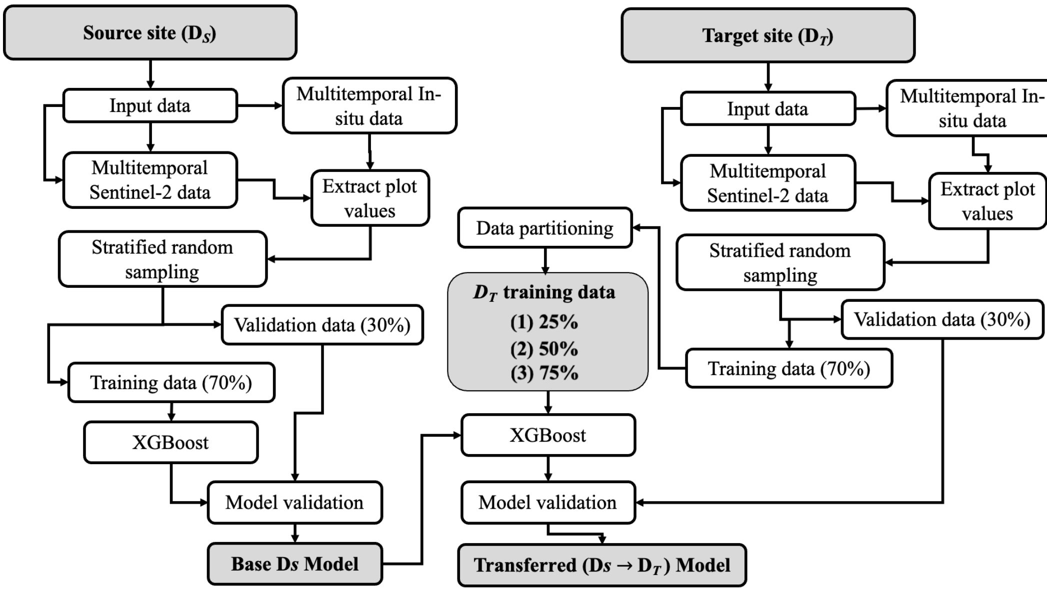


Figure 6-2. Flowchart summarising the methodology of the study in a one-directional transfer setting, e.g., Bothaville ($\mathcal{D}_S$) → Harrismith (D_T).

2.5. Model training and evaluation

Multi-temporal in situ data from each respective Source site $\mathcal{D}_S$ were used to train and evaluate the transferability of the LAI, LC$_{ab}$, and CCC models. Specifically, the Source site $\mathcal{D}_S$ data were split into training (i.e., 70%) and validation (i.e., 30%) using a stratified-random sampling approach and used for training and validating the Base $\mathcal{D}_S$ models, respectively. Then, for re-training transferred models, the

Target site D_T training data (i.e., 70%) were randomly split into 25%, 50%, and 75% of samples used in training the Base $\mathcal{D}_S$ models (**Table 6-1**), and 30% of D_T samples were held-out for evaluating the transferability of the model (i.e., its performance in a new [unseen] site) (**Figure 6-2**). The prediction accuracies of each BV model in the Source and Target sites were assessed with the measures recommended by Richter et al. (2012), i.e., Root Mean Squared Error (RMSE), Coefficient of determination (R^2), Relative Root Mean Squared Error (RRMSE), Mean Absolute Error (MAE), and Percent Bias (%Bias) (Equations 6-2 to 6-6). The modelling, validation, and mapping were performed using the R-Statistics package: mlr v2.19 (available online: https://mlr.mlr-org.com/, accessed on 5 March 2022) (Bischl et al., 2016).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - y_i)^2} \quad (6\text{-}2)$$

$$R^2 = \frac{\sum(y_i^N - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (6\text{-}3)$$

$$RRMSE = \frac{RMSE}{\bar{x}_i} \quad (6\text{-}4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (6\text{-}5)$$

$$\%BIAS = \sum_{i=1}^{n}(x_i - y_i)/\sum_{i=1}^{n}(x_i) \quad (6\text{-}6)$$

where x_i is the observed BV (e.g., LAI), and y_i is the predicted biophysical parameter (e.g., LAI), $\bar{x}_i$, and $\bar{y}_i$ are the mean of observed and predicted BV, respectively; n is the sample size, and N is the number of errors.

Table 6-1. Descriptive statistics of LAI (m^2 m^{-2}), LCab (µg cm^{-2}) and CCC (µg cm^{-2}) training samples for Source site (i.e., Base $\mathcal{D}_S$) different transfer scenarios, i.e., Base $\mathcal{D}_S + 25\%$, Base $\mathcal{D}_S +$ 50%, and Base $\mathcal{D}_S + 75\%$.

		Bothaville			Harrismith		
		LAI	**LC_{ab}**	**CCC**	**LAI**	**LC_{ab}**	**CCC**
Base $\mathcal{D}_S$	*n*	166	166	166	147	147	147
	Min	1.54	3.70	11.86	1.16	10.77	20.44
	Mean	3.62	34.53	126.98	3.45	26.83	92.56
	Max	5.90	74.02	288.22	6.32	56.82	333.63
	SD	1.04	15.57	66.99	0.94	10.89	51.25
Base $\mathcal{D}_S$ +25%	*n*	202	202	202	188	188	188
	Min	1.54	3.71	11.86	1.16	3.71	11.86
	Mean	3.57	33.07	120.18	3.53	28.35	102.25
	Max	6.35	74.02	288.22	6.32	71.35	333.63
	SD	1.01	15.06	65.33	0.98	12.33	59.63
Base $\mathcal{D}_S$ +50%	*n*	239	239	239	229	229	229
	Min	1.54	3.71	11.86	1.16	3.71	11.86
	Mean	3.56	32.04	115.80	3.51	29.97	106.89
	Max	6.35	74.02	288.22	6.32	74.01	333.63
	SD	0.99	14.65	63.66	0.98	13.82	62.47
Base $\mathcal{D}_S$ +75%	*n*	276	276	276	270	270	270
	Min	1.54	3.71	11.86	1.16	3.71	11.86
	Mean	3.55	31.36	113.41	3.55	30.58	110.48
	Max	6.35	74.01	333.63	6.32	74.01	333.63
	SD	1.00	14.30	63.31	1.01	14.06	63.15

3. Results

The evaluation of the spatial transferability of the Machine Learning Regression (MLR) models for estimating crop biophysical and biochemical variables (BVs) is critical for determining their utility in different geographical areas, with a few training samples, different spatial distributions, and spectral shifts owing to acquisition conditions and crop types. In this study, we evaluated the spatial transferability of generic BV models based on XGBoost by applying them to different target experimental sites, under different transfer scenarios, designed based on the proportion of the training data available in the Target sites, i.e., 25%, 50%, and 75%. Specifically, we applied pre-trained Source site models (from Bothaville and Harrismith), i.e., Base $\mathcal{D}_S$, to Target sites (i.e., Harrismith and Bothaville) with 25%, 50%, and 75% available, randomly selected training samples, yielding BV models with Base $\mathcal{D}_S + 25\%$, Base $\mathcal{D}_S + 50\%$, and Base $\mathcal{D}_S + 75\%$.

The performances of the Base $\mathcal{D}_S$ and transferred models to different Target sites D_T (i.e., Bothaville and Harrismith) are presented in **Figure 6-3** to **Figure 6-5** for LAI, LC_{ab}, and CCC, respectively. As shown in **Figure 6-3**a,e, Bothaville Base $\mathcal{D}_S$ LAI model had better performance (i.e., RMSE: 0.61 m^2 m^{-2}, R^2: 0.71) than the Harrismith Base $\mathcal{D}_S$ LAI model (i.e., RMSE: 0.66 m^2 m^{-2}, R^2: 56). The Bothaville Base $\mathcal{D}_S$ LAI models with 25% (i.e., Base $\mathcal{D}_S + 25\%$) and 50% (i.e., Base $\mathcal{D}_S + 50\%$) training samples in the Target site D_T (i.e., Harrismith) have a better RMSE of 0.61 m^2 m^{-2} (R: 0.59) compared to the Harrismith Base $\mathcal{D}_S$ LAI model, but exhibited relatively higher underestimations, i.e., percent bias of 5% (**Figure 6-3**b,c). In contrast, when the Bothaville Base $\mathcal{D}_S$ LAI model was re-trained with 75% of the samples in the D_T (i.e, Harrismith, **Figure 6-3**d), the results maintained a better RMSE of 0.64 m^2 m^{-2} when compared to the Harrismith Base $\mathcal{D}_S$ LAI model (i.e., RMSE: 0.66 m^2 m^{-2}), but the differences were not significant. Moreover, the transferred model explained only 52% of LAI variability in the Target site D_T (i.e., Harrismith).

When considering Harrismith as the Source site $\mathcal{D}_S$ and Bothaville as the Target site D_T, the results using Base $\mathcal{D}_S + 25\%$ LAI model (**Figure 6-3**f) showed an inferior RMSE of 0.67 m^2 m^{-2} relative to the Bothaville Base $\mathcal{D}_S$ LAI model, i.e., RMSE: 0.61 m^2 m^{-2} (**Figure 6-3**a). Similarly, the transferred model (Base $\mathcal{D}_S + 25\%$) explained relatively lower LAI variability in the Target site D_T (**Figure 6-3**f), i.e., 64%, and marginally high underestimations, i.e., by 1%. The higher proportions of training samples in D_T (i.e, Bothaville) seems to have benefitted model accuracy, with Base $\mathcal{D}_S + 50\%$ (**Figure 6-3**g) and Base $\mathcal{D}_S + 75\%$ (**Figure 6-3**h) achieving relatively better RMSEs of 0.65 m^2 m^{-2} and 0.63 m^2 m^{-2}, respectively, and explaining higher LAI variability, i.e., 66% and 67%, respectively, when compared to Base $\mathcal{D}_S + 25\%$ model (**Figure 6-3**f). While underestimations for Base $\mathcal{D}_S + 50\%$ (**Figure 6-3**g) were similar to Base $\mathcal{D}_S + 25\%$ (i.e., %Bias: 2%, **Figure 6-3**f), a 0% percent bias was found for Base $\mathcal{D}_S + 75\%$ (**Figure 6-3**h).

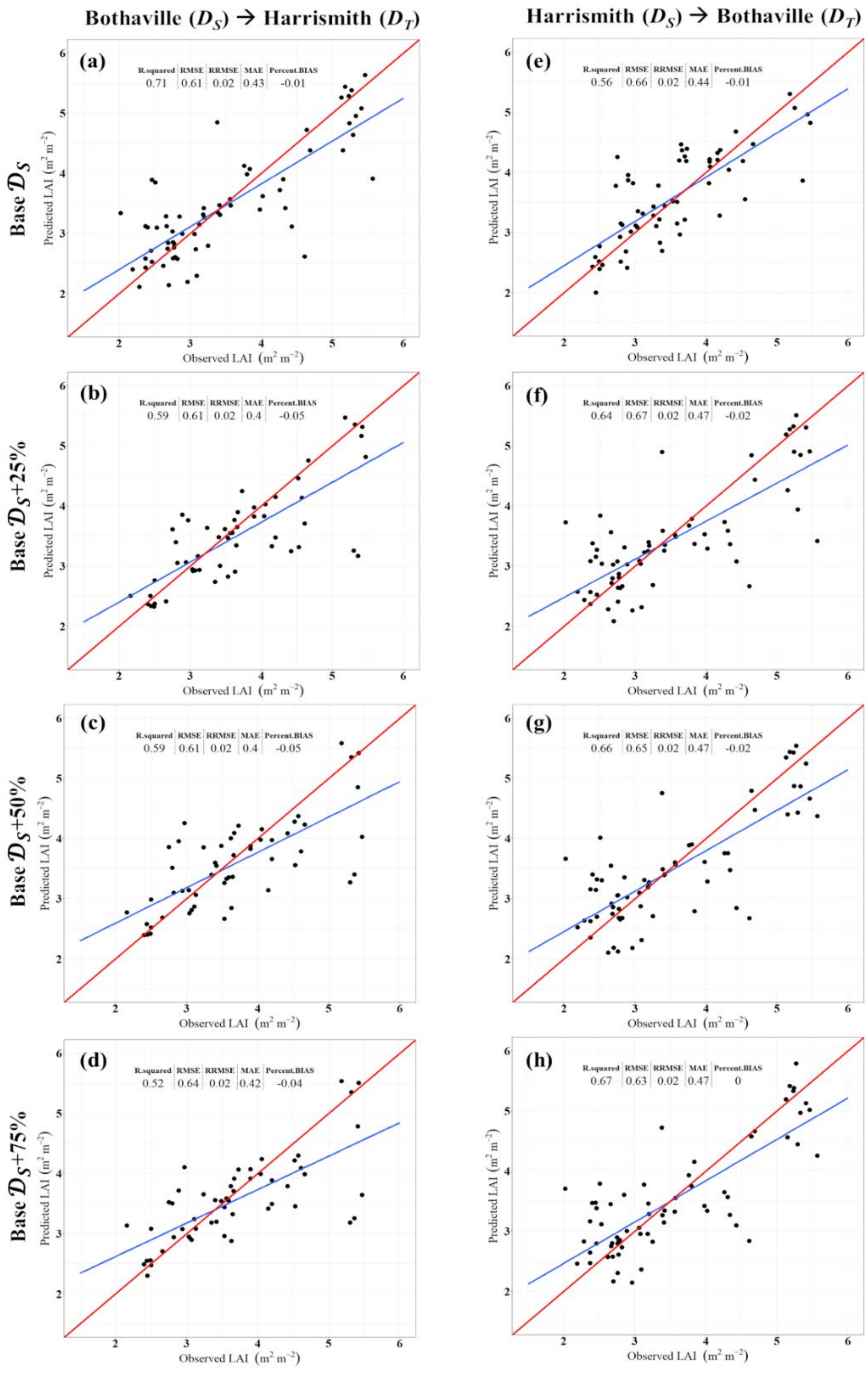


Figure 6-3. Scatterplots for Leaf Area Index (LAI, m^2 m^{-2}) indicating the performance of the XGBoost models in the Source $\mathcal{D}_S$ and Target D_T sites considering scenarios where only 25% (Base $\mathcal{D}_S$ + 25%, (b,f), 50% (Base $\mathcal{D}_S$ + 50%, (c,g), and 75% (Base $\mathcal{D}_S$ + 75%, (d,h) of the training data are available in the Target D_T site. Base $\mathcal{D}_S$ models for cases where Bothaville and Harrismith are the Source sites $\mathcal{D}_S$ are shown in panels (a,e), respectively. The 1:1 and the regression lines between the observed and predicted values are shown in red and blue, respectively.

Leaf Chlorophyll Content (LC$_{ab}$) Base $\mathcal{D}_S$ models in Bothaville and Harrismith also had different performances. The Bothaville Base $\mathcal{D}_S$ LC$_{ab}$ model (**Figure 6-4**a) explained the greatest variability, i.e., 83% compared to only 58% achieved by the Harrismith Base $\mathcal{D}_S$ LC$_{ab}$ model (**Figure 6-4**e). However, the RMSE of the Bothaville Base $\mathcal{D}_S$ LC$_{ab}$ model was higher, i.e., 8.1 µg cm^{-2}, than that of Harrismith Base $\mathcal{D}_S$ LC$_{ab}$ model, which achieved a relatively lower RMSE of 7.06 8.1 µg cm^{-2}. When the Bothaville Base $\mathcal{D}_S$ LC$_{ab}$ model was transferred to Harrismith (i.e., Bothaville ($\mathcal{D}_S$) → Harrismith (D_T)) with 25% training samples in D_T(i.e., Base $\mathcal{D}_S + 25\%$), the results (**Figure 6-4**b) indicate an RMSE, i.e., 8.41 µg cm^{-2} and explained 46% of LC$_{ab}$ variability in Harrismith. The Base $\mathcal{D}_S + 25\%$ also led to an overestimation of 4% when compared to validation data. When the proportion of training samples in the Target site D_T (i.e., Harrismith) were increased to 50%, i.e., Base $\mathcal{D}_S + 50\%$ (**Figure 6-4**c), the RMSE and R^2 improved slightly to 8.18 and 0.49, respectively. Similarly, an increase in the Target site D_T samples to 75%, i.e., Base $\mathcal{D}_S + 75\%$ (**Figure 6-4**d), improved the RMSE to 7.09 (i.e., equivalent to the one achieved by the fully trained model) and 58% of LC$_{ab}$ variability was explained in the Target site D_T (i.e., Harrismith).

In contrast, the Harrismith Base $\mathcal{D}_S$ LC$_{ab}$ model transferred to Bothaville (i.e., Harrismith ($\mathcal{D}_S$) → Bothaville (D_T)) showed improvements in RMSE, even with the smallest proportion of samples, i.e., 25%, in D_T(Base $\mathcal{D}_S + 25\%$), explaining 83% of LC$_{ab}$ variability in D_T (i.e., Bothaville) and RMSE of 8.16 µg cm^{-2} (**Figure 6-4**f), which was equivalent to the Bothaville Base $\mathcal{D}_S$ LC$_{ab}$ model (Figure 6-4a). The addition of 50% (**Figure 6-4**g) and 75% (**Figure 6-4**h) of D_T (i.e., Bothaville) training samples resulted in relatively poor LC$_{ab}$ retrievals, with the former (i.e., Base $\mathcal{D}_S + 50\%$) yielding RMSE of 8.68 µg cm^{-2} and R^2 of 0.81 (**Figure 6-4**g), while the latter (i.e., Base $\mathcal{D}_S + 75\%$) achieving inferior RMSE of 8.96 µg cm^{-2}, R^2 of 0.79 and higher underestimations of 4%. Nonetheless, all the transferred models had RRMSE of <5% in both cases, i.e., Bothaville ($\mathcal{D}_S$) → Harrismith (D_T) and Harrismith ($\mathcal{D}_S$) → Bothaville (D_T).

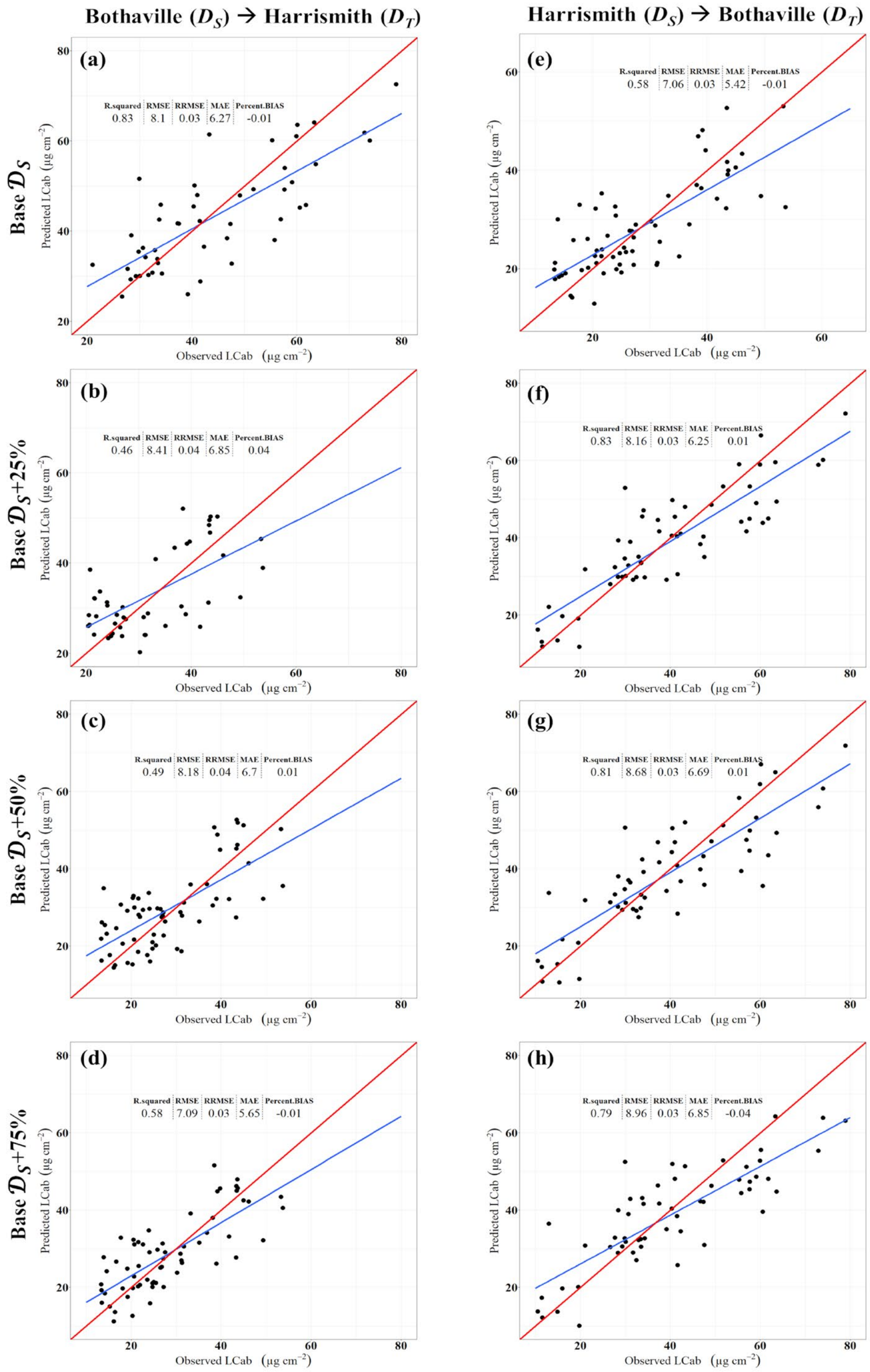


Figure 6-4. Scatterplots for Leaf Chlorophyll Content (LC$_{ab}$, μg cm^{-2}) indicating the performance of the XGBoost models in the Source $\mathcal{D}_S$ and Target D_T sites considering scenarios where only 25% (Base $\mathcal{D}_S$ + 25%, (b,f), 50% (Base $\mathcal{D}_S$ + 50%, (c,g), and 75% (Base $\mathcal{D}_S$ + 75%, (d,h) of the training samples are available in the Target D_T site. Base $\mathcal{D}_S$ models for cases where Bothaville and Harrismith are the Source sites $\mathcal{D}_S$ are shown in panels (Figures 6-4a,e), respectively. The 1:1 and the regression lines between the observed and predicted values are shown in red and blue, respectively.

The Bothaville Base $\mathcal{D}_S$ CCC model had better RMSE, i.e., 37.12 µg cm^{-2}, and explained greater variability, i.e., 79%, (**Figure 6-5**a) than the Harrismith Base $\mathcal{D}_S$ CCC model, which achieved RMSE of 38.57 µg cm^{-2} and R^2 of 0.56 (**Figure 5**e). Similar to LC$_{ab}$ models, Bothaville CCC models, i.e., Bothaville ($\mathcal{D}_S$) → Harrismith (D_T), resulted in poor retrieval when only 25% of D_T samples were used, i.e., RMSE: 40.43 µg cm^{-2}, R^2: 0.52 (**Figure 6-5**b). While proportions, i.e., 50% (i.e., Base $\mathcal{D}_S + 50\%$) and 75% (i.e., Base $\mathcal{D}_S + 75\%$), led to better retrieval accuracies, i.e., RMSE: 40.18 µg cm^{-2} (**Figure 5**c) and 36.3 µg cm^{-2} (**Figure 6-5**d), respectively. The Base $\mathcal{D}_S + 50\%$ model explained 52% of CCC variability in the Target site D_T (i.e., Harrismith), while Base $\mathcal{D}_S + 75\%$ explained 61% of CCC variability. Moreover, the $\mathcal{D}_S + 25\%$ model also had higher underestimations of CCC, i.e., %Bias of 4%, than other transferred models, which had a %Bias of only 1%. The Harrismith ($\mathcal{D}_S$) → Bothaville (D_T) model, re-trained with 25% of the samples in the Target site D_T (i.e., Bothaville) (**Figure 6-5**f), showed better CCC retrieval accuracy, i.e., RMSE: 40.25 µg cm^{-2}, R^2: 0.77, than other transfer scenarios, i.e., Base $\mathcal{D}_S + 50\%$ (RMSE: 41.39 µg cm^{-2}, R^2: 0.76) and Base $\mathcal{D}_S +$ 75% (RMSE: 42 µg cm^{-2}, R^2: 0.74). The Base $\mathcal{D}_S + 25\%$ model (**Figure 6-5**f) and Base $\mathcal{D}_S + 75\%$ model (**Figure 6-5**h) had underestimations of 3%, while the Base $\mathcal{D}_S + 50\%$ had a %Bias of 0%. All the transferred models were significant ($p < 2.2 \times 10^{-16}$).

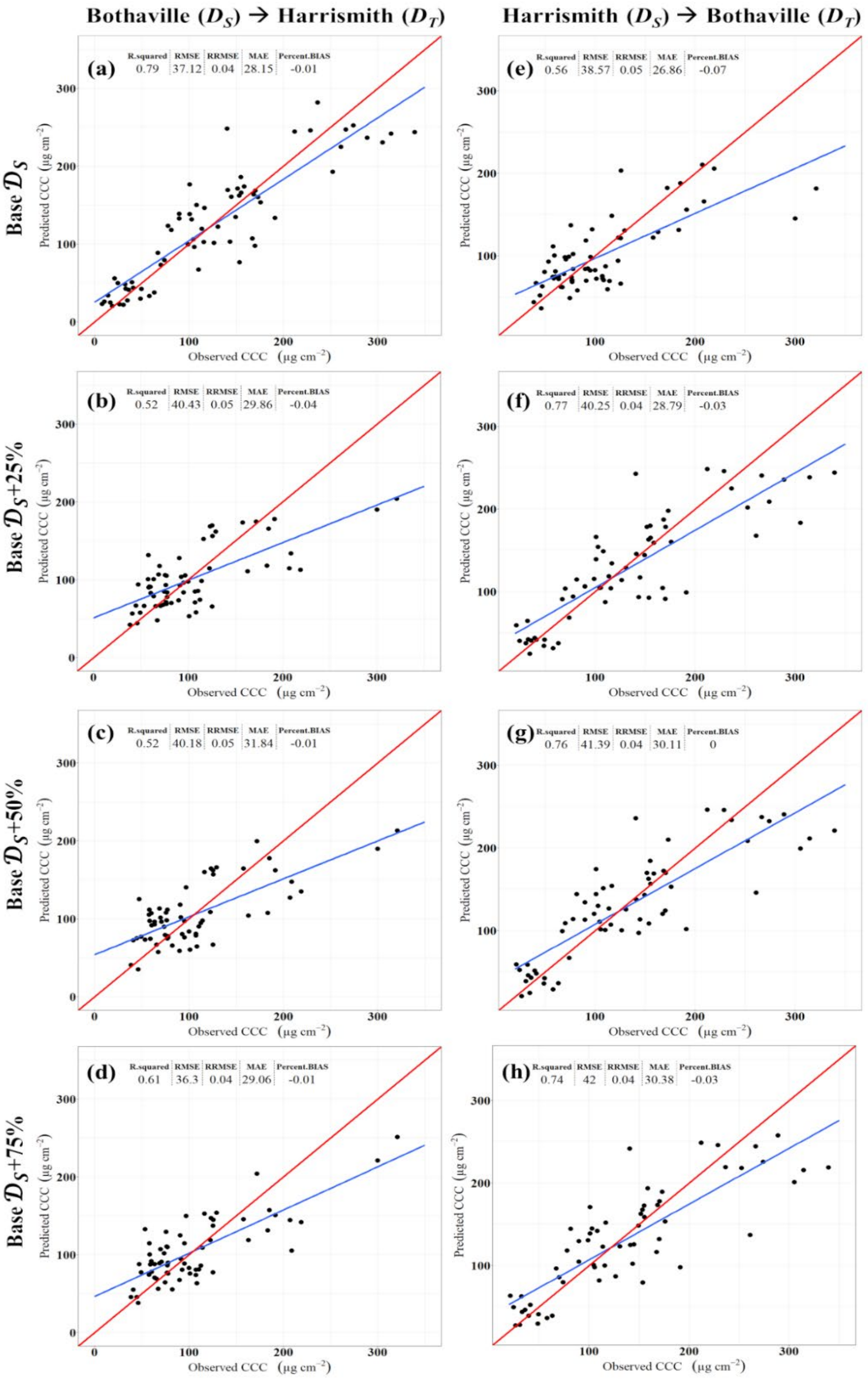


Figure 6-5. Scatterplots for Canopy Chlorophyll Content (CCC, μg cm^{-2}) indicating the performance of the XGBoost models in the Source $\mathcal{D}_S$ and Target D_T sites considering scenarios where only 25% (Base $\mathcal{D}_S$ + 25%, b,f), 50% (Base $\mathcal{D}_S$ + 50%, (c,g), and 75% (Base $\mathcal{D}_S$ + 75%, (d,h) of the training data are available in the Target D_T site. Base $\mathcal{D}_S$ models for cases where Bothaville and Harrismith are the Source sites $\mathcal{D}_S$ are shown in panels (a,e), respectively. The 1:1 and the regression lines between the observed and predicted values are shown in red and blue, respectively.

The spatial distribution maps of biophysical and biochemical variables retrieved with Source (i.e., Base $\mathcal{D}_S$) and best transferred models (D_T) are presented in **Figure 6-6** and **Figure 6-7**. Figure 6-6 shows Bothaville ($\mathcal{D}_S$) and Harrismith ($\mathcal{D}_S$) → Bothaville (D_T) transferred retrievals ((d)–(e)), while Figure 6-7 shows Harrismith ($\mathcal{D}_S$, (a)–(c)) and Bothaville ($\mathcal{D}_S$) → Harrismith (D_T) transferred retrievals ((d)–(e)). As shown, the retrievals from transferred models show underestimations of all BVs in Bothaville when compared to the Source site (i.e., Base $\mathcal{D}_S$) spatial distribution maps. On the other hand, Bothaville ($\mathcal{D}_S$) → Harrismith (D_T) maps (Figure 6-7) show relatively similar spatial distributions for all BVs.

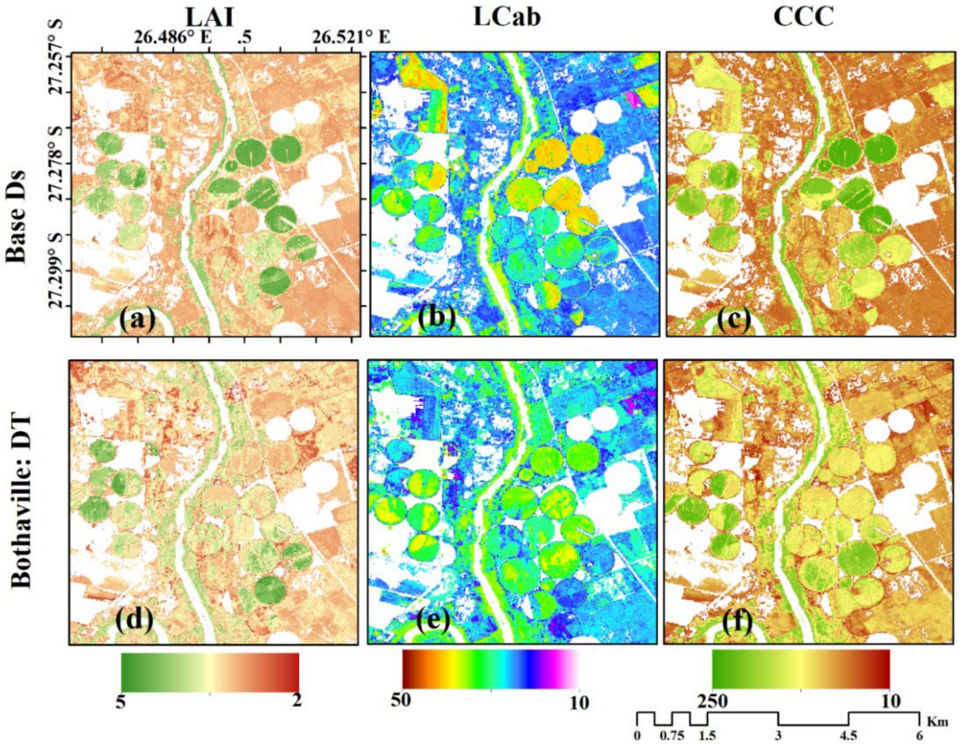


Figure 6-6. Spatial distribution maps of LAI (i.e., (a,d), LC_{ab} (i.e., (b,e), and CCC (i.e., (c,f) retrieved using XGBoost and Sentinel-2. (a–c) were retrieved with Bothaville Source (i.e., Base $\mathcal{D}_S$) models, while (d–f) were retrieved with transferred models to Bothaville D_T (i.e., Harrismith ($\mathcal{D}_S$) → Bothaville (D_T)) that achieved superior transferability, i.e., Base $\mathcal{D}_S$ + 75% for LAI (d) and Base $\mathcal{D}_S$ + 25% for LC_{ab} (e) and CCC (f).

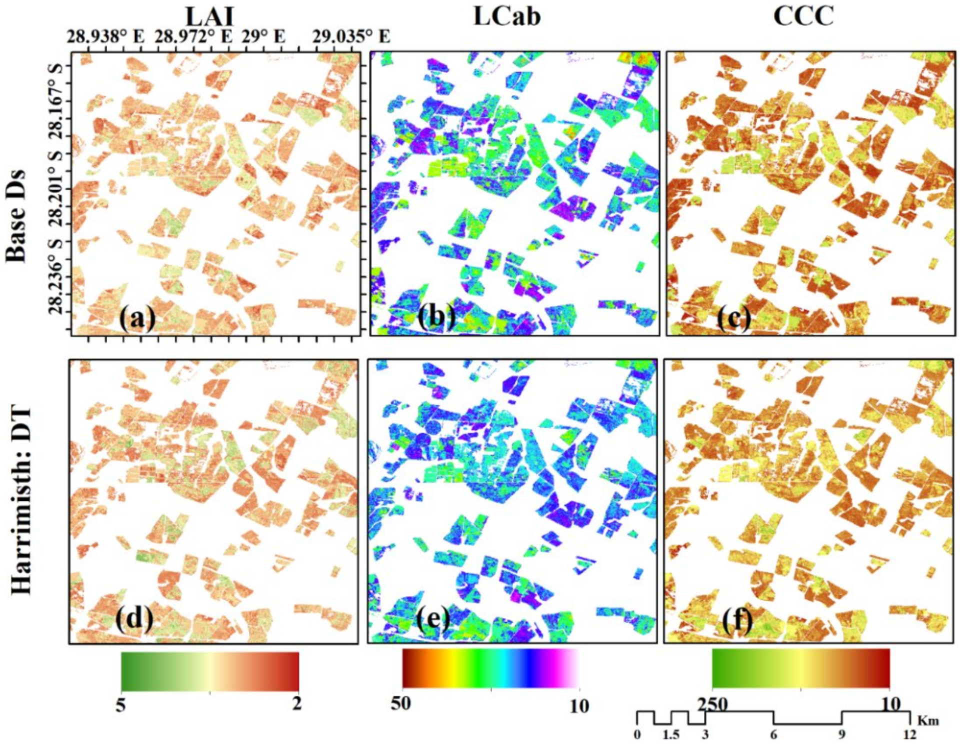


Figure 6-7. Spatial distribution maps of LAI (i.e., (a,d), LC$_{ab}$ (i.e., (b,e), and CCC (i.e., (c,f) retrieved using XGBoost and Sentinel-2. (a–c) were retrieved with Harrismith Source (i.e., Base $\mathcal{D}_S$) models, while (d–f) were retrieved with transferred models to Harrismith D_T (i.e., Bothaville ($\mathcal{D}_S$) → Harrismith (D_T)) that achieved superior transferability, i.e., Base $\mathcal{D}_S$ + 25% for LAI (d) and Base $\mathcal{D}_S$ + 75% for LC$_{ab}$ (e) and CCC (f).

4. Discussion

The transferability of retrieval approaches has been a major challenge, yet this is critical for reducing the need for large training sets, reducing computational costs associated with calibrating MLRAs. It is also important for improving the operational access to important crop BVs in data-scarce areas. For this study, we considered a case of spatial transferability technique, where a trained Source model (i.e., Base $\mathcal{D}_S$) is transferred to a new site (i.e., Target site $\mathcal{D}_T$) with different proportions of training samples available (relative to the ones used in the Base $\mathcal{D}_S$) for re-training the transferred model. Then, the performance of the transferred models was compared against the fully trained model in the Target site $\mathcal{D}_T$ to determine the optimal proportion of samples that delivers comparable results. Moreover, the effect of the Source site model (i.e., Base $\mathcal{D}_S$) accuracy on its transferability was evaluated.

4.1. Effect of training samples available in the target site

The assessment of the impact of the available training samples on the Target site $\mathcal{D}_T$ on the transferability of the MLRA biophysical variables (BVs) retrieval models is significant for determining the optimal proportion of training samples required for successfully transferring models to new (unseen) areas, critical for operational agronomic applications. Generally, the results showed that Bothaville ($\mathcal{D}_S$) → Harrismith ($\mathcal{D}_T$) models required lower proportions, respectively, i.e., 25% or 50%, of the training samples to make reliable retrievals of LAI in the Target site $\mathcal{D}_T$ (Harrismith). However, this was not true for Harrismith ($\mathcal{D}_S$) → Bothaville ($\mathcal{D}_T$) models, where the results show that the transferred models with higher proportions of training samples, i.e., 75%, were more accurate (**Figure 6-3**h). This observation implies that the proportion of samples required to make reliable LAI retrievals depends on the characteristics of the Source site $\mathcal{D}_S$. For example, the crop types varied between Bothaville and Harrismith, where the former is characterised by maize, peanuts, sunflower, and beans, while Harrismith is mainly dominated by maize and beans. Therefore, Bothaville ($\mathcal{D}_S$) → Harrismith ($\mathcal{D}_T$) models explained the variability caused by different crop types in the Target site $\mathcal{D}_T$ (i.e., Harrismith) relatively well, while Harrismith ($\mathcal{D}_S$) → Bothaville ($\mathcal{D}_T$) models required a considerable amount of samples (i.e., up to 75%) to learn new (i.e., unseen) distributions caused by the new crop types (i.e., peanuts and sunflower) with distinct structural characteristics which were encountered in the Target site $\mathcal{D}_T$ (i.e., Bothaville). Besides the different crop types, the different edaphic factors such as soil types (sandy-loamy soils vs. clay-loamy) and soil moisture between sites—known to affect surface reflectance and LAI variability (Darvishzadeh et al., 2008)—may have contributed to the requirement for a large proportion of samples to improve transferability of the Harrismith ($\mathcal{D}_S$) → Bothaville ($\mathcal{D}_T$) LAI model.

Although cropping at the two sites occurs within the same crop calendar and in situ measurements were collected when plants were in their physiological maturity, it is anticipated that there were some spectral shifts caused browning of some leaves in the canopies which may have been pronounced in Bothaville (fieldwork in April) than in Harrismith (fieldwork in March). Without the additional samples, the model resulted in relatively higher errors (**Figure 6-3**f,g). However, the improvements brought by increasing the $\mathcal{D}_T$ training samples in the case of Harrismith ($\mathcal{D}_S$) → Bothaville ($\mathcal{D}_T$) were incremental but marginal, thus signalling that the contrasting canopy architectures of maize and beans found in Harrismith played a key role in the model transferability but were not sufficiently representative to explain variability caused by other crop types present in the $\mathcal{D}_T$ (i.e., Bothaville). Nonetheless, the transferred LAI model achieved comparable accuracies (RMSE: 0.63 m^2 m^{-2}; R^2: 0.61) to a Gradient Boosting Machine model in a related previous study over $\mathcal{D}_T$ (Kganyago et al., 2021). Moreover, all the transferred LAI models performed better than the Artificial Neural Networks pre-trained with PROSAIL-simulated Look-Up-Tables (LUTs) which achieved RMSE >1 m^2 m^{-2} in Bothaville (Kganyago et al., 2020a) and elsewhere (Bochenek et al., 2017; Pasqualotto et al., 2019).

For chlorophyll content models (**Figure 6-4** and **Figure 6-5**), the Bothaville ($\mathcal{D}_S$) → Harrismith ($\mathcal{D}_T$) models required a significant proportion of $\mathcal{D}_T$ training samples to achieve a comparable retrieval accuracy to the model trained with all the training samples in Harrismith. In contrast, the Base $\mathcal{D}_S$ model experienced significant accuracy losses when transferred and re-trained with 25% (i.e., Base $\mathcal{D}_S + 25\%$) and 50% (Base $\mathcal{D}_S + 50\%$). This implies that the smaller training samples (i.e., 25% and 50%) were insignificant to account for the variability in the Target site $\mathcal{D}_T$ (i.e., Harrismith), which was probably caused by the different crop growth stages and plant structural forms present at the two experimental sites. In the case of Harrismith ($\mathcal{D}_S$) → Bothaville ($\mathcal{D}_T$), the lower proportion of $\mathcal{D}_T$ training samples (i.e., 25%) were sufficient to achieve the optimal retrieval accuracies while increasing the proportion of samples caused a decline. Generally, the results showed that the XGBoost models could be transferred to different areas with various crop types and crop phenology with minimal losses to the retrieval accuracies; however, the number of samples required for re-training the source models depended on the BV and the Source site $\mathcal{D}_S$ characteristics. The RRMSEs for all the transferred BV models under all spatial transfer scenarios were 5% and below, hence below the recommended limits by the Global Climate Observing System (GCOS) (Mason & Reading, 2007).

4.2. Source Model Accuracy and Its Effect on Transferred XGBoost Models

It is crucial to understand the effect of the source model accuracy on its transferability. Here, we evaluate the magnitude of this effect, if any, under the different transfer scenarios. The results indicated that the accuracy of the source model (i.e., Base $\mathcal{D}_S$) is important to ensure lower accuracy losses when the model is transferred to an unseen (i.e., new) site. For example, when Bothaville Base $\mathcal{D}_S$ LAI model (with RMSE: 0.61 m^2 m^{-2}; R^2: 0.71) was transferred to Harrismith (i.e., Bothaville ($\mathcal{D}_S$) → Harrismith (D_T)), only a few training samples (i.e., 25%) were required in the Target site D_T to obtain improved RMSE and R^2 by 0.05 m^2 m^{-2} and 3%, respectively, relative to the fully trained model in Harrismith (**Figure 6-3**b,e). Conversely, in the case of Harrismith ($\mathcal{D}_S$) → Bothaville (D_T), the Harrismith Base $\mathcal{D}_S$ LAI model—with RMSE of 0.66 m^2 m^{-2} and R^2 of 0.56—required a significant proportion of training samples (i.e., 75%) to achieve better accuracy, which remained lower than the fully trained model in the Source site $\mathcal{D}_S$ (**Figure 6-3**a,h). Therefore, when the Source model is highly accurate, its likelihood to perform better in the Target site (with few additional samples) increases. Essentially, the higher the variability explained by the source model, the lower the proportion of samples needed to account for the new variability in the target site.

The chlorophyll content is a highly dynamic BV resulting from to its sensitivity to various biotic (such as pests and diseases) and abiotic factors (such as water, temperature, and nutrients), hence may vary significantly within and between fields containing similar or different crop types relative to LAI, which may remain the same despite the minor changes in chlorophyll content. Interestingly, the lower variability explained by the Harrismith Base $\mathcal{D}_S$ LC_{ab} and CCC models (**Figure 6-4**e) did not negatively impact the performance of the transferred models and required only a few training samples (i.e., 25%) to explain the LC_{ab} and CCC variabilities in the Target site D_T (i.e., Bothaville) (**Figure 6-4**f and **Figure 6-5**f), and the accuracies were comparable to the fully trained models (**Figure 6-4**a and **Figure 6-5**a). While the Bothaville Base $\mathcal{D}_S$ LC_{ab} and CCC models (**Figure 6-4**a and **Figure 6-5**a) explained 83% and 79% of LC_{ab} and CCC variability, respectively, they suffered accuracy losses when transferred to the Target site D_T

(i.e., Harrismith), particularly when only 25% and 50% of D_T training samples were used (**Figure 6-4**b,c and **Figure 6-5**b,c). As a result, a high proportion of samples, i.e., 75%, were required to achieve comparable or better accuracies to the respective fully trained models in Harrismith (**Figure 6-4**e and **Figure 6-5**e). These findings may be attributed to the differences in the ranges of LC$_{ab}$ values; in the case of Bothaville ($\mathcal{D}_S$) → Harrismith (D_T), the source model could not explain unseen LC$_{ab}$ values below 20 µg cm^{-2} which were evidently present in the Target site D_T (i.e., Harrismith) but not learned by the Base $\mathcal{D}_S$ model. Whereas the Base $\mathcal{D}_S$ CCC model had a greater range, the values encountered when transferring the model could not be well interpolated using fewer training samples, i.e., 25% and 50%. Hence, for both BVs, the source models required a significant amount of training data to achieve comparable accuracies to the fully trained model in the Target site D_T (i.e., Harrismith). However, despite a relatively low LC$_{ab}$ range in the case of Harrismith ($\mathcal{D}_S$) → Bothaville (D_T), the values above 50 µg cm^{-2} were limited and well accounted for when the model was transferred and re-trained with 25% of the D_T training samples. Moreover, the overestimations of values ~20 µg cm^{-2} observed in the scatterplot of the source model from Harrismith (**Figure 6-5**e) seem to have benefitted the performance of the transferred model, since these overestimations were well within the range in the Target site D_T (Bothaville).

Because the source model contained values below ~20 µg cm^{-2} (i.e., not measured in the D_T), the transferred model increased the range of LC$_{ab}$ retrievals, which was beneficial to the transferred model. Similarly, the saturation of the Base $\mathcal{D}_S$ CCC model in Harrismith at ~200 µg cm^{-2} was averted by the additional re-training with 25% of the Target site D_T training samples and benefitted the retrieved CCC range. Overall, the results indicate that the source model (i.e., Base $\mathcal{D}_S$) accuracy is important for BVs that do not vary much at a specific crop phenological stage such as LAI at physiological maturity, while for rapidly changing BVs such as LC$_{ab}$ and CCC, the source model (i.e., Base $\mathcal{D}_S$) accuracy could be averted by re-training with some proportion of training samples. However, in both $\mathcal{D}_S$ and D_T, the range of BV values played a significant role, where ranges are a different significant number of training samples, i.e., 75%, were required.

4.3. Sources of Uncertainty

Despite the promising results, there were several potential sources of uncertainty. The degree of uncertainty observed when the LC_{ab} and CCC models were transferred can be attributed to the use of different chlorophyll instruments, i.e., Minolta SPAD-502 chlorophyll meter and MC-100 Chlorophyll Concentration Meter, between the two periods, i.e., 2019 and 2021, respectively. While MC-100 measures the absolute chlorophyll values, the SPAD-502 chlorophyll meter measures an index related to LC_{ab} and requires site-specific calibration using destructive and lab-measured chlorophyll content for different crop types and leaf structural types. In our study, empirical calibration equations developed elsewhere were used instead, based on the lack of laboratory calibration data. Unfortunately, such calibration equations may yield inconsistent results since the SPAD values saturate at $LC_{ab} > 40$ µg cm^{-2} and vary by species, growth stages, and distribution of LC_{ab} (Uddling et al., 2007; Padilla et al., 2018; Dong et al., 2019). Future studies should consider cross-calibration of the two instruments and attempt a systematic error adjustment of the field measurements taken by the two instruments. This may lead to improved accuracy of the transferred model and the replicability of the results. Moreover, the edaphic factors at each site such as soil types and moisture, as well as other site characteristics such as differences in planting dates (and thus crop phenology within and between fields), may have affected the measured crop biophysical parameters, presenting varying BV values to XGBoost per site.

One of the known limitations of MLRAs is that they are not capable of estimating beyond the input values used during training; therefore, sufficient data covering multiple years, crop phenology, and areas are needed to further calibrate the models. However, there is generally a lack of consistent and dedicated field campaigns in semi-arid areas of sub-Saharan Africa. This is hindering the development of crop biophysical variables (BVs) and the adoption of precision farming. Nonetheless, the results obtained in this study demonstrate that the XGBoost models can be spatially transferable under the assessed scenarios and are better than those obtained by Vuolo et al. (2013) who found RMSEs between 0.7 m^2 m^{-2} and 0.9 m^2 m^{-2} when Support Vector Machine and Random Forest LAI models were transferred. For all the BVs, the image quality, particularly caused by different atmospheric conditions between the acquisition dates, has affected the source model (Base $\mathcal{D}_S$) accuracies and their spatial transferability. For example, $\mathcal{D}_S$ and D_T acquisition conditions were different, i.e., >50% clouds in Harrismith for the 2019 image, which propagated to the BV models and hindered the accuracy of Harrismith source models and hence their transferability. Although Sen2cor is the Sentinel-2-specific atmospheric correction technique, its performance has been questionable.

The validation and inter-comparison studies (Martins et al., 2017; Doxani et al., 2018; Sola et al., 2018; Kganyago et al., 2020) show that it performs inconsistently for different spectral bands, land cover types, and environments. For the 2019 image over Harrismith used here, (Kganyago et al., 2020) showed that under such partly-cloud conditions, Sen2Cor performed poorly, i.e., R^2 < 20%, in the visible (VIS) and SWIR bands when validated against field data, while it had better performance, i.e., R^2 > 60%, in the red-edge and NIR bands. Previous studies show that both VIS and SWIR bands are highly influential in LAI, LC$_{ab}$ and CCC retrieval (Ollinger, 2011; Verrelst et al., 2012; Kganyago, Mhangara & Adjorlolo, 2021). Therefore, the observed lack of transferability with only 25% of samples for Harrismith ($\mathcal{D}_S$) → Bothaville (D_T) LAI models and Bothaville ($\mathcal{D}_S$) → Harrismith (D_T) LC$_{ab}$ and CCC models could be due to the residual errors after Sen2Cor atmospheric correction and lower sunlit intensity (Verrelst et al., 2013). In future, predictors should be optimised for various crop BVs according to their relative influence and quality as it has been recently shown that fewer input predictors improve spatial transferability in classification problems (Orynbaikyzy et al., 2022).

5. Conclusions

Evaluating model transferability is critical for determining the utility of the model in different geographical areas, with a few training samples, different geographical distributions, and spectral shifts resulting from acquisition conditions and crop types. In this study, we evaluated the spatial transferability of the XGBoost BV models by applying them to different experimental sites under different spatial transfer scenarios, i.e., when 25%, 50%, and 75% of samples are available in the new (unseen) site. Consequently, the effect of different proportions of available training samples in the Target site $\mathcal{D}_T$ (i.e., 25%, 50%, and 75%) were assessed, where the $\mathcal{D}_T$ consisted of the two semi-arid experimental sites, i.e., Bothaville and Harrismith. Moreover, the effect of the Source site $\mathcal{D}_S$ (trained) model accuracy on the Target site $\mathcal{D}_T$ (unseen) model uncertainty was also evaluated. The results showed some dependence on the Source site $\mathcal{D}_S$ characteristics such as crop types and plant structural characteristics for LAI retrievals, where models from Bothaville—trained with a variety of crop types—could be transferred with only a few samples, i.e., 25%, required in the $\mathcal{D}_T$ for obtaining reliable LAI retrievals. On the contrary, the Source site $\mathcal{D}_S$ LAI models from Harrismith required significantly higher proportions of training samples in the $\mathcal{D}_T$ to obtain reliable retrievals. When it comes to chlorophyll content at both leaf and canopy levels, i.e., LC_{ab} and CCC, respectively, the factors determining transferability can be attributed to its highly dynamic nature, variability within and between fields and crop types, and resulting from different field instruments used between the years, i.e., SPAD vs. MC-100. However, it was established that the data range of the trained source models matters, where the unseen low or high BV ranges may deter the transferability of the model. Moreover, the Source site model accuracy has a lesser role when the trained and unseen data ranges differ. The results obtained in this study are better than those in previous related studies, thus demonstrating prospects for achieving reliable retrievals of essential crop BV without requiring many additional samples in the target sites.

CHAPTER 7

TESTING SENTINEL-2 SPECTRAL CONFIGURATIONS FOR ESTIMATING RELEVANT CROP BIOPHYSICAL AND BIOCHEMICAL PARAMETERS FOR PRECISION AGRICULTURE USING TREE-BASED AND KERNEL-BASED ALGORITHMS

This chapter is based on the following manuscript:

Kganyago, M., Adjorlolo, C., Sibanda, M., Mhangara, P., Laneve, G., & Alexandridis, T. (**2022**). Testing Sentinel-2 Spectral Configurations for Estimating Relevant Crop Biophysical and Biochemical Parameters for Precision Agriculture Using Tree-based and Kernel-based Algorithms. *GeoCarto International*. https://doi.org/10.1080/10106049.2022.2146764

Abstract

Sentinel-2 spectral configurations, S2-10 m and S2-20 m were evaluated for retrieving essential crop biophysical and biochemical parameters and their effect on the performance of three Machine Learning Regression Algorithms (MLRAs) in two African semi-arid sites. The results were benchmarked against all spectral bands (S2-All). The results show that the S2-20 m was more robust in retrieving Leaf Area Index (LAI) ($RMSE_{cv}$: 0.58 m^2 m^{-2}, 0.47 m^2 m^{-2}), while the S2-10 m provided optimal retrievals Leaf Chlorophyll *a* + *b* (LC_{ab}) ($RMSE_{cv}$: 6.89 µg cm^{-2}, 7.02 µg cm^{-2}) for the two sites, respectively. In contrast, S2-20 m performed better in retrieving Canopy Chlorophyll Content (CCC) in Bothaville to an $RMSE_{cv}$ of 35.65 µg cm^{-2}, while S2-10 m yielded relatively low uncertainties ($RMSE_{cv}$ of 26.84 µg cm^{-2}) in Harrismith. Moreover, various MLRAs were sensitive to the various spectral configurations, and performance varied by site. GPR and XGBoost were more robust, and thus have the most potential for crop biophysical and biochemical parameter retrieval in both sites. Based on the benchmark results, the two configurations can be used independently. The results obtained here are relevant for the rapid development of essential crop biophysical and biochemical parameters for Precision Agriculture using Sentinel-2's 10 m or 20 m bands, without the need for resampling.

1. Introduction

Food and nutrition security improvement has been the principal mandate for every nation within the Sustainable Development Goals (SDGs) framework for alleviating hunger and poverty in light of population growth (Mango et al. 2017). The most significant population growth is occurring in developing countries (Walker 2016). These countries are currently affected by a marginal mismatch between the demand for food and agricultural production (Charles et al., 2014). For instance, southern Africa is facing urbanisation, and population growth which are constantly and increasingly scaling up the demand for food and emerging challenges presented by climate change and natural resources constraints. Meanwhile, agriculture is still the mainstay of many economies in southern Africa, contributing 35% of gross domestic product, employing between 70% and 80% of the population, and producing ~30% of foreign exchange. It further sustains about 70% of the livelihoods of smallholder farmers (Mango et al., 2017). Although South Africa produces surplus food, household and individual food insecurities are still glaring especially in the rural communities. The agricultural sector plays an invaluable role, and therefore, the sector needs to be optimised to bridge the gap between national and household food insecurities. There is a need for time-efficient monitoring frameworks grounded on spatially explicit technologies for near real-time monitoring of crop production indicators. Crop production indicators and attributes include the extent of cropland, irrigated cropland, crop structure, and growth parameters (i.e., chlorophyll, leaf area index, biomass) and yield (Delegido et al., 2011).

Traditional in situ, laboratory-based, and empirical point-based sampling techniques have been used to assess crop productivity. These field-based techniques are highly accurate. However, they are laborious, time-consuming, and inadequate in spatially and temporally characterising plant productivity. Therefore, they are not suitable for assessing expansive croplands. Remote sensing has emerged mainly as a non-invasive, resource-efficient method of monitoring crop productivity elements through time and space in a spatially explicit manner (Lawley et al., 2016). Specifically, the premise of monitoring crops using remotely sensed data is based on the spectral signatures or properties of crops which tend to vary with growth stage, health state, and crop type. Through time, remote sensing of crops has developed from airborne systems in the 1970s (Tucker & Maxwell, 1976; Collins, 1978) to more sophisticated satellite-based sensors such as Landsat, which offered an efficient means to repeatedly monitor agricultural crop productivity at larger scales. Although Landsat missions have been successfully used to estimate crop productivity elements in previous studies (Peng & Gitelson, 2012; Gao et al., 2017; Ma et al., 2018; Croft et al., 2020), these sensors do not cover all the critical sections of the electromagnetic spectrum such as the red-edge that is instrumental in characterising crop productivity; it is strongly associated with chlorophyll content and Leaf Area Index (LAI) variability (Chemura et al., 2017). The Earth observation community has recently witnessed the launching of the Sentinel-2 Multi-Spectral Instrument (MSI) which closes this gap, making it particularly suitable for crop productivity elements mapping.

The MSI sensors onboard Sentinel-2 2A and 2B satellites provide 13 spectral bands covering the visible (VIS), red-edge (RE), near-infrared (NIR), and shortwave infrared (SWIR) spectrums. Their revisit frequency of five days and the spatial resolutions of 10 m and 20 m present better prospects in crop biophysical and biochemical retrieval (Delegido et al., 2011). The traditional broad (i.e., 30–115 nm) VNIR bands are available at 10 m (S2-10 m), while the strategically located narrow RE and NIR bands (15–20 nm), as well as SWIR bands have 20 m resolution (S2-20 m). In this regard, data fusion techniques such as Super-Resolution for Multispectral Multiresolution Estimation (SupReMe) (Lanaras et al., 2017) and DSen2 (Lanaras et al., 2018) have been proposed for improving the spatial resolution of S2-20 m bands to match the relatively high resolution of S2-10 m without compromising the spectral consistency. Although the highest spatial resolution is often desired, Kganyago et al. (2020) show that the difference in LAI accuracy between Sentinel-2 MSI bands resampled to 10 m and 20 m spatial resolutions is negligible. Nonetheless, spatial resolutions of up to 20 m are regarded as sufficient for Precision Agriculture applications (Mulla, 2013).

While numerous studies show that LAI, Leaf Chlorophyll Content (LC_{ab}) and Canopy Chlorophyll Content (CCC) can be retrieved with the entire spectral coverage of Sentinel-2 MSI (Xie et al., 2019; Kobayashi et al., 2020; Segarra et al., 2020; Silva Junior et al., 2020), others (Delegido et al., 2013; Verrelst et al., 2016; Clevers et al., 2017) show that only a few bands are necessary for achieving high accuracies. Clevers et al. (2017), for example, found that vegetation indices constructed using S2-10 m (i.e., VNIR) were better at retrieving LAI, LC_{ab}, and CCC of potato crops, while Delegido, et al. (2011) found that the exclusion of S2-20 m RE bands resulted in systematic errors in the retrieval of LAI and CCC for multiple crops with simulated Sentinel-2 data. In other studies (Verrelst et al., 2015; Chrysafis et al., 2020; Kganyago et al., 2021), Sentinel-2 SWIR bands were identified among the most influential variables in various machine learning models for LAI, LC_{ab}, and CCC retrieval. Therefore, it is essential to evaluate the individual performance of the different Sentinel-2 spectral configurations at 10 m, i.e., characterised by broad VNIR bands (hereafter, S2-10 m), and 20 m, i.e., characterised by RE-NIR-SWIR bands (hereafter, S2-20 m) spectral bands in biophysical and biochemical parameter retrieval to demystify these inconsistencies. This is a worthy endeavour, especially since various biophysical and biochemical traits affect the various regions of the electromagnetic spectrum differently.

Meanwhile, the literature also underscores the importance of Machine Learning Regression Algorithms (MLRAs) in building models for characterising the spatial distribution of crop productivity elements. Generally, MLRAs are categorised into three types according to their architectural designs, i.e., tree-based or tree ensembles (e.g., Random Forest [RF]), kernel-based (e.g., Support Vector Machines [SVM]), and deep learning (e.g., Artificial Neural Networks [ANN]) (Rivera-Caicedo et al., 2017). Among these, tree-based, and kernel-based MLRAs are often applied for estimating crop BVs in previous studies because they are relatively uncomplicated, computationally fast, have good accuracy, and require comparatively few intuitive hyperparameters when compared to deep learning MLRAs (Li Wang et al., 2018; Shah et al., 2019; Kganyago et al., 2021). For example, LI et al. (2017) found R^2 of 88% and an RMSE of 0.195 m^2 m^{-2} in retrieving grassland LAI using RF, Landsat Enhanced Thematic Mapper (TM+) and Operational Land Imager (OLI) data. Others (Verrelst et al., 2012; Verrelst et al., 2013; Verrelst et al., 2016; Camacho et al., 2021) show that kernel-based algorithms such as Gaussian Process Regression (GPR) outperform other popular algorithms of the same family such as SVM and Kernel Ridge Regression (KRR), as well as ANN, and therefore offer better prospects for biophysical and biochemical retrieval based on its superior accuracy and unique capability to provide uncertainty estimates of the response variable. These uncertainty estimates allow the assessment of the robustness of the retrievals for operational applications. Despite the optimal performance of these MLRAs, the literature also states that no algorithm is suitable for all contexts (Ndlovu et al., 2021). Thus, their performance varies based on crop conditions and types, environments, and sensors (according to their spectral and spatial configurations) (Delloye et al., 2018). Related studies (Verrelst et al., 2012; Delloye et al., 2018) were conducted in the temperate maritime and Mediterranean climate, using simulated data, and compared complex, unexplainable, i.e., "black box", algorithms such as ANN and Kernel Ridge Regression (KRR).

In this regard, there is still a need to compare and identify relevant and effective algorithms (including less complex, more robust, and explainable algorithms) for specific contexts such as crop biophysical and biochemical parameters retrieval in semi-arid environments. Therefore, the objectives of this study were: (1) to evaluate the performance of the Sentinel-2 spectral configurations, i.e., S2-10 m (VNIR), and S2-20 m (RE-NIR-SWIR), benchmarked against all spectral bands (S2-All), in estimating crop biophysical and biochemical parameters; and (2) to determine the effect of Sentinel-2 spectral configurations on the performance of three MLRAs, i.e., Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Gaussian Process Regression (GPR), in retrieving LAI, LC_{ab} and CCC. These MLRAs were chosen based on their competitive accuracy achieved in previous studies, as well as other advantages such as their robustness, low complexity. They also require only a few hyperparameters (Verrelst et al., 2015, 2016; Rivera-Caicedo et al., 2017; Estévez et al., 2020; Pathy, Meher & P, 2020; Amin et al., 2021; Kganyago et al., 2021; Mansaray et al., 2022). The study was conducted over maize (*Zea mays L.*), beans (*Phaseolus vulgaris*), and peanuts (*Arachis hypogaea L*) characterised by contrasting physiological pathways, leaf and canopy structures, and architectures, thus offering generic models that may be widely applicable. The generic models are critical in African contexts where intercropping and mixed crop management practices are dominant. The contribution of this study is in elucidating the optimal Sentinel-2 configuration and MLRA combinations for estimating specific crop BVs in semi-arid areas. The results could inform future satellite-based product development and operational solutions for Precision Agriculture.

2. Materials and Methods

The flowchart summarising the methods followed in the current study is presented in **Figure 7-1**.

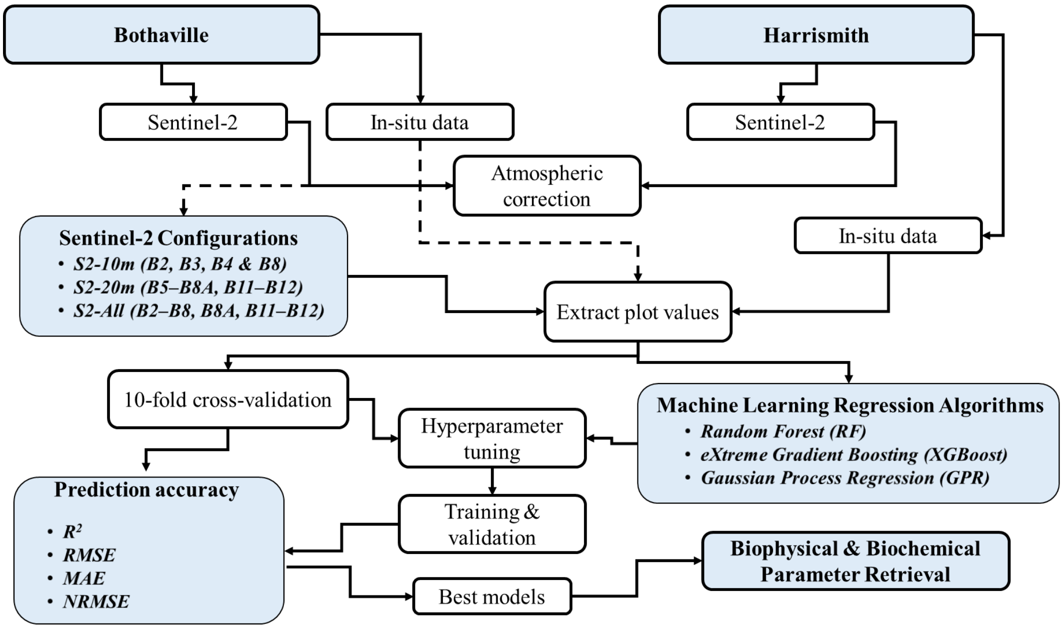


Figure 7-1. Summary of the methods followed in the study.

2.1. Experimental sites

This study was conducted in two experimental sites located in Bothaville and Harrismith in Free State province, South Africa (**Figure 7-2**). The experimental sites are situated in the main agricultural production zone of the country, i.e., Free State, with more than 3 million Ha of land cultivated. Bothaville—used as a test site in this study—is located at latitudes: 27°13′0″S to 28°8′0″S, and longitudes: 26°0′0″E to 27°05′0″E, while Harrismith—used as a validation site in this study—borders Lesotho in the South via Drakensberg Mountains and is located at latitudes: 28°0′0″S to 29°0′0″S and longitudes: 28°0′0″E to 29°8′0″E. The two sites experience warm and wet summers, with mean temperatures of ~18 °C and ~19.2

°C and annual mean rainfall of ~584 mm and ~115 mm, respectively. The summer season represents the main cropping season (i.e., from December to May or June). The Free State is dominated by medium- to large-scale commercial farming, with an average field size of 2 336 Ha (http://www.ard.fs.gov.za/wp-content/uploads/2019/10/APP-FINAL-2019-22.pdf), where the typical main crops are maize, sunflower, and groundnuts in Bothaville and maize, soybeans, and dry beans in Harrismith. The crops in Bothaville are grown on sandy to sandy-loamy soils on generally flat slopes, while Harrismith soils are clay-loamy with higher water-retention capacities on undulating slopes.

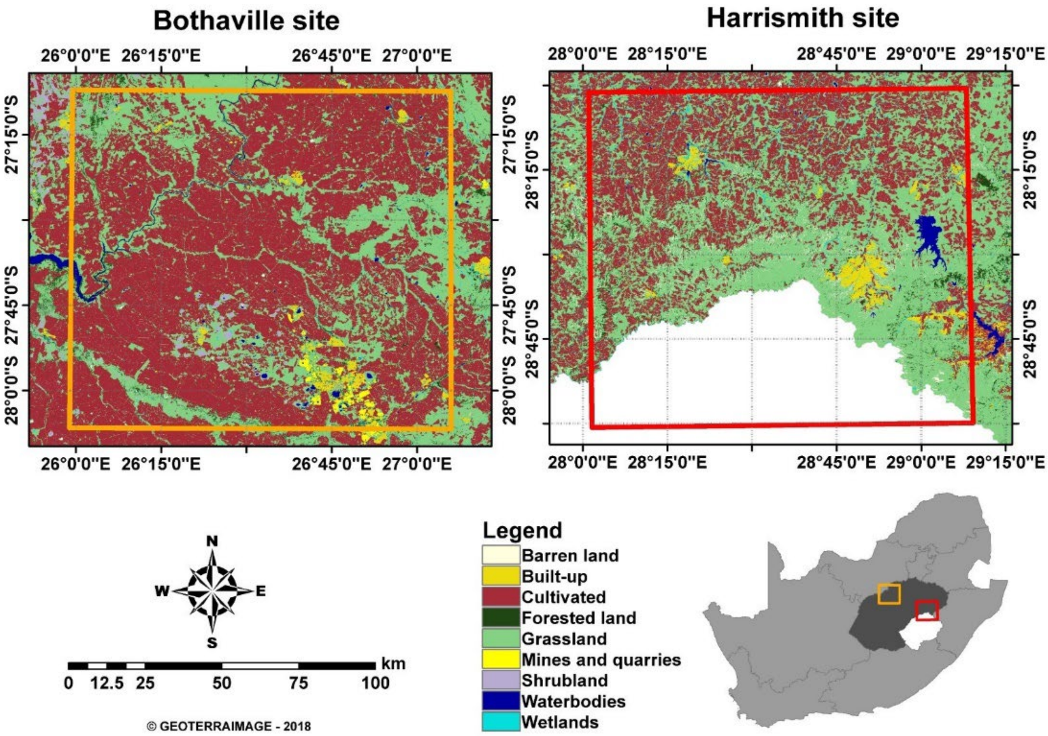


Figure 7-2. Land cover types and locations of Bothaville (orange), and Harrismith (red), in Free State province (dark grey), South Africa. Study area map adopted from Kganyago (2021).

2.2. Data

2.2.1. In situ data

The in situ LAI and LC$_{ab}$ and CCC data were collected in the field from 15 to 26 March 2021 in Harrismith and from 11 to 23 April 2021 in Bothaville. LAI and LC$_{ab}$ measurements were collected non-destructively within 40 m × 40 m plots, selected along randomly transects. Trimble® TDC600 handheld Data Collector, with global navigation satellite systems (GNSS) accuracy of 1.5 m, was used to Geo-tag the centroid of each plot and take plot pictures. Each plot consisted of an average of six to eight random measurements for each of the main crops at each site, i.e., maize (*Zea mays L.*), beans (*Phaseolus vulgaris*), peanuts (*Arachis hypogaea L*) in Bothaville and maize and beans in Harrismith. These crop types, therefore, allowed the development of generic MLRA models (i.e., with a potential for wide application) since they have contrasting physiological pathways, leaf and canopy structures, and architectures. For LAI measurements, we used LI-COR 2200c Plant Canopy Analyzer (LI-COR, Inc., Lincoln, NE, USA) in both field campaigns, with a 180° view cap to shield the influence of the operator and unequal sky conditions on the measurements. In contrast, LC_{ab} measurements were an internal average of eight to nine sun-exposed leaves at each sampling point and were collected with MC-100 Chlorophyll Concentration Meter (Apogee Instruments, Inc., Logan, UT, USA). The MC-100 is calibrated to measure chlorophyll concentration in absolute units, i.e., µmol m^{-2}, achieved through crop-specific and generic calibration coefficients which are applied to the measured ratio of transmission at 931 nm to 653 nm (Parry et al. 2014). To be consistent with previous studies, the chlorophyll concentration values in µmol m^{-2} were converted to µg cm^{-2}. The Canopy Chlorophyll Content (CCC) for each plot was estimated as a product of LC_{ab} and LAI (LC_{ab} × LAI) (Jacquemoud et al., 2009). Since our aim was not to develop crop-specific biophysical and biochemical parameters retrieval models, the field data for all crops found at each site were combined.

2.2.2. Remotely sensed data

Sentinel Hub Cloud API for Satellite Imagery (Sinergise Laboratory for geographical information systems, Ltd., Ljubljana, Slovenia) was used to retrieve the Sentinel-2 Multi-Spectral Imager (MSI) reflectance image (granule: 35JMK), acquired on 14 April 2021 over Bothaville and 22 March 2021 (granule: 35JPJ) over Harrismith. These acquisition dates coincided with the dates of field data collection at each experimental site. Sentinel-2A and 2B conjunctively provide a 5-days revisit period and carry the identical MSI sensors. MSI sensors acquire images in 13 bands at 10 m (i.e., Band 2:490 nm, Band 3:560 nm, Band 4:665 nm, and Band 8:842 nm), 20 m (i.e., Band 5:705 nm, Band 6:740 nm, Band 7:783 nm, Band

8A:865 nm, Band 11:1 610 nm, and Band 12:2 190 nm), and 60 m (i.e., Band 1:443 nm, Band 9:945 nm, and Band 10:1 375 nm) spatial resolution. The bands at 60 m were dedicated for atmospheric correction and cloud screening using Sen2cor (Drusch et al., 2012).

Sen2cor is a Sentinel-2 dedicated atmospheric correction (including cirrus clouds and terrain correction) processor. The algorithm uses the libRadtran database of Look-Up-Tables (LUTs) generated for a wide variety of atmospheric conditions, solar geometries, and ground elevations to convert the Level-1C Top-of-Atmosphere (TOA) image data to Bottom-of-Atmosphere (BOA) reflectance. The image data were corrected using parameters: atmospheric model "mid-latitude summer", aerosol type "rural" and two-band water volume retrieval (i.e., 940 nm and 1 130 nm). Further details on Sen2Cor can be obtained from Louis et al. (2016). For further analysis, the spectral bands were grouped according to their native spatial resolutions, i.e., S2-10 m (i.e., B2, B3, B4, and B8) and S2-20 m (i.e., B5, B6, B7, B8A, B11, and B12). S2-All bands consisted of the 10 m bands and 20 m bands resampled bands to 10 m using the nearest neighbour resampling technique in SNAP software v8.0 (Sentinel Application Platform, http://step.esa.int) because of its ability to maintain the spectral fidelity of the data.

2.3. Crop and green-vegetation masking

A crop mask derived from the National Crop Boundaries Dataset (CropEstimatesConsortium, 2017) was used to mask non-croplands on the Sentinel-2 bands. However, this dataset did not necessarily represent the active crop fields during the period of the current study (i.e., March and April 2021) since it is generated from SPOT 5 and 6 data acquired in 2014 and 2015. Therefore, a vegetation mask generated from the NDVI (calculated from each respective image), was used to mask non-vegetated pixels (i.e., those with NDVI < 0.2) from further analysis. This constrained further analysis to the planted crop fields in the 2021 summer growing season.

2.4. Machine Learning Regression Algorithms

The MLRAs used in this study were chosen based on their good accuracy achieved in previous studies (Verrelst et al., 2015; Verrelst et al., 2016; Rivera-Caicedo et al., 2017; Estévez et al., 2020; Pathy et al., 2020; Amin et al., 2021; Kganyago et al., 2021; Mansaray et al., 2022).

2.4.1. Random Forest

Random Forest (Breiman, 2001) is an ensemble tree-based machine learning algorithm for classification and regression and an improvement of CART. In contrast to Classification and Regression Trees (CART), Random Forest (RF) uses bagging (or bootstrapping) to iteratively and independently build a large number of decision trees (*ntree*) based on a random subset of training samples created by resampling with replacement from the original sample (Fawagreh et al., 2014, Breiman, 2001). Then, for each bootstrap sample, a decision tree is fit using randomly selected features (*mtry*), which are used to split each node in the tree (i.e., binary partitioning). Therefore, the trees grown from different and random subsets ensure increased diversity of decision trees and reduced bias of the regression (Gislason et al., 2006; Rodriguez-Galiano et al., 2012). The final regression output is obtained as an average across all trees (Gislason et al., 2006). The remaining training samples from each created random sample by bagging are called out-of-bag (OOB) data and are used for regression evaluation (Gislason et al., 2006). The optimal RF hyperparameters (i.e., *mtry* and *ntree*) for each configuration and response variable (i.e., LAI, LC_{ab}, and CCC) were tuned using the Grid-search strategy, and the optimal models were selected as those that have the lowest $RMSE_{cv}$. The *mtry* ensures that the trees in the ensemble have low bias, high variance, and lower correlation; and thus, preventing overfitting (Loggenberg et al., 2018). On the other hand, while the prediction accuracy will generally improve with increasing *ntree* up to a certain point, previous studies show that this parameter has low impact on the accuracy and can be as high as possible (Guan et al., 2013; Du et al., 2015).

2.4.2. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) is an improved implementation of Gradient Boosting Machines (GBM), also known as Gradient Boosted Regression Trees (GBRT) (Friedman, 2001), bringing several additional features and advantages. It uses gradient boosted decision trees and a more regularised formalisation to avoid overfitting, handles missing values (or sparse data) more efficiently, and employs parallel and distributed computing for rapid tree construction and building of large models, respectively. Further, it can fit new data added to the trained model. Thus, XGBoost is computationally effective and often outperforms other algorithms (Chen & Guestrin, 2016; Beltran et al., 2019). Provided with the training dataset containing predictor and response variables, XGBoost generally works as follows:

1. Sort the predictors and search for the optimal node splits,
2. Choose an optimal split from the predictor that optimises the objective function, which consists of the loss function (d) and a regularisation term (β) (Equation 7-1).

$$\Omega(\theta) = \sum_{i=1}^{n} d(y_i, \hat{y}_i) + \sum_{k=1}^{K} \beta(f_k) \qquad (7\text{-}1)$$

Where $\hat{y}_i$ is the predictive value, n is the number of instances in the training data, K is the number of trees, f_k is a tree from the ensemble of trees. In this study, the Mean Squared Error (MSE, Equation 7-2) was used as the loss function.

$$MSE = (y_i - \hat{y}_i^{(t-1)})^2 \qquad (7\text{-}2)$$

3. Repeat steps 1 and 2 until the most extreme tree depth is achieved,
4. Assign the prediction scores to the leaves, and prune any negative nodes using a bottom-up approach,
5. Repeat the above steps in a value-adding manner until the predetermined number of iterations is reached.

The XGBoost algorithm requires parameterisation of several parameters, which include the following pertinent ones for the tree booster: learning rate (eta, shrinks the feature weights and prevents overfitting), maximum tree depth (max_depth, controls the complexity of the model where a higher value result in a complex and deep tree), minimum sum of instance weight (min_child_weight, controls the partitioning of trees below which further tree partitioning would terminate), sampling ratio per tree (subsample, helps to prevent overfitting), minimum loss reduction (gamma, controls further partitioning of the tree leaf nodes where the larger value will result in a conservative model), and L1 and L2 regularisation terms on weights (alpha and lambda, respectively). The optimal hyperparameters were selected using the lowest Root Mean Squared Error of cross-validation ($RMSE_{cv}$) based on the 10-fold cross-validation (cv) resampling strategy. We refer the interested readers to excellent mathematical descriptions of XGBoost, which can be found in the original publication, Chen & Guestrin (2016), and others (Ayumi, 2017) and (Gupta et al., 2017).

2.4.3. Gaussian process regression

The Gaussian process regression (GPR) (Rasmussen, 2003) is a kernel-based probabilistic approach that establishes a relation between explanatory variables (e.g., spectral bands) and the output variable (e.g., LAI). To infer an unknown functional relationship from a training dataset, GPR elicits a prior GPR to constrain the possible form of the unknown function. Then, it updates the prior GPR in the light of training samples to generate the posterior GPR as the final functional model. A scaled Gaussian kernel is commonly used, which required hyperparameters, signal (v, σ_b) and noise σ_n, i.e., $\theta = \{v, \sigma_b, \sigma_n\}$. These hyperparameters θ combat model overfitting and are typically selected by Type-II Maximum Likelihood, using the analytical marginal likelihood (also called evidence) of the observations (Verrelst et al., 2016). Often, the derivatives of the log-evidence are also analytical; thus, conjugated gradient ascent is typically used for optimisation (Camps-Valls et al., 2009). The GPR has recently gained popularity owing to its competitive accuracy and capability to provide uncertainty estimates of the response variables (Camps-Valls et al., 2009; Verrelst et al., 2012; Verrelst et al., 2012; Verrelst et al., 2013; Verrelst et al. 2016; Camacho et al. 2021). It was selected in the current study because of its high accuracy, robustness to overfitting, and rapid training speeds. The GPR hyperparameters for this study were automatically optimised in ARTMO software (Available online: https://artmotoolbox.com/, accessed: 27 October 2021) based on the training data, using 10-fold cv, where the optimal combination of hyperparameters used for training the models was selected as the one that minimised the prediction error ($RMSE_{cv}$). For a detailed account of GPR in remote sensing, we refer readers to Camps-Valls et al. (2019) and others that applied it to biophysical and biochemical retrieval (Verrelst et al., 2012; Verrelst et al., 2013; Verrelst et al. 2015; Verrelst et al., 2016; Estévez et al., 2020; Amin et al., 2021).

2.5. Model training and validation

For training and validation, this study used k-fold cross-validation, i.e., k = 10 for this study, to ensure that all data are used for both training and validation instead of the traditional split into 70% training vs 30% validation (Snee, 1977; Verrelst et al., 2015; Shah et al., 2019). Prior to model training and validation, average pixel values were extracted from the intersecting image pixels within plot blocks of 40 m × 40 m. During the k-fold cross-validation (cv), the dataset is randomly divided into equal k sub-datasets. Then, a training dataset is formed by k-1 sub-datasets, while a validation dataset is formed by a one k sub-dataset. The final estimation value is a combination of the iterative validation steps, i.e., k times, using one of the k sub-datasets each time.

Table 7-1. Descriptive statistics of measured LAI (m^2 m^{-2}), LC_{ab} (µg cm^{-2}) and CCC (µg cm^{-2}) at the two sites.

	Bothaville					Harrismith				
	n	Min	Mean	Max	SD	n	Min	Mean	Max	SD
LAI	172	1.78	3.37	5.75	0.90	179	1.16	3.54	6.17	0.88
LC_{ab}	172	3.32	29.09	63.62	14.75	179	10.77	27.71	56.83	10.57
CCC	172	7.87	104.25	339.09	71.23	179	20.44	96.81	282.54	43.68

The prediction accuracies of each MLR model and the experimental scenario were assessed using 10-fold cv with the Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Normalised RMSE (NRMSE) (Equation 7-1 to 7-4) as recommended by Richter et al. (2012).

$$R^2 = \frac{\sum(y_i^n - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (7\text{-}1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \quad (7\text{-}2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (7\text{-}3)$$

$$NRMSE = 100 \times \left(\frac{RMSE}{y_{max} - y_{min}}\right) \quad (7\text{-}4)$$

Where y_i and $\bar{y}_i$ in Equation 7-1 denote the biophysical or biochemical predictions and mean of the observed (or measured) biophysical or biochemical parameter (e.g., LC_{ab}), respectively, while x_i and y_i in Equations 7-2 to 7-3 denote the observed and predicted biophysical or biochemical parameter (e.g., LC_{ab}), respectively, and n is the number of samples. y_{max} and y_{min} in Equation 7-4 denote the maximum and minimum values of the observed values.

All model building, prediction accuracy assessment, and biophysical and biochemical parameter mapping were performed in MATLAB based software application, i.e., ARTMO version 3.29 (Available online: https://artmotoolbox.com/, accessed: 27 October 2021), using MLRA Toolbox (Camps-Valls et al., 2013).

3. Results

This study evaluated the performance of the various Sentinel-2 configurations, i.e., S2-10 m (VNIR) and S2-20 m (RE-NIR-SWIR), in estimating LAI, LC_{ab}, and CCC using three Machine learning regression algorithms, i.e., RF, XGBoost, and GPR. The resulting accuracies for each crop biophysical and biochemical parameter were benchmarked against all spectral bands (S2-All) resampled to 10 m—the highest spatial resolution available from Sentinel-2.

3.1. Crop biophysical and biochemical parameter retrieval accuracies using MSI configurations

The two Sentinel-2 MSI configurations, i.e., S2-10 m and S2-20 m, showed varying performances for different biophysical and biochemical parameters (**Table 7–2** and **Table 7–3**). For LAI, S2-20 m resulted in consistently superior performance between the two sites, where the highest $RMSE_{cv}$ of 0.58 and 0.47 m^2 m^{-2} were achieved for Bothaville and Harrismith, respectively. Consistently, S2-20 m explained the greatest variability, i.e., 58% and 72%, when compared to S2-10 m, which explained only 52% and 64% for the two sites, respectively. A benchmark against the full MSI spectral data (i.e., S2-All) indicated consistently similar performances with S2-20 m between the two sites.

The results for LC_{ab} (also shown in **Table 7–2** and **Table 7–3**) showed that S2-10 m was superior to S2-20 m in Bothaville, with $RMSE_{cv}$ of 6.89 µg cm^{-2} (R^2: 0.79), while S2-20 m only achieved $RMSE_{cv}$ of 7.34 µg cm^{-2} (R^2: 0.75). However, in Harrismith, the two configurations resulted in equivalent retrieval accuracies, with $RMSE_{cv} \approx 7.0$ µg cm^{-2} ($R^2 \approx 0.55$). When benchmarking S2-10 m LC_{ab} results (in Bothaville) with S2-All, the results show that it outperforms S2-All., while in

Harrismith, S2-All slightly outperformed both S2-10 m and S2-20 m. Lastly, S2-20 m resulted in the most robust estimates of CCC in Bothaville, with $RMSE_{cv}$ of 35.65 µg cm^{-2} and explained 76% of CCC variability when compared to S2-10 m ($RMSE_{cv}$: 37.66 µg cm^{-2}; R^2: 0.73). However, contradictory results were found in Harrismith, where S2-20 m was relatively poor, achieving $RMSE_{cv}$ of 28.17 µg cm^{-2} (R^2: 0.58) when compared to the relatively good estimates of S2-10 m ($RMSE_{cv}$: 26.84 µg cm^{-2}; R^2: 0.62). The benchmarking (i.e., S2-All) results were worse than those obtained for Bothaville with S2-20 m and Harrismith with S2-10 m. Overall, both spectral configurations (i.e., S2-10 m and S2-20 m) also achieved $NRMSE_{cv}$ of <20%, with the highest $NRMSE_{cv}$, i.e., ≈11%, being achieved for LAI and CCC in Bothaville, and all biophysical and biochemical parameters in Harrismith.

Table 7-2. The performance of S2-10 m (VNIR), S2-20 m (RE-NIR-SWIR), S2-All (all) spectral bands for estimating LAI (m^2 m^{-2}), LC_{ab} (µg cm^{-2}), and CCC (µg cm^{-2}) with three MLRAs in Bothaville. The **bold** formatted numbers indicate the lowest $RMSE_{cv}$ achieved for each biophysical and biochemical parameter and spectral configuration.

		S2-10 m			S2-20 m			S2-All		
		RF	XGBoost	GPR	RF	XGBoost	GPR	RF	**XGBoost**	GPR
LAI	R2	**0.52**	0.52	**0.52**	**0.58**	0.57	0.54	0.55	**0.59**	0.52
	$RMSE_{cv}$	**0.62**	0.63	**0.62**	**0.58**	0.59	0.61	0.60	**0.58**	0.63
	MAE_{cv}	**0.41**	0.43	**0.44**	**0.38**	0.41	0.42	0.40	**0.42**	0.44
	$NRMSE_{cv}$	**15.72**	15.78	**15.66**	**14.80**	14.93	15.44	15.22	**14.67**	15.75
LCab	R2	0.77	**0.79**	0.77	0.74	0.75	**0.75**	0.75	0.75	**0.75**
	$RMSE_{cv}$	7.08	**6.89**	7.11	7.50	7.47	**7.34**	7.36	7.38	**7.30**
	MAE_{cv}	5.48	**5.40**	5.46	5.66	5.66	**5.53**	5.51	5.63	**5.61**
	$NRMSE_{cv}$	11.74	**11.42**	11.78	12.44	12.39	**12.18**	12.21	12.24	**12.10**
CCC	R2	0.69	**0.73**	0.70	0.74	**0.76**	0.70	0.73	0.74	0.72
	$RMSE_{cv}$	39.57	**37.66**	38.92	36.54	**35.65**	39.17	36.88	36.84	37.67
	MAE_{cv}	26.71	**26.73**	26.66	25.23	**25.32**	26.52	25.31	26.48	25.71
	$NRMSE_{cv}$	11.95	**11.37**	11.75	11.03	**10.76**	11.83	11.14	11.12	11.37

Table 7-3. The performance of S2-10 m (VNIR), S2-20 m (RE-NIR-SWIR), S2-All (all) spectral bands for estimating LAI (m^2 m^{-2}), LC_{ab} (μg cm^{-2}), and CCC (μg cm^{-2}) with three MLRAs in Harrismith. The **bold** formatted numbers indicate the lowest RMSE*cv* achieved for each biophysical and biochemical parameter and spectral configuration.

		S2-10 m			S2-20 m			S2-All		
		RF	XGBoost	GPR	RF	XGBoost	GPR	RF	XGBoost	GPR
LAI	R^2	0.54	0.55	**0.64**	0.54	0.54	**0.72**	0.61	0.59	**0.71**
	RMSE*cv*	0.59	0.60	**0.53**	0.60	0.61	**0.47**	0.56	0.58	**0.48**
	MAE*cv*	0.45	0.47	**0.40**	0.39	0.46	**0.31**	0.43	0.45	**0.32**
	NRMSE*cv*	11.89	12.00	**10.52**	11.98	12.14	**9.31**	11.26	11.56	**9.51**
LC_{ab}	R^2	0.55	0.54	**0.56**	0.53	0.53	**0.57**	0.55	0.57	**0.57**
	RMSE*cv*	7.09	7.14	**7.03**	7.21	7.29	**7.02**	7.10	6.96	**6.92**
	MAE*cv*	5.53	5.50	**5.48**	5.63	5.69	**5.32**	5.48	5.37	**5.29**
	NRMSE*cv*	15.40	15.52	**12.25**	15.65	15.82	**15.24**	15.42	15.11	**15.03**
CCC	R^2	0.60	0.61	**0.62**	**0.58**	0.57	0.57	0.59	**0.61**	0.59
	RMSE*cv*	27.41	27.46	**26.84**	**28.17**	28.81	28.71	27.95	**27.55**	28.08
	MAE*cv*	20.12	20.44	**19.93**	**20.93**	21.44	20.85	20.99	**20.64**	20.62
	NRMSE*cv*	10.46	10.48	**10.24**	**10.75**	10.99	10.96	10.67	**10.51**	10.71

3.2. Comparison of MLRAs accuracies under various spectral configurations

The three MLRAs, i.e., RF, eXtreme Gradient Boosting (XGBoost), and GPR, were evaluated for their retrieval accuracy under various Sentinel-2 MSI spectral configurations, i.e., S2-10 m, S2-20 m, and S2-All. This was particularly crucial for elucidating the effect of various Sentinel-2 MSI spectral configurations on the performance of these MLRAs. The results for Bothaville (**Table 7-2**) showed that all the MLRAs considered here performed proportionately in estimating LAI with S2-10 m, achieving $RMSE_{cv} \approx 0.62$ m^2 m^{-2} and equivalent R^2 of 0.58. The analysis in Harrismith (**Table 7-3**)—performed to confirm the consistency in the performance of MLRAs under the same spectral configurations—generally showed similar patterns to Bothaville, showing that the retrieval accuracy between MLRAs was marginal with a maximum $RMSE_{cv}$ difference of 0.07 m^2 m^{-2}.

When the MLRAs were evaluated under the S2-20 m and S2-all configurations, the results showed similar patterns to the S2-10 m results, especially in Bothaville where the $RMSE_{cv}$ differences between MLRAs were only up to 0.02

m^2 m^{-2} and 0.03 m^2 m^{-2}, respectively (**Table 7-2**). In Harrismith (**Table 7–3**), the same is observed between RF and XGBoost, with both configurations (i.e., S2-20 m and S2-All) achieving $RMSE_{cv}$ differences of only 0.01 m^2 m^{-2} and 0.02 m^2 m^{-2}, respectively. Conversely, there were marked differences between GPR and RF with $RMSE_{cv}$ differences of 0.15 m^2 m^{-2} and 0.10 m^2 m^{-2} for the S2-20 m and S2-All, respectively. Overall, the S2-20 m-RF and S2-All-XGBoost models were equivalently the best models for the retrieval of LAI in Bothaville with $RMSE_{cv}$ of 0.58 m^2 m^{-2} (R^2: 0.58), while S2-20 m-GPR and S2-All-GPR models offered the best performances in Harrismith with $RMSE_{cv}$ of 0.47–0.48 m^2 m^{-2} (R^2: 0.72–0.71).

In general, the retrieval of chlorophyll content at the leaf level (i.e., LC_{ab}) and canopy level (i.e., CCC) with RF, XGBoost, and GPR showed no superior single MLRA across different MSI configurations and sites. The results showed marginal differences, i.e., < 1 μg cm^{-2} in $RMSE_{cv}$ between MLRAs across all MSI configurations and sites, except for the CCC-XGBoost model in Bothaville which exhibited higher $RMSE_{cv}$ differences between all MLRAs with a magnitude of 1.91 μg cm^{-2} when using S2-10 m and 3.52 μg cm^{-2} between XGBoost and GPR when using S2-20 m. For LC_{ab}, the best retrieval accuracies across all configurations were achieved with the S2-10 m-XGBoost model ($RMSE_{cv}$: 6.89 μg cm^{-2}; R^2: 0.79) and S2-All-GPR model ($RMSE_{cv}$: 6.92 μg cm^{-2}; R^2: 0.57) in Bothaville and Harrismith, respectively. In contrast, for CCC, S2-20 m-XGBoost ($RMSE_{cv}$: 35.65 μg cm^{-2}; R^2: 0.76) and S2-10 m-GPR ($RMSE_{cv}$: 26.84 μg cm^{-2}; R^2: 0.62) were the best models across all configurations in Bothaville and Harrismith, respectively.

In summary, the optimal MLRAs for retrieving crop biophysical and biochemical parameters in Bothaville (**Figure 7-3**) and Harrismith (**Figure 7-4**) were achieved with XGBoost and GPR, respectively. In Bothaville, the MSI spectral configurations for optimal LAI, LC_{ab}, and CCC retrievals were S2-All, S2-10 m, and S2-20 m, respectively, achieving $RMSE_{cv}$ of 0.58 m^2 m^{-2}, 6.89 μg cm^{-2} and 35.65 μg cm^{-2}. In Harrismith, S2-20 m, S2-All, and S2-10 m were the optimal MSI configurations, providing $RMSE_{cv}$ of 0.47 m^2 m^{-2}, 6.92 μg cm^{-2} and 26.84 μg cm^{-2} for the three biophysical and biochemical parameters, respectively. These models (consisting of optimal MSI configurations and MLRAs) were applied to mapping the biophysical and biochemical parameters at the two sites (**Figure 7-5** and **Figure 7-6**). Across all the evaluated MLRAs and MSI spectral configurations, $NRMSE_{cv}$ for LAI, LC_{ab} and CCC were generally below 16%.

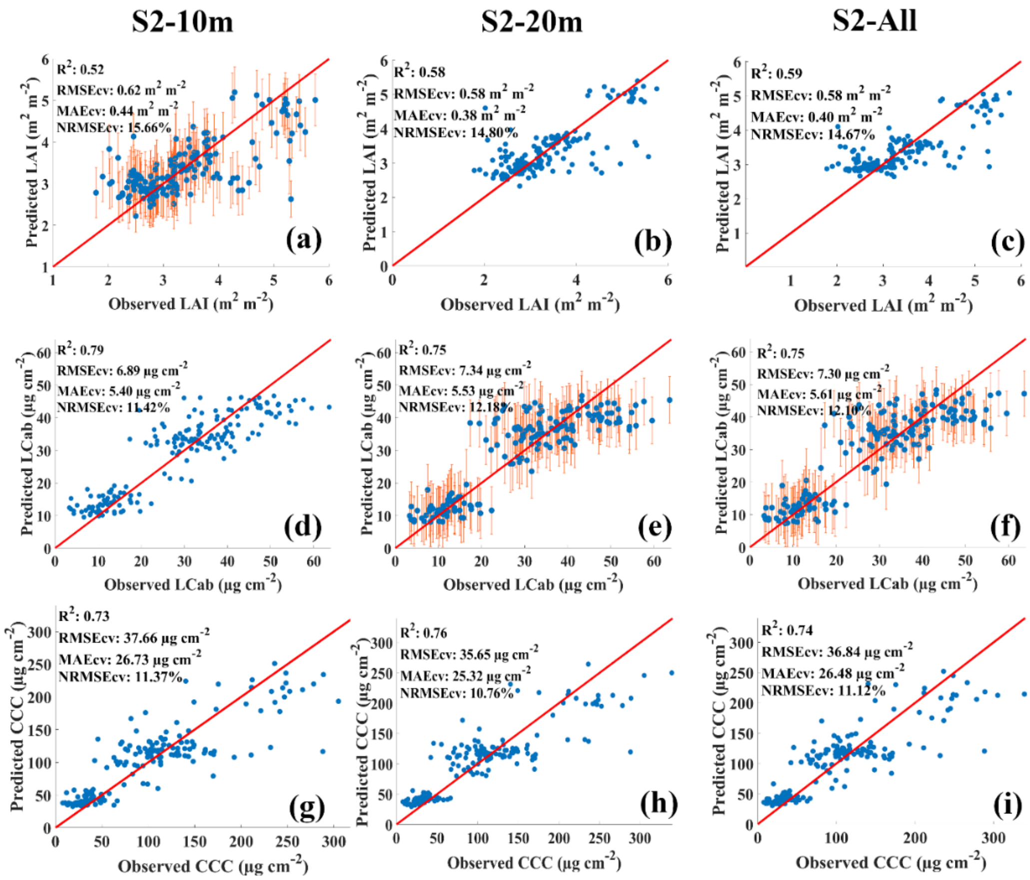


Figure 7-3. Scatterplots of the best MLRAs for each of the spectral configurations, i.e., S2-10 m ([a], [d], [g]), S2-20 m ([b], [e], [h]) and S2-All ([c], [f], [i]) in Bothaville. (a), (b), and (c) show the best LAI results obtained by GPR using S2-10 m, RF using S2-20 m, and XGBoost using S2-All, respectively. (d), (e), and (f) show the best LC_{ab} results obtained by XGBoost using S2-10 m, GPR using S2-20 m and S2-All, respectively. Lastly, (g), (h), and (i) show the best CCC results obtained by XGBoost using S2-10 m, S2-20 m, and S2-All, respectively.

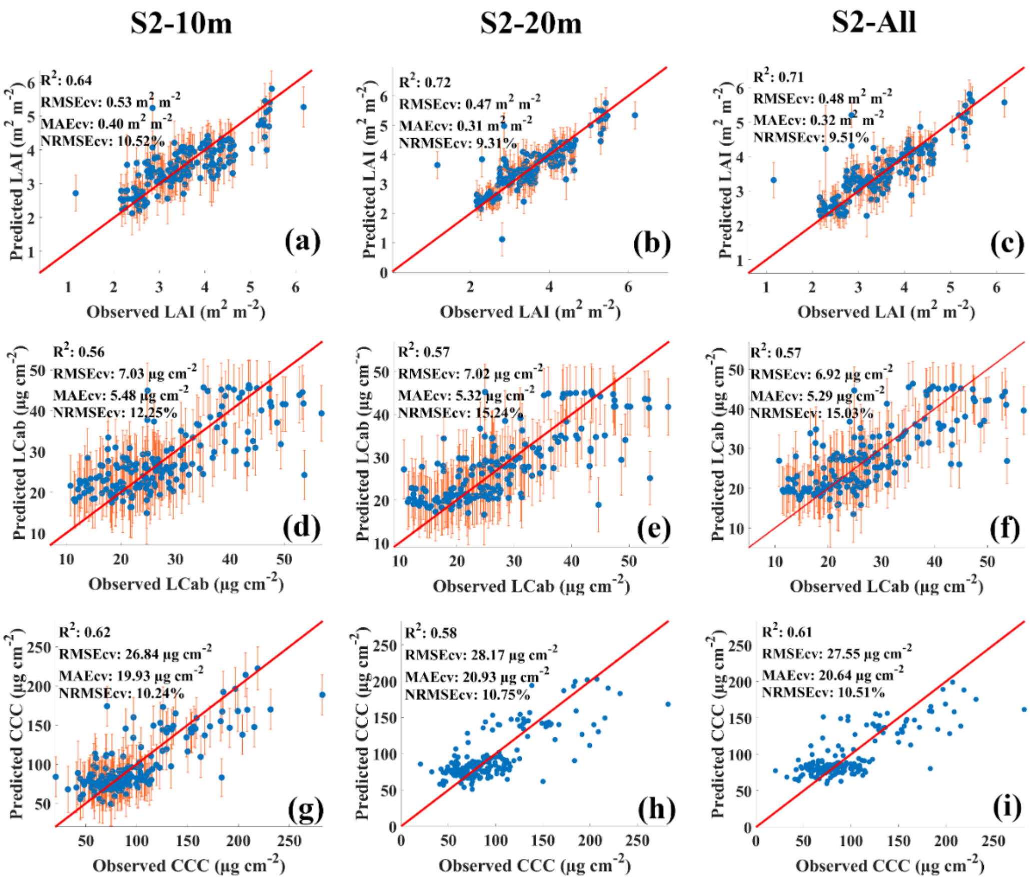


Figure 7-4. Scatterplots of the best MLRAs for each of the spectral configurations, i.e., S2-10 m ([a], [d], [g]), S2-20 m ([b], [e], [h]) and S2-All ([c], [f], [i]) in Harrismith. The best LAI ([a] – [c]) and LC_{ab} results (d) - (f) were by GPR for all MSI configurations, i.e., using S2-10 m, S2-20 m, and S2-All. (g), (h), and (i) show the best CCC results obtained by GPR using S2-10 m, RF using S2-20 m, and XGBoost using S2-All, respectively.

3.3. Spatial distribution maps for optimal MSI spectral configurations and MLRA models

The spatial distribution maps of LAI, LC_{ab} and CCC from S2-10 m and S2-20 m and the best MLRA models (i.e., corresponding to the scatter plots above) at the two sites are given in **Figure 7-5**, while the best GPR models and their associated uncertainty layers (i.e., coefficient of variation, CV) are presented in **Figure 7-6**. **Figure 7-5**a shows the detailed within-field LAI spatial variations achieved by the S2-10 m-GPR-LAI model. As shown in **Figure 7-6**d, higher LAI values (i.e., >4 m^2 m^{-2}) over circular irrigated fields had relatively low uncertainties, i.e., CV < 20%, while the surrounding regular (usually rainfed fields) had relatively high uncertainties, i.e., 20%>CV<40%. In contrast, the S2-20 m-RF-LAI results, i.e., F**igure 7-5**d, display relatively low within-field variability. In Harrismith, the S2-10 m-GPR-LAI model, i.e., **Figure 7-5**g, shows relatively low LAI values, while the S2-20 m-GPR-LAI model, i.e., **Figure 7-5**j, shows relatively high values for most fields. The S2-20 m-GPR-LAI model which achieved the best RMSE, i.e., **Figure 7-6**j, shows uncertainties similar to Bothaville, where higher LAI values (i.e., >4 m^2 m^{-2}) exhibited lower uncertainties, i.e., CV <20%, while LAI values of ~3 to 4 m^2 m^{-2} had a CV of between 20 and 40% (**Figure 7-6**j). These uncertainties were mainly owing to the presence of senescent (brown) leaves at the time of the field measurements, associated with the physiological maturity stage, while other fields were almost completely senescent. These fields may have had higher NDVI values than the threshold used to mask green vegetation, i.e., 0.2.

The spatial distribution of LC_{ab} between the two configurations was different, with S2-20 m showing higher values over irrigated (circular) fields (**Figure 7-5**e), while S2-10 m values over the same fields were relatively low (**Figure 7-5**b). The rainfed (regular) fields also exhibited relatively low LC_{ab} values. The Bothaville results using the S2-20 m configuration were achieved with GPR, while the S2-10 m results were obtained with XGBoost. Generally, the same patterns can be observed in Harrismith using both configurations and GPR. The uncertainty maps obtained with the best GPR models only, i.e., **Figure 7-6**e and **Figure 7-6**k, also show higher uncertainties (i.e., 20%>CV<40%) where LC_{ab} values are relatively low (<20 μg cm^{-2}), and better uncertainties (CV <20%) over irrigated fields with relatively high LC_{ab} values (> 40 μg cm^{-2}).

The spatial distribution maps of CCC obtained with XGBoost for both S2-10 m and S2-20 m configurations, in Bothaville, show no obvious differences (**Figure 7-5**c and f). In Harrismith, some differences between the two configurations are evident particularly over rainfed fields with relatively low CCC values (**Figure 7-4**i and **Figure 7-6**l), which can be two attributed to the fact that they were generated by two different algorithms and spectral configurations. The GPR uncertainties, which were applicable for Harrismith only (**Figure 7-6**l), show CV over 60% in some parts of the rainfed fields. While the spatial resolution was not of interest here, it may have played a role in the variations in spatial distributions of the retrieved biophysical and biochemical parameters. S2-10 m provided finer details with greater within-field variability than S2-20 m.

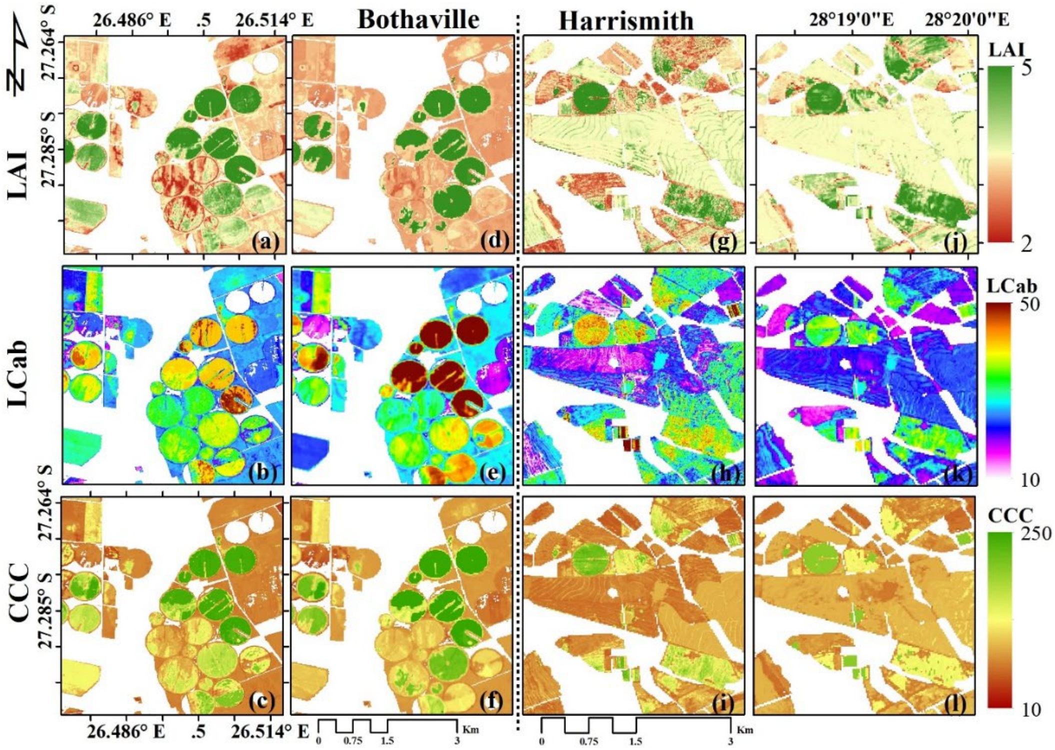


Figure 7-5. Maps generated by the best models with S2-10 m and S2-20 m data in Bothaville (a–f)) and Harrismith (g–l). (a) and (d) show the best LAI (m^2 m^{-2}) results using GPR (S2-10 m) and RF (S2-20 m), (b) and (e) show the best LC_{ab} (μg cm^{-2}) results using XGBoost (S2-10 m) and GPR (S2-20 m), and (c) and (f) show the best CCC (μg cm^{-2}) results using XGBoost for both S2-10 m and S2-20 m in Bothaville. (g) and (j) show the best LAI model using GPR for both S2-10 m and S2-20 m, (h) and (k) show the best LC_{ab} model results using GPR with both S2-10 m and S2-20 m, and (j) and (l) show the best CCC results using GPR (S2-10 m) and RF (S2-20 m), respectively.

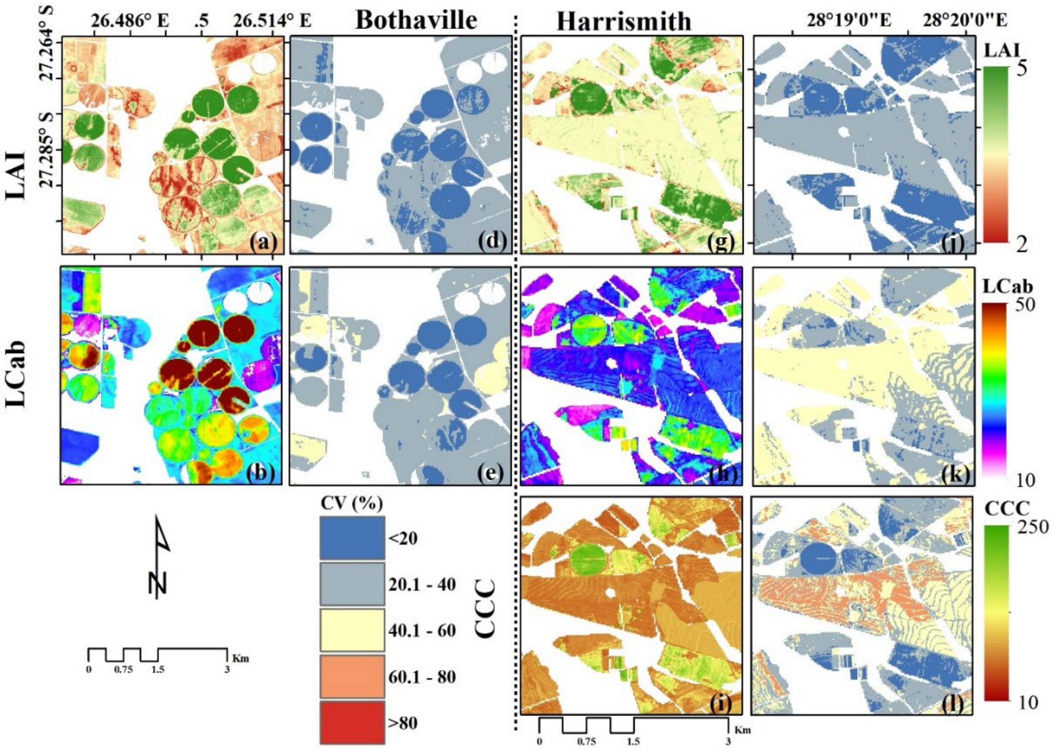


Figure 7-6. Biophysical and biochemical parameters maps generated by the best GPR models with S2-10 m and S2-20 m data in Bothaville (a–b) and Harrismith (e–g). (c, d, h–j) are the pixel-wise uncertainty coefficient of variation (CV, %).

4. Discussion

The advent of Copernicus Sentinel-2 twin-satellites has provided prospects to improve crop biophysical and biochemical retrieval accuracy as well as the frequency and level of detail relevant to Precision Agriculture and crop monitoring needs. Its improved spectral configuration, i.e., with new RE bands, centred at 705 nm, 740 nm, and 783 nm, has increased interest in their utility for crop biophysical and biochemical parameters retrieval using various methods. Of interest here, is the performance of the two Sentinel-2 spectral configurations, i.e., providing four standard multi-spectral bands in the VNIR region at 10 m spatial resolution (i.e., S2-10 m) and six bands in the RE-NIR-SWIR regions at 20 m spatial resolution (i.e., S2-20 m), in retrieving LAI, LC_{ab} and CCC using three robust MLRAs in two semi-arid agricultural sites, i.e., Bothaville and Harrismith. The two configurations were benchmarked against the full MSI spectral data (S2-All), consisting of ten bands, covering VIS, RE, NIR, and SWIR spectral regions all resampled to a spatial resolution of 10 m.

4.1. Performance of MSI configurations for crop biophysical and biochemical parameters retrieval

The VNIR spectral region of the electromagnetic spectrum (i.e., 350–649 nm), sampled by S2-10 m spectral bands, contains the fundamental vegetation absorption regions that have allowed image-based vegetation characterisation for decades (Tucker, 1979; Pinty & Verstraete, 1992; Myneni & Williams, 1994; Myneni et al., 2002; Brown et al., 2006; Zhu et al., 2013). Sentinel-2 B2:490 nm and B4:665 nm coincide with the widely known intense absorption by pigments such as xanthophyll, anthocyanin, chlorophyll, and carotenoid, while B8:560 nm exhibits a high scattering effect caused by the canopy structure, spongy mesophyll cells, and water content in leaves (Jensen, 1983; Blackburn, 1998). Therefore, it is not surprising that VNIR bands are one of the predominant predictors of biophysical and biochemical parameters in various retrieval approaches and environmental settings, despite the spectral resolution, range, and number of bands of the input dataset (Delegido et al., 2011; Verrelst et al., 2016). For example, Delegido et al. (2011) found that a normalised difference index (NDI) constructed with CHRIS (Compact High-Resolution Imaging Spectroscopy) hyperspectral bands centred at 674 nm (i.e., near S2-B4:665 nm) and 712 nm (i.e., near S2-B5:705 nm) was not only the best predictors of LAI but was also portable to a simulated Sentinel-2 image. This resulted in uncertainty (i.e., RMSE) of 0.6 $m^2 m^{-2}$. In the current study, the S2-10 m configuration—characterised by broad bandwidth bands (i.e., B2:490 nm, B3:560 nm, B4:665 nm and B8:842 nm)—resulted in comparable LAI uncertainty (i.e., $RMSE_{cv}$) of between 0.62 $m^2 m^{-2}$ and 0.53 $m^2 m^{-2}$ in Bothaville and Harrismith, respectively. Consistent with these findings, Verrelst et al. (2016) found that the bands centred at 462 nm (blue) and 1 327 nm (NIR) were among the four optimal bands (of the 125 HyMap spectral bands) for LAI retrieval at low uncertainties (i.e., RMSE*cv* of 0.37 $m^2 m^{-2}$ and R^2 of 0.95). However, as shown, their results were significantly better than the ones obtained here because they used hyperspectral data with narrow bandwidths (i.e., 11 nm and 21 nm) and two of the four optimal bands (i.e., centred at 708 nm and 723 nm) are positioned in the red-edge region.

Although the utility of hyperspectral data has been extensively demonstrated in the literature (Zhao et al., 2011; Yi et al., 2014; Yu et al., 2017), the lack of operational space-based sensors hinders its practical application. The narrow bands of Sentinel-2, i.e., B5:705 nm, B6:740 nm, B7:783 nm, and B8A:865 nm at 20 m spatial resolution are therefore a good compromise and essential for the detailed (field-level) characterisation of essential biophysical and biochemical parameters for agronomic applications. The contribution of the red-edge bands is shown in the results of the S2-20 m configuration (in this study)—characterised by the three red-edge bands, one narrow NIR band and two SWIR spectral bands (i.e., B11:1 610 nm and B12:2 190 nm)—which were robust in consistently retrieving LAI with relatively low uncertainties (RMSE*cv*) of 0.58 m^2 m^{-2} and 0.47 m^2 m^{-2} in Bothaville and Harrismith, respectively. The S2-20 m LAI uncertainties were slightly better than those obtained with S2-10 m at both sites. Therefore, the results show a combined effect of chlorophyll content, plant structure, and foliar moisture content. These control the reflectance in the RE, NIR, and SWIR regions and were strongly influential in the retrieval of LAI. The RE region is sensitive to changes in chlorophyll, thus averting the saturation effect caused by this pigment in the VIS region. For example, VNIR data often saturates and fails to accurately characterise medium (i.e., ~3 m^2 m^{-2}) to high (i.e., >5 m^2 m^{-2}) LAI values, while the inclusion of RE bands improves the dynamic range of these biophysical and biochemical parameters (Peng & Gitelson, 2011). A benchmark of S2-20 m's performance to S2-All showed proportionate performance, which implies that broadband VNIR can be discarded in the retrieval of LAI.

These results are comparable to Campos-Taberner et al. (2016) who found similar uncertainties over Mediterranean rice in Spain, i.e., RMSE: 0.39 m^2 m^{-2} and 0.51 m^2 m^{-2}, and Italy, i.e., 0.38 m^2 m^{-2} and 0.47 m^2 m^{-2}, using Landsat and SPOT-5 data, respectively. This also shows that the Sentinel-2 SWIR bands, which are similar to those of Landsat and SPOT data, were also essential in achieving low uncertainties with S2-20 m in this study. The contribution of SWIR bands (B11:1 610 nm and B12:2 190 nm) to LAI accuracy is mostly because they are affected by foliar moisture content, which plays an important role in the critical developmental (vegetative and productive) stages of crops. This in turn controls the abundance of biophysical and biochemical traits such as canopy structure and chlorophyll content (Curran, 2001; Verrelst et al., 2015). Essentially, the availability or deficiency of water determines the productivity and yield of an agricultural system. When crops reach physiological maturity (as in our case), moisture content declines steadily, thus causing a decline in leaf chlorophyll content and loss of greenness, while LAI may remain moderately high. In a related study utilising the entire Sentinel-2 spectral data (resampled to 20 m) in Bothaville (Kganyago et al. 2021), B11:1 610 nm and B12:2 190 nm were in the top five most influential bands in the LAI model, helping achieve a comparable (to the current study) RMSE of 0.5 $m^2 m^{-2}$ using RF algorithm.

Verrelst et al. (2015) also found that Sentinel-2 SWIR bands were consistently among the relevant spectral bands for retrieving LAI with the Variational Heteroscedastic GPR (VH-GPR) model with RMSE*cv* of 0.44 $m^2 m^{-2}$ and R^2 of 0.90. In the current study, the benchmarking results using S2-All further ascertained the relative contribution of RE, NIR, and SWIR bands, leading to the assumption that the location, bandwidth, and spectral regions where the bands were sampled (i.e., the spectral configuration) were more important than spatial resolution for LAI retrieval. This is consistent with Kganyago et al. (2020) who found no significant difference between Sentinel-2 resolutions in retrievals of LAI using a pre-trained hybrid Radiative Transfer Model (RTM) and Artificial Neural Networks (ANN) model.

The results also showed that LC_{ab} could be retrieved with relatively low uncertainties of 6.89 μg cm^{-2} and 7.02 μg cm^{-2} with S2-10 m at the two sites, respectively. This finding is consistent with Clevers et al. (2017) who found that vegetation indices constructed from VNIR (i.e., S2-10 m) spectral bands such as the Weighted Difference Vegetation Index (WDVI), Green Chlorophyll Index (CI_{green}), and Chlorophyll Vegetation Index (CVI) were more robust than those computed from red-edge (i.e., S2-20 m) spectral bands such as Red-edge Chlorophyll Index ($CI_{red\text{-}edge}$) (the ratio of Transformed Chlorophyll in Reflectance Index and Optimised Soil-adjusted Vegetation Index [TCARI/OSAVI]) in retrieving LAI, LC_{ab}, CCC for potato crops. Moreover, using GPR-BAT (GPR-based band analysis tool) on the field hyperspectral data, Verrelst et al. (2016) found that LC_{ab} could be accurately estimated with bands centred at 482 nm (blue), 500 nm and 564 nm (i.e., green peak), 710 nm and 714 nm (red-edge) and a region between 878–980 nm (NIR) with $NRMSE_{cv}$ < 10%. The red-edge spectral bands in Verrelst et al. (2016), i.e., 710 nm and 714 nm, are closer to Sentinel-2 B5:705 nm, which was found to be one of the most influential bands (alongside B3:560 nm, B4:665 nm, B11:1 610 nm, B12:2 190 nm), in the MLRA retrieval of LC_{ab}, achieving uncertainties of 7.57 µg cm^{-2} (Kganyago et al., 2021). In the current study, the contribution of these spectral bands (i.e., S2-20 m) in the retrieval of LC_{ab} was evident, achieving equivalently lower uncertainties as S2-10 m, i.e., 7.02 µg cm^{-2}, in Harrismith.

The benchmark results using S2-All did not result in any significant variations in estimates in relation to S2-10 m, demonstrating the usefulness of VNIR bands. The results imply that spectrally limited datasets such as those from SPOT 6/7, PlanetScope Doves and low-cost UAV platforms can be used for crop nitrogen management since studies have established that LC_{ab} is highly correlated with nitrogen content (Jia et al. 2013; Vincini et al. 2016). This also means that small-crop damage resulting from biotic (e.g., pests and diseases) and abiotic (e.g., water, temperature, and nutrients) stress factors can be detected early (i.e., before it becomes widespread) based on the high detail provided by these systems, thus potentially providing better prospects of early crop stress mitigation and higher yields. However, as shown by the results (**Table 7-2** and **Table 7-3**), S2-20 m and S2-All also provided equally good results; therefore, where RE, NIR, and SWIR spectral bands are available, they should be used to reduce systematic errors and improve the range of retrieved values in line with previous studies (Verrelst et al., 2012; Vincini et al., 2016).

For CCC, the best configurations were different for the two sites, where S2-20 m was better in Bothaville ($RMSE_{cv}$: 35.65 µg cm^{-2}) and S2-10 m was better in Harrismith (i.e., $RMSE_{cv}$: 26.84 µg cm^{-2}). The inconsistencies may be due to slightly different conditions at the two sites, where CCC (a product of LAI and LC_{ab}) in Harrismith was mainly influenced by chlorophyll content than the one in Bothaville, where plant structure and water content played a major role. This is reasonable since fieldwork dates for Harrismith and Bothaville differed slightly, i.e., March and April, respectively. Since the crop calendar is the same for both sites, Bothaville had relatively low LC_{ab}, and its influence on CCC was minimal when compared to LAI. Using S2-All did not improve the CCC results by S2-20 m (in Bothaville) and S2-10 m (in Harrismith), implying that either 10 m or 20 m MSI spectral bands can be applied without the need to use all resampled bands. The results of S2-20 m and S2-10 m achieved here are slightly better than those found in a related previous study (Kganyago et al., 2021), where CCC retrieval with resampled Sentinel-2 bands to 20 m achieved RMSE of 39.49 µg cm^{-2}.

The utility of S2-10 m and S2-20 m for various parameters is essential for the rapid assessment of the crop biophysical and biochemical parameters, without delays caused by additional pre-processing steps such as down-sampling the S2-10 m or up-sampling the S2-20 m spectral bands, and applying super-resolving techniques before retrieval (Zhang et al., 2019); thus, the results from this study have operational significance. In our study, up-sampling to 10 m caused 7 min 17.787 s delay for a single Sentinel-2 tile consisting of a width and height of 10980 pixels on an Intel® Core™ i7-8700 CPU and 64 GB RAM. Moreover, S2-10 m results obtained here are significant for informing biophysical and biochemical parameter retrieval using other sensors such as Planet Doves or low-cost UAVs, which only have VNIR bands and higher or flexible temporal resolution. However, all configurations had a relatively low R^2 in Harrismith, explaining the variability of between 54% to 72%, 53% to 57%, and 57% to 62%, for LAI, LC_{ab}, and CCC, respectively. In contrast, only LAI in Bothaville achieved a similar accuracy (R^2) of 52% to 58%, while the variability of LC_{ab} and CCC was relatively well-explained by the two configurations, with R^2 of 75% to 79% and 69% to 76%, respectively. The lower R^2 may be linked to the diverse structural forms within the same area emanating from different crop types and planting times. Nonetheless, all R^2 were above 50%, while NRMSE*cv* was below 20%, thus within limits recommended by the Global Climate Observing System (GCOS) (GCOS, 2011).

4.2. Effect of various MSI configurations on the performance of MLRAs

The above results were achieved with three state-of-art MLRAs, i.e., RF, XGBoost, and GPR. As shown by the results (**Figure 7-3** and **Figure 7-4**), the MLRAs considered here were generally sensitive to various Sentinel-2 MSI configurations, i.e., S2-10 m (i.e., four bands), S2-20 m (i.e., six bands), and S2-All (10 bands), outcompeting each other for each configuration and biophysical and biochemical parameter. Although RF and XGBoost had similar performances (attributed to their similar tree-based origin) XGBoost was superior in most cases in retrieving crop biophysical and biochemical parameters in both Bothaville and Harrismith and using all Sentinel-2 configurations. While RF uses bagging, randomly selected variables at each split, and many trees for predictions. In contrast, XGBoost introduces gradient-boosted decision trees and computational efficiency for thousands of trees, thus having better flexibility, improved efficiency to avoid overfitting, and is sparsity-aware (Chen & Guestrin, 2016).

The slightly better performance of XGBoost found here is consistent with previous studies (Bahrami et al., 2021; Zhang et al., 2021). Generally, tree-based algorithms are attractive because they are: simple to understand; transparent and explainable (i.e., tree structure, splitting points, and variables for each decision); and influential variables can be interrogated to understand how they operate in different scenarios. However, the results showed that GPR was more robust in most cases (**Figure 7-3** and **Figure 7-4**), resulting in better estimates even when only four bands (S2-10 m) were used. Despite its "black-box" nature, the strength of GPR lies in providing the per-pixel uncertainty estimates, which can be used to decide an uncertainty threshold in operational settings (Amin et al., 2021). In previous studies, GPR uncertainty measures, i.e., standard deviation and coefficient of variation, had been used to also exclude uncertainty from fallow and non-crop areas (Verrelst et al., 2013). Overall, MLRAs evaluated here showed sensitivity to different datasets (i.e., S2-10 m, S2-20 m, and S2-All) and the experimental sites (Bothaville and Harrismith). This implies that it is essential to evaluate various MLRAs, before choosing the optimal one for specific spectral configuration, application, and crop conditions.

Consequently, software tools such as ARTMO Machine Learning Regression Algorithm (MLRA) toolbox (Verrelst et al., 2012; Caicedo et al., 2014) (which provide an intuitive platform for rapidly and simultaneously computing multiple MLRAs) are essential to achieving improved crop biophysical and biochemical parameters and their rapid dissemination to users. Recent studies show the integration of ARTMO-generated coefficients with satellite data cloud APIs such as Google Earth Engine to enable rapid upscaling of crop biophysical and biochemical parameters such as LAI (Pipia et al., 2021; Estévez et al., 2022). Therefore, it would be interesting to extend the results obtained here, i.e., with different Sentinel-2 configurations, to other areas. In such a case, hybrid models (e.g., RTM-MLRA) should be considered, since experimental data are limited to the measured crop types and conditions and are affected by prevailing climatic and environmental conditions.

Although comparable with previous studies, our results for the S2-10 m configuration were likely impacted by the Sentinel-2 B2:490 nm, which is known to exhibit residual atmospheric effects that may have introduced uncertainties in the crop biophysical and biochemical retrievals using MLRAs. Another source of uncertainty may be the high correlation between the B2:490 nm and B4:665 nm, which may have introduced collinearity owing to their similar vegetation spectral response in these bands (i.e., pronounced absorption), as well as saturation resulting from chlorophyll absorption. Nonetheless, the usefulness of the blue band has been demonstrated in the Enhanced Vegetation Index (EVI) formulation to account for atmospheric effects and avoid the saturation effect of NDVI at high (i.e., 6 m^2 m^{-2}) and low (i.e., <2 m^2 m^{-2}) LAI values. Moreover, it featured prominently in the biophysical and biochemical parameters retrieval models in recent studies (Kganyago et al., 2021; Verrelst et al., 2016). Therefore, the sensitivity of the MLRA retrieval models to B2:490 nm effects must be evaluated in greater detail, in tandem with the efforts to quantify the magnitude of these residual errors from various atmospheric correction techniques (i.e., including Sen2Cor used here). Lastly, although there was a fair balance between maize (i.e., 63.94% and 62.01%) and beans (i.e., 32.56% and 49.72%) at the two sites respectively, peanuts (present in Bothaville only) were the least represented (3.49%). Besides the machine learning algorithms being renowned for robustness to imbalanced training samples, we cannot eliminate the possibility that this may have affected performance of the MLRA models. Crop-specific models will be considered in our future studies.

5. Conclusion

This study assessed the utility of the two Sentinel-2 spectral configurations that provide four standard multi-spectral bands in the VNIR region at 10 m spatial resolution (i.e., S2-10 m) and six bands in the RE-NIR-SWIR regions at 20 m spatial resolution (i.e., S2-20 m), in retrieving crop biophysical and biochemical parameters, i.e., LAI, LC_{ab} and CCC, using three robust MLRAs in two semi-arid agricultural sites, i.e., Bothaville and Harrismith. The results were compared to those obtained with all spectral bands (S2-All). In summary, the results showed that the S2-20 m configuration—with four narrow bands and two SWIR bands—was more robust, when compared to S2-10, in retrieving LAI with low uncertainties (i.e., $RMSE_{cv}$: 0.58 m^2 m^{-2} and 0.47 m^2 m^{-2}) in the two sites, respectively. In contrast, the S2-10 m configuration was relatively good at retrieving LC_{ab} in both sites ($RMSE_{cv}$: 6.89 µg cm^{-2} and 7.02 µg cm^{-2}). However, S2-20 m was equally robust in Harrismith, in achieving equivalent uncertainties as S2-10 m, i.e., $RMSE_{cv}$: 7.02 µg cm^{-2}. This shows the relevance of red-edge bands in biophysical and biochemical parameters retrieval as shown by previous studies (Mutanga & Skidmore, 2007; Verrelst et al., 2012). However, the results in the current study showed that VNIR bands could perform better than red-edge bands when it comes to retrieving LC_{ab}. Regarding CCC, the performance of the two configurations was not consistent in the two sites, where S2-20 m performed better in Bothaville with $RMSE_{cv}$: 35.65 µg cm^{-2}, but not in Harrismith, where S2-10 m yielded relatively low uncertainties with $RMSE_{cv}$ of 26.84 µg cm^{-2}. The obtained results are slightly better than those of a related study utilising resampled Sentinel-2 bands at 20 m (Kganyago et al., 2021). Moreover, all the configurations yielded accuracies that were slightly better or equivalent to the benchmark dataset consisting of resampled bands to 10 m, i.e., S2-All.

The better performance of S2-10 m in the retrieval of LC_{ab} and CCC found here may inform biophysical and biochemical parameters retrieval from similar high-resolution data with VNIR data from SPOT 6/7, PlanetScope Doves and low-cost UAV platforms, essential for field-level crop nitrogen management. However, it should be noted that the S2-10 m results obtained here, may have been affected by the inclusion of the blue band (i.e., B2), which contains residual atmospheric effects correlated to the red band (i.e., B4). This is owing to similar vegetation spectral response and saturation effects in the red band resulting from high chlorophyll content. Future studies should assess the sensitivity of the MLRA retrieval models to the blue band effects in greater detail.

The results imply that the two Sentinel-2 configurations can be used independently since there was no marked difference between all configurations (i.e., S2-10 m and S2-20 m) and the resampled bands (S2-All). Further analyses in other areas are required to ascertain the findings in the current study since the biophysical and biochemical retrieval models developed here used experimental data, which are limited to the measured crop types and conditions and affected by prevailing climatic and environmental conditions. While GPR was robust in most cases, RF and XGBoost were robust in others, indicating that all MLRAs evaluated here are sensitive to various spectral configurations and study areas. Therefore, it is essential to evaluate various MLRAs before choosing the optimal one for specific biophysical and biochemical parameters. The results inform future retrieval from the two Sentinel-2 configurations of essential crop biophysical and biochemical parameters that can support time-sensitive precision agronomic applications.

CHAPTER 8
EVALUATING THE PROSPECTS OF INCORPORATING PER-PIXEL VIEW AND ILLUMINATION GEOMETRY FOR IMPROVING THE ACCURACY OF RETRIEVING CROP BIOPHYSICAL AND BIOCHEMICAL PARAMETERS

This chapter is based on the following published manuscript:

Kganyago, M., G. Ovakoglou, Mhangara, P., Adjorlolo, C., Alexandridis, T., Laneve, G. & Beltran, J. S. (**2023**). Evaluating the Contribution of Sentinel-2 View and Illumination Geometry to the Accuracy of Retrieving Crop Biophysical and Biochemical Parameters. *GIScience and Remote Sensing*. https://doi.org/10.1080/15481603.2022.2163046

Abstract

Wide Field-of-View (FOV) sensors (such as Sentinel-2) exhibit per-pixel view and illumination geometry variation that may influence the retrieval accuracy of essential crop biophysical and biochemical variables (BVs) for Precision Agriculture. However, based on the existing literature, this aspect has scarcely been studied. Hence, the current study aimed to evaluate the contribution of view and illumination geometries to the accuracy of retrieving Leaf Chlorophyll *a* and *b* (LC_{ab}), Canopy Chlorophyll Content (CCC), and Leaf Area Index (LAI) using Random Forest (RF). The experiments were performed on various input variable scenarios where per-pixel geometric covariates, i.e., View and Sun Zenith Angles (VZA and SZA, respectively), and Relative Azimuth Angle (RAA), are excluded and then included in spectral bands (*SB*) and spectral vegetation indices (*SVIs*), respectively, in two semi-arid areas. The results showed that spectral bands or vegetation indices combined with geometric covariates improved the R^2 by 10%–15% for LAI and 3%–5% for CCC. In contrast, negligible improvements of 1%–2% were achieved for LC_{ab} with cross-validation test data and independent held-out dataset, respectively. In agreement with previous studies, VZA and SZA were among the most influential variables in the RF models for estimating LAI, LC_{ab}, and CCC. Collectively, per-pixel geometric variables explained more than 30% of the variability in surface reflectance for all Sentinel-2 spectral bands ($p < 2.2e\text{-}16$). Overall, the results showed that incorporating geometric covariates improved the accuracy of retrieving BVs; thus, it provided additional information that improves the predictive power of *SB* and *SVIs*. The significant benefits of the geometric variables were mainly realised for canopy-level BVs (i.e., LAI and CCC) than for LC_{ab}. Therefore, it is recommended to incorporate per-pixel view and illumination geometry in estimating LAI and CCC, especially when using wide-view sensors such as Sentinel-2. However, further testing in different phenology, sites, and acquisition conditions is needed to confirm the contribution of the geometric covariates to facilitate reliable retrieval of BVs from remotely sensed data that can improve agronomic decisions.

1. Introduction

Improving the estimation of crop biophysical and biochemical variables (BVs) is critical to support agricultural management practices, aid the optimisation of farm inputs, and monitor agricultural production at the landscape to regional level (Alexandridis et al., 2020). Plant parameters such as the Leaf Chlorophyll *a* and *b* (LC_{ab}), Leaf Area Index (LAI), and Canopy Chlorophyll Content (CCC) are often used as input to models monitoring the physiological status of plants and their health (Luet al., 2018). The LAI is considered an essential parameter for quantitatively analysing a series of biophysical processes which describe vegetation dynamics (Liu & Jin, 2017; Alexandridis et al., 2020). However, LC_{ab} and CCC are critical for characterising the photosynthetic activity of crops. Because they are both related to nitrogen (N) content (Clevers et al., 2017), CCC and LC_{ab} are essential agronomic parameters to monitor N content at canopy and leaf scales. Moreover, the photosynthetic activities of crops hinge on water availability. Thus, crop stress caused by declining soil moisture can be detected rapidly by assessing the changes in chlorophyll content.

LAI and LC_{ab} are traditionally measured using destructive and lab-based methods (Cihlar et al., 1997; Liang et al., 2011), which entail prohibitive operational costs for vast areal and temporal coverages. Therefore, optical handheld LAI instruments, such as Accu-Par Ceptometer, digital hemispherical photography, SunScan, and LAI-2200C and chlorophyll meters such as MC-100 Chlorophyll Concentration Meter and SPAD-502, provide reasonably accurate measurements of effective LAI and LC_{ab} in a rapid and non-invasive manner. When coupled with satellite-based remotely sensed data, the detailed retrieval of crop BVs can be realised over vast areas at frequent intervals, thus suitable for Precision Agriculture needs.

For years, spectral vegetation indices (*SVIs*) have shaped the scientific advancements in the retrieval of BVs based on their high correlation with vegetation traits such as leaf pigments, nitrogen, and LAI (Haboudane et al., 2008; Lee et al., 2008; Clevers & Gitelson, 2013; Clevers et al., 2017). Such SVIs were developed using visible and near-infrared (VNIR) bands from earlier missions such as AVHRR (Advanced Very High-Resolution Radiometer), MODerate Resolution

Imaging Spectroradiometer (MODIS), and Landsat. Hence, SVIs such as the NDVI (Rouse Jr, 1973) and its variants—designed to overcome background effects, atmospheric interference, and saturation—are popular in agronomic and other vegetation applications. However, VNIR indices are known to suffer from underestimations at various (i.e., low or high) biomass, chlorophyll content, and LAI. Recently, red-edge SVIs based on quasi-hyperspectral sensors, such as Worldview-2 and Multi-Spectral Instrument (MSI) on board Sentinel-2, were shown to enhance the accuracy of BVs and overcome challenges of VIS/NIR indices (Fitzgerald et al., 2010; Delegido et al., 2013; Dimitrov et al., 2019; Sibanda et al., 2019). Other studies have demonstrated the potential of Radiative Transfer Models (RTMs) (Jacquemoud et al. 2009). However, such models are beset with errors resulting from poor parameterisation and complexity, which may result in considerable uncertainty in the retrieved BVs (Durbha et al., 2007; M. Xu et al., 2019).

Alternatively, MLRAs (Machine Learning Regression Algorithms) were tailored to learn complex relationships between measured BVs and reflectance data from satellite imagery (Mao et al., 2019). The application of MLRAs for the retrieval of important crop BVs is considered ideal based on their high accuracy and flexibility to process multiple features and datasets (Apolo-Apolo et al., 2020). It is, therefore, essential to examine whether such approaches could improve future product development, aid informed agricultural management decisions and accelerate Precision Agriculture, especially in developing regions such as Africa where operational use of Earth observation for Precision Agriculture is still in its infancy.

Vegetation canopy reflectance is affected by: biophysical and biochemical traits; phenological and environmental changes; signal attenuation by the atmosphere; topographic effects; and acquisition conditions, among others (Ollinger, 2011). In particular, satellite-based radiometric measurements exhibit non-negligible effects from view and illumination geometry or Bidirectional Reflectance Distribution Function (BRDF) that may introduce ambiguity in retrieving BVs. BRDF refers to the ratio of the solar incident irradiance E_i from a given direction on a target surface to its contribution to the reflected radiance L_r

from the same surface in another direction. Mathematically, it is expressed as $f_r(\theta_i, \phi_i; \theta_r, \phi_r) = \frac{dL_r(\theta_i,\phi_i;\theta_r,\phi_r;E_i)}{dE_i(\theta_i,\phi_i)}$ $[sr^{-1}]$, where θ and ϕ indicate a direction, i and r are the quantities associated with the incident and reflected radiant flux, respectively. E_i, L_r, and d denote the incident irradiance, reflected radiance, and differential quantity, respectively (Nicodemus et al., 1977). Essentially, the BRDF refers to the variation in surface reflectance with changing view and illumination geometry (Nagol et al., 2015). Owing to the different acquisition angles between sensors and orbits and the sun-angle variations across different seasons, latitudes, and terrains, previous studies dedicated considerable effort to reducing BRDF effects (Roy et al., 2017). The reduction of BRDF effects is essential for obtaining temporally consistent data across different sensors to enhance the accuracy and temporal consistency of land cover maps (Nagol et al., 2015), canopy structural parameters (Sharma et al., 2013), and BVs (Pocewicz et al., 2007).

In that regard, techniques such as the c-factor (Roy et al., 2017) (based on a semi-empirical Ross Thick/Li Sparse-Reciprocal BRDF model) (Schaaf et al., 2002) and integrated topographic and angular normalisation (Yin et al., 2020) (based on the C-factor and Path Length Correction [PLC]) have been proposed for Landsat and Sentinel-2. Other techniques, such as the MODIS BRDF/Albedo algorithm, have for decades been used to produce Nadir BRDF-Adjusted surface Reflectance (NBAR) (Schaaf et al., 2002). For high-resolution sensors, e.g., Sentinel-2, atmospheric correction tools such as Sen2Cor provide an option to perform empirical BRDF corrections (Louis et al., 2016). However, to be consistent with other methods, the BRDF correction option is often left deactivated in validation studies (Sola et al., 2018). So far, the radiometric variation introduced by the sensor view and illumination geometry has merely been regarded as a nuisance. However, such geometry-induced spectral variation may provide additional structural information content owing to the dependence of reflected radiance on the geometric configurations (Kneubühler et al., 2008). This is especially highlighted in the studies utilising multi-angular data (Verrelst et al., 2008, Duveiller et al., 2015, Vierling et al., 1997), showing its utility for characterising structural variation of various layers of vegetation (e.g., foliage clumping) that affects LAI. However, high-resolution multi-angular data are costly and can only be acquired by commercial satellites (Roosjen et al., 2018). Other studies showed that view and illumination geometry have a considerable effect on vegetation indices and phenology metrics (Ranson, 1986, Middleton, 1991, Pocewicz et al., 2007, Verrelst et al., 2008, Ma et al., 2020).

Meanwhile, studies (Nagol et al., 2015, Roy et al., 2017) also show that popular multi-spectral sensors (e.g., Landsat and Sentinel-2) exhibit within-scene angular effects. In particular, Sentinel-2's wide Field-of-View (FOV), i.e., ~20.6° (~295 km), have been shown to exhibit considerable within-scene surface anisotropy because of varying per-pixel view and illumination geometry for each of its 12 detectors (Gascon et al., 2017, Roy et al., 2017). Fortunately, Sentinel-2 data provides a pixel-wise and band-wise view and illumination geometry, thus providing some prospects to address the questions such as: (1) how much variability in crop canopy reflectance is explained by the view and illumination geometry variations? (2) What is the contribution of incorporating per-pixel view and illumination geometry to the accuracy of retrieving crop BVs using MLRAs? (3) Are geometric variables influential in improving modelling accuracy when coupled with commonly used radiometric input variables, i.e., spectral bands and SVIs? These will clarify whether acquisition conditions (i.e., sun-view geometry) provide additional information or noise to the estimation of crop BVs. The operational Sentinel-2 Land biophysical processor (SL2P) (available in Sentinel Application Platform, SNAP), based on pre-trained ANN, uses per-pixel geometric information as input to the biophysical retrieval process (Weiss & Baret, 2020). However, such complex "black box" (or opaque) models do not provide model agnostics, which would otherwise reveal the influence of per-pixel geometric variables on the estimations. Interpretable MLRAs such as RF are essential for obtaining new causal links and novel insights between predictor and response variables. Moreover, they enable the detection and diagnosis of biases in the MLRA models and associated input data, hence essential for improved satellite-based value-added products (Azodi et al., 2020).

In that context, this study evaluated the contribution of view and illumination geometries on crop biophysical and biochemical retrieval using the Random Forest (RF) regression algorithm and various combinations of Sentinel-2 covariates. The evaluated covariates consisted of spectral bands (*SB*), *SVIs*, *SB* combined with sensor view-illumination geometries (*SB + Angles*), and *SVIs* combined with sensor view and illumination geometries (*SVIs + Angles*), and all variables (*Avar*). The specific objectives were three-fold: (1) To quantify the amount of spectral variability caused by the per-pixel view and illumination geometry in Sentinel-2 spectral bands, (2) To determine the contribution of per-pixel view and illumination geometry to the accuracy of retrieving LAI, LC_{ab} and CCC with RF, and (3) To identify the most important (i.e., influential) variables contributing to the model accuracy, considering different combinations of covariates. The study area comprised of two semi-arid African agricultural sites, located in South Africa.

2. Study Area

Bothaville and Harrismith sites are located in the Free State province, South Africa (**Figure 8-1**). The sites are characterised mainly by commercial farming of crops and livestock. The summer growing season, which constitutes most cropping activities in both sites, has mean annual temperatures of about 18 °C and 19.2 °C and mean annual rainfall of about 584 mm and 115 mm, respectively (Kganyago, Mhangara, et al. 2020, Kganyago, Ovakoglou, et al. 2020). The crops at the sites are similar, i.e., predominantly maize and beans, while sunflower and groundnuts are dominant only in Bothaville. Cropping takes place on flat slopes with sandy-loamy and sandy soils in Bothaville and undulating slopes with clay-loamy soils in Harrismith. The summer cropping calendar at both sites is from December to May or June.

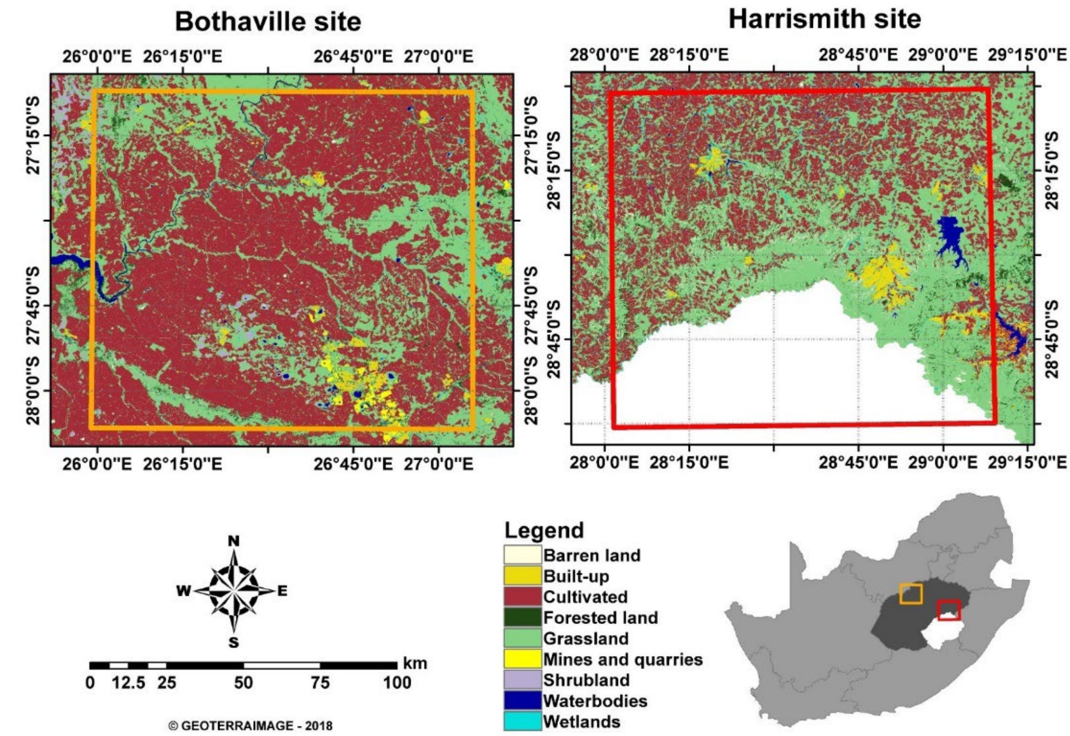


Figure 8-1. Land cover distributions in Bothaville and Harrismith (orange and red squares), Free State province (dark grey), South Africa (adopted from Kganyago, 2021). The insert maps are projected to UTM (Universal Transverse Mercator) with WGS-84 (World geodetic System 1984) Datum. The land cover dataset was obtained from the Department of Forestry, Fisheries and the Environment (https://egis.environment.gov.za/gis_data_downloads, accessed 20 March 2021)

3. Materials and Methods

The summary of the methods followed in this study is given in **Figure 8-2.**

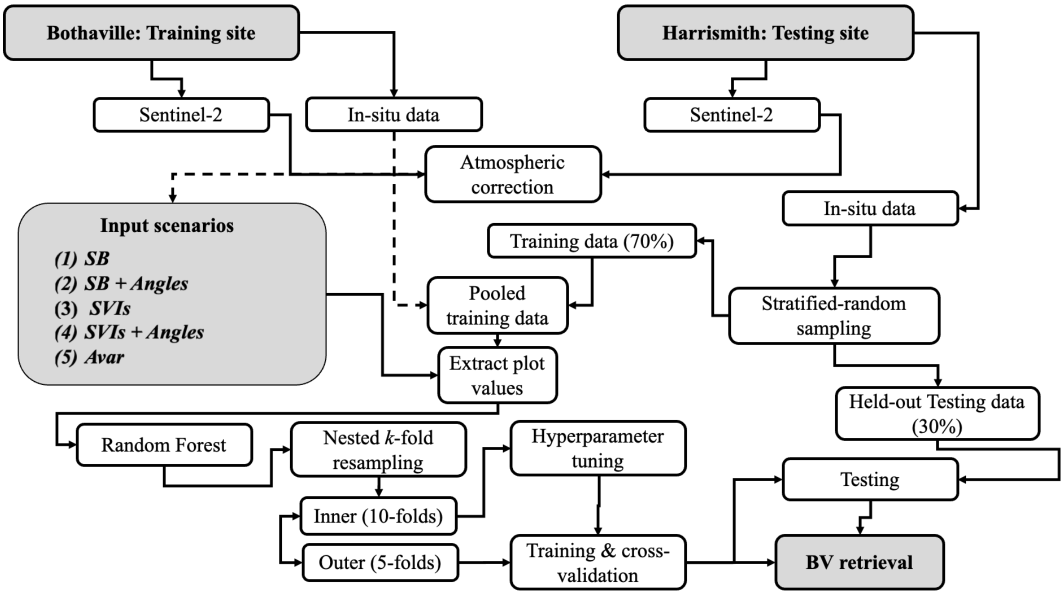


Figure 8-2. The schematic representation of the methods used in this study. Five experimental scenarios, i.e., spectral bands (*SB*), SB and Angles (*SB + Angles*), Spectral vegetation indices (*SVIs*), *SVIs* and *Angles* (*SVIs + Angles*) and All variables (Avar) were used for crop biophysical and biochemical variable (BV) retrieval using two test datasets, i.e., cross-validation test dataset and an independent held-out test (i.e., unseen) dataset.

3.1. Data

3.1.1. In situ data

LAI and LC_{ab} in situ data were collected non-destructively in Harrismith between 15 and 26 March 2021 and in Bothaville between 11 and 23 April 2021, respectively. A similar sampling strategy was followed at the two sites, where the plots (each with 40 m × 40 m) were designed along transects chosen randomly within various crop fields. Six to eight random measurements were captured within each plot and averaged. Each plot's centroids were geo-tagged with a GPS coordinate and a picture using TDC600 handheld Data Collector (Trimble® Inc., Westminster, CO, USA) with 1.5 m GNSS accuracy. Plant Canopy Analyzer 2200c (LI-COR, Inc., Lincoln, NE, USA), equipped with an 180° FOV view cap, was used to measure LAI. The view cap ensured that any possible uncertainty caused by the operator and uneven sky conditions were avoided. In contrast, MC-100 (Apogee Instruments, Inc., Logan, UT, USA) was used to acquire LC_{ab} measurements on sun-exposed leaves. The CCC was obtained as a product of plot-averaged LC_{ab} and LAI measurements (LC_{ab} × LAI) (Jacquemoud et al., 2009). Because our aim was to develop generic models, we combined the in situ data from each site, i.e., beans, maize, and peanuts, and beans and maize found in Bothaville and Harrismith, respectively. Our approach is aligned with Nguy-Robertson et al., (2012) who note that crop types with different physiological pathways (i.e., C3 and C4), plant structures and architectures are representative of the physiological and anatomical traits of many other crop types. Therefore, the developed models are likely widely applicable to multiple crops, critical for sub-Saharan Africa, where intercropping agricultural practices dominate.

3.1.2. Sentinel-2 data

Sentinel-2 MSI images (granule: 35JMK and 35JPJ), acquired on 2021-03-22T07:46:11.024Z and 2021-04-14T07:56:01.024Z, over Harrismith (Sensing orbit number: 135) and Bothaville (Sensing orbit number: 35), respectively, were retrieved from Sentinel Hub (Sinergise Laboratory for geographical information systems, Ltd., Ljubljana, Slovenia) at processing Level-2A, i.e., surface reflectance. Sentinel-2A and 2B have the identical MSI sensors and provide a 5-day revisit period. The images used here were acquired by Sentinel-2A satellite in a descending orbit direction. Out of the 13 MSI spectral bands, which are provided in various spatial resolutions, we used only the 10 m and 20 m spectral bands which are designed for land applications: 490 nm (B2:Blue), 560 nm (B3:Green), 665 nm (B4:Red), 705 nm (B5:Red-edge), 740 nm (B6:Red-edge), 783 nm (B7:Red-edge) , 865 nm (B8A:NIR), 842 nm (B8:NIR), 1 610 nm (B11:SWIR-1), and 2 190 nm (B11:SWIR-2). The details of the Sentinel-2 calibration are reported by Gascon et al. (2017). In addition to radiometric data, each tile, i.e., 100 km × 100 km, is packaged within various folders containing the ancillary, spectral data, and band-wise and mean per-pixel view and illumination geometry, i.e., with: View Zenith Angle (VZA) and Sun Zenith Angle (SZA), respectively; and View Azimuth Angle (VAA) and Sun Azimuth Angle (SAA), respectively, on a 5 km grid (Gascon et al., 2017). The mean SZA, mean VZA, mean SAA, mean VAA, and spectral bands acquired by different detectors were resampled to a standard spatial resolution. This was achieved with a 20 m resolution using the bilinear resampling technique.

Other land covers, i.e., other than croplands, were removed from the images by applying a crop mask generated from the National Crop Boundaries data

(CropEstimatesConsortium, 2017). To limit the analyses to the active croplands of 2021 summer growing season, we applied an NDVI-based vegetation mask, where non-vegetated pixels were considered as those with NDVI < 0.2.

3.2. Multiple linear regression

As shown in **Figure 8-3**, the SZA, VZA, and RAA vary spatially within a single Sentinel-2 MSI tile. Gascon et al. (2017) indicate that Sentinel-2 acquisition geometries differ according to the 12 detectors for each spectral band owing to the sensor's push-broom design. This spatial variability is clear in **Figure 8-3** and **Figure 8-4**, where it is evident that the SZA, VZA, and RAA vary by pixel across the scene. Therefore, it is also expected that pixel-wise surface reflectance over crops would vary by sensor view and illumination geometry (Verrelst et al., 2008, Jay et al., 2017, Galvao et al., 2011).

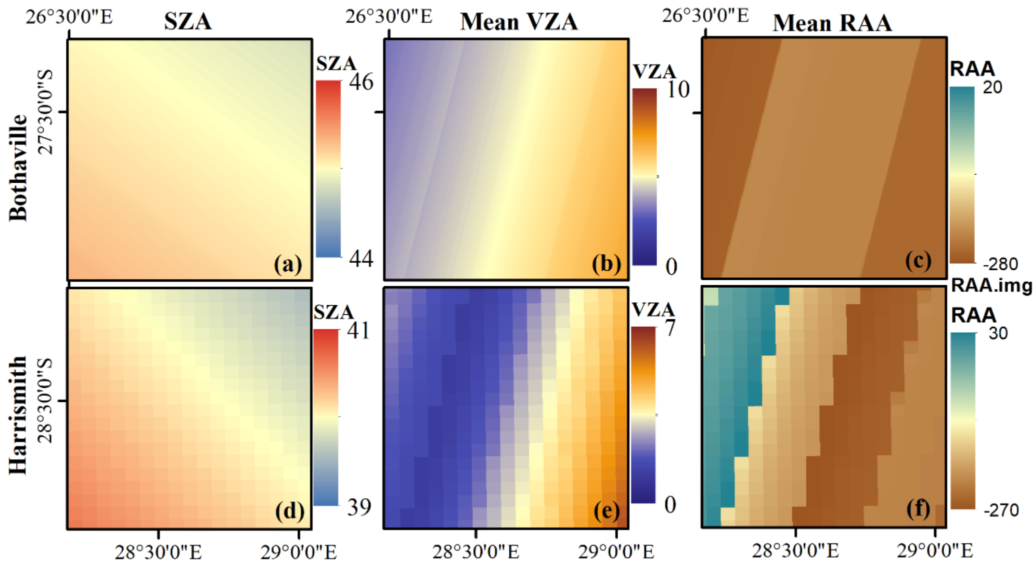


Figure 8-3. Spatial (per-pixel) variation of Sentinel-2 illumination and view angles over the two study sites, i.e., Bothaville (a–c) and Harrismith (d–f). (a) and (d) show per-pixel Sun Zenith Angle (SZA), (b) and (e) show per-pixel mean View Zenith Angle (VZA), and (c) and (f) show per-pixel mean Relative Azimuth Angle (RAA).

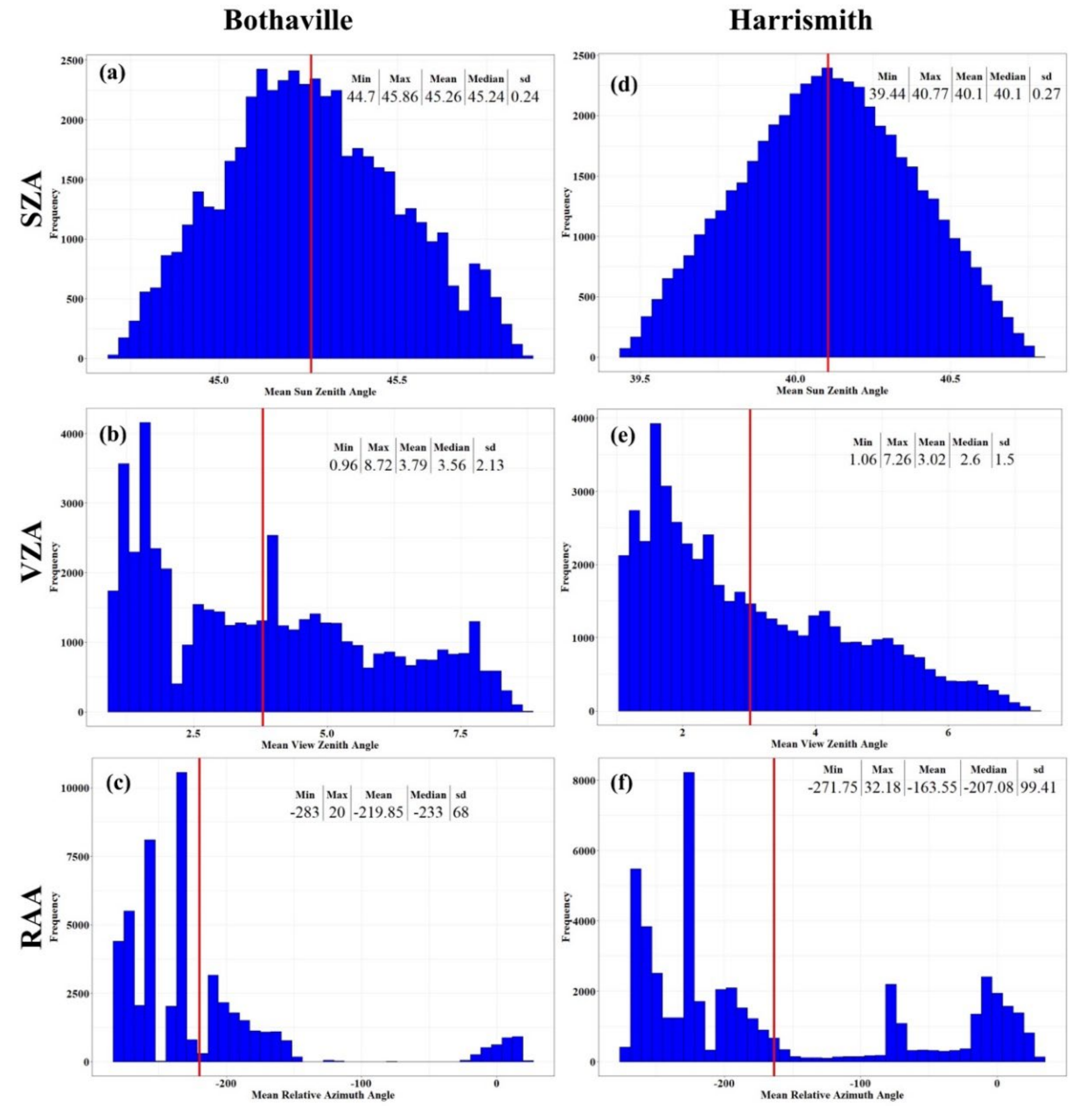


Figure 8-4. Summary statistics Sentinel-2 illumination and view angles in Bothaville (a–c) and Harrismith (d–f). The red line indicates the mean value for each geometric variable over 50 000 randomly sampled pixels.

In this study, multiple linear regression (Equation 8-1) and coefficient of determination adjusted for the degree of freedom, adjusted R^2, were used to evaluate how much of the reflectance variation is explained by sensor view and illumination per-pixel variation using mean SZA, VZA, and RAA as predictor variables (x), and each spectral band as the response variable (y). RAA was calculated as a difference between SAA and VAA.

$$y = m_1x_1 + m_2x_2 + \cdots + m_nx_n + b \quad (8\text{-}1)$$

Where y is the response variable (i.e., surface reflectance of each spectral band), and $x_{1\ldots n}$ are predictor variables (SZA, VZA, RAA), m_1 represent regression coefficients, while b is the intercept. The regression coefficient, i.e., the partial derivative of the response variable for each predictor variable, measures the linear sensitivity of y to inputs $x_{i \ldots n}$ (Mohanty & Codell, 2002). Adjusted R^2 was used to assess the relationship between the response variable and predictor variables. In contrast, the F-statistic *p-value* was used to determine the significance of the relationships at $\alpha = 0.05$.

3.3. Statistical analysis experiments

3.3.1. Random Forest Regression Algorithm

The retrieval of BVs considered here, i.e., LAI, LC_{ab}, and CCC, was performed using RF (Breiman, 2001) within the "randomForest" R-statistics package (Tierney, 2012; Liaw & Wiener, 2002). RF is a popular tree-patterned MLRA that utilises bagging to repeatedly and independently build multiple decision trees using a subset of random training samples generated using resampling with replacement out of the original sample (Fawagreh et al., 2014, Breiman, 2001). From all the training data (i.e., 70%), about 64% are kept for model building (i.e., in-bag samples), while model performance evaluation and variable importance are performed with the remaining 36% (i.e., out-of-bag or OOB samples) (Gislason et al., 2006). For each predictor, variable importance is computed by assessing the Percent Increase in Mean Squared Error (%IncMSE) when permuting OOB samples while keeping all other variables constant. The predictive power of the variables is assessed with %IncMSE, i.e., also used as a ranking measure for the various variables. For example, a variable is deemed significant if its exclusion from the model causes a considerable reduction in prediction accuracy. Essentially, a predictor with a high importance score strongly correlates with the response variable or predictive power (Mutanga et al., 2012). The tuning parameters, i.e., *ntree* (number of trees) and *mtry* (the number of variables for each split), were determined using the Grid-search approach using values from 100 to 500 with an interval of 10, and from 1 to p (i.e., number of predictors) with a single interval, respectively.

3.3.2. Input variables

The input variables to the RF regression algorithm were divided into five (5) experimental scenarios (**Table 8-1**). The per-pixel view and illumination geometries, i.e., SZA, mean of VZA and RAA (i.e., calculated as the difference between SAA and VAA), were obtained as raster layers from the Sentinel-2 MSI product. At the same time, the MSI spectral bands were used to compute carefully selected red-edge *SVIs*, see **Table 8-2**) based on their previous empirical relationship with the BVs considered in the current study (Viña et al., 2011; Schlemmer et al., 2013; Pasqualotto et al., 2019a; Delegido et al., 2011; Nguy-Robertson et al., 2012). All SVIs were calculated in the Sentinel-2 Toolbox within *Sentinel Application Platform (*SNAP) *v8.0* software (http://step.esa.int, accessed 10 November 2020). Since the variables had varying values, they were normalised using scale and centre approaches in the "mlr" R-package.

Table 8-1. The various experimental scenarios were designed by excluding and including view and illumination geometry from commonly used radiometric inputs (i.e., spectral bands and vegetation indices).

Scenario	Input variables	Description of input variables
SB (*p* = 10)	Spectral Bands	Only spectral bands (defined in 3.1.2) were used.
SB + *Angles* (*p* = 13)	Spectral Bands and Angles	Spectral bands combined with the cosine of view/illumination geometries (*Angles*), i.e., Sun Zenith Angle (SZA), mean View Azimuth Angle (VAA), and the Relative Azimuth Angle (RAA).
SVIs (*p* = 10)	Spectral Vegetation Indices	Only the spectral vegetation indices were used.
SVIs + *Angles* (*p* = 13)	SVIs and Angles	*SVIs* combined with the *Angles*, i.e., SZA, VAA, and mean RAA, were used.
Avar (*p* = 23)	All Variables	All input variables (described above) were used.

Table 8-2. The red-edge indices from Sentinel-2 MSI data.

Index	Equation	Reference
Normalised Difference Index	$NDI45 = \frac{R_{705} - R_{665}}{R_{705} + R_{665}}$	Delegido et al. (2011)
MERIS Terrestrial Chlorophyll Index	$MTCI = \frac{R_{740} - R_{705}}{R_{740} - R_{665}}$	Dash and Curran (2004)
Modified Chlorophyll Absorption Ratio Index	$MCARI = [(R_{705} - R_{665}) - 0.2 \times (R_{705} - R_{560})] \times (\frac{R_{705}}{R_{665}})$	Daughtry et al.(2000)
Red-Edge Inflection Point Index	$REIP = 700 + 40 \times (\frac{(R_{665} - R_{783})/2 - R_{705})}{(R_{740} + R_{705})}$	Guyot & Baret (1988)
Sentinel-2 Red-Edge Position Index	$S2REIP = 705 + 35 \times (\frac{((R_{665} - R_{783})/2 - R_{705})}{(R_{740} + R_{705})}$	(Frampton et al., 2013)
Inverted Red-Edge Chlorophyll Index	$IRECI = \frac{R_{783} - R_{665}}{R_{705} + R_{740}}$	(Frampton et al., 2013)
Red-edge Chlorophyll Index ($CI_{red\text{-}edge}$)	$CI_{red-edge} = \frac{R_{NIR}}{R_{RE}} - 1$	(Gitelson et al., 2003)
Red-Edge Relative Indices (RERI) @ 705 nm	$RERI_{[705]} = \frac{R_{705} - R_{665}}{R_{865}}$	Xu et al. (2019)
Red-Edge Relative Indices (RERI) @ 783 nm	$RERI_{[783]} = \frac{R_{783} - R_{705}}{R_{865}}$	Xu et al. (2019)
Simple Sentinel-2 LAI Index	$SeLI = \frac{R_{865} - R_{705}}{R_{865} + R_{705}}$	Pasqualotto et al. (2019b)

3.3.3. Model calibration and testing

The in situ data from Harrismith data were split into 70% calibration (i.e., training) *vs* 30% testing (i.e., validation) using stratified-random sampling to represent all crops at the study site proportionally. Then, the 70% calibration data were combined with Bothaville data to form the calibration data for the study, while 30% was held-out for independent testing of the models. Using the pooled calibration data from the two sites, nested *k*-fold cross-validation (*cv*), where two nested *k*-fold *cv* resampling, were used for hyperparameter tuning ($k = 10$) and model calibration and testing ($k = 5$), respectively. *k*-fold cross-validation (*cv*) randomly divides the dataset into equal number of sub-datasets according to *k*. Then, *k*-1 sub-datasets constitute a training dataset, while one k^{th} sub-dataset is the validation dataset. Using one of the *k* sub-datasets at each iteration, i.e., *k* times, the final estimation value is formed by aggregating all the validation steps. The nested *k*-fold *cv* resampling ensured that all pooled data from both study sites were used to calibrate and test the RF models (Shah et al., 2019;Verrelst et al., 2015). The *cv* resampling for model calibration and testing was performed with 100 iterations to reduce variability and increase the model's predictive ability on unseen data.

3.3.4. Prediction accuracy metrics

The prediction accuracies for each Random Forest (RF) model, constructed using various experimental scenarios, were assessed with the recommended metrics by Richter et al. (2012). These metrics are R^2 (coefficient of determination), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Relative RMSE (RRMSE) (Equation 8-2 to 8-6). The R^2 calculates the amount of the response variable (e.g., CCC) variance explained by the predictors (e.g., spectral bands). The RMSE and MAE, on the other hand, measure the degree of error between predictions and observations in the units of the BVs, i.e., $m^2\ m^{-2}$ for LAI and $\mu g\ cm^{-2}$ for LC_{ab} and CCC. The RRMSE is used for comparing different variables or ranges. RRMSE values of ≤ 10% are indicative of Excellent predictions, and 10% < RRMSE ≤ 20% are Good predictions (Richter et al., 2012). Lastly, percentage Bias (%Bias) measures the model's tendency to overestimate or underestimate a specific BV. A model is deemed accurate when the %Bias is close to 0% (Gara et al., 2019). Prediction accuracy assessment was assessed using R-statistics version 4.1.2 (Tierney, 2012) packages "Metrics" (https://cran.r-project.org/web/packages/Metrics/index.html, accessed 18 July 2021) and "Fgmutils" (https://rdrr.io/cran/Fgmutils/, accessed 18 July 2021).

$$R^2 = \frac{\sum(y_i^n - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (8\text{-}2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - y_i)^2} \quad (8\text{-}3)$$

$$RRMSE = \frac{RMSE}{\bar{x}_i} \quad (8\text{-}4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (8\text{-}5)$$

$$\%BIAS = \sum_{i=1}^{n}(x_i - y_i)/\sum_{i=1}^{n}(x_i) \quad (8\text{-}6)$$

Where x_i denotes the observed BV (e.g., CCC), and y_i denotes the predicted BV (e.g., CCC), $\bar{x}_i$ and $\bar{y}_i$ are the means of the observed (or measured) and predicted BV, respectively. *N* in Equation 8-3 is the number of errors and *n* is the sample size.

4. Results

4.1. Spectral variability explained by the view and illumination geometry

To determine the amount of spectral variability explained by the view and illumination geometry variables, we assessed the pixel-wise relationship of the geometries and each spectral band using multiple linear regression and coefficient of variation adjusted for the degree of freedom (adjusted R^2). The results (**Table 8-3**) show that there is a significant relationship between view and illumination geometric covariates and Sentinel-2 spectral bands (*p-value* < 2.2e-16). The adjusted R^2 values were between 0.30 and 0.43; therefore, view and illumination geometry covariates (i.e., SZA, VZA, RAA) explained 30% to 43% of spectral variability. B4, B2, and B7 had the lowest adjusted R^2 values of 0.30≤0.34, while B3, B5, B6, B8, B8A and B12 had adjusted R^2 values of 0.35<0.4, and B11 had adjusted R^2 value of >0.4. The same is observed in Harrismith for VIS bands B2, B3, and B4 and NIR band B8, but the geometric variables explained <30% of surface reflectance variability in B5, B6, B7 and B8A, while they explained >40% of surface reflectance variability in B11 and B12. In Harrismith, the significant geometric variables at $\alpha < 0.05$ were RAA in B2, RAA and VZA in B3, and SZA and VZA in B7, while all the other bands were influenced significantly by all geometric variables, i.e., SZA VZA, and RAA. Similarly, in Bothaville, all the geometric variables were significant at $\alpha < 0.05$.

Table 8-3. The amount of spectral variability explained by the geometric variables based on multiple linear regression. Adj. R^2 denotes Adjusted R^2.

			B2	B3	B4	B5	B6	B7	B8	B8A	B11	B12
B	(	Adj. R^2	0.32	0.36	0.30	0.35	0.35	0.34	0.35	0.36	0.43	0.37

	p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
Harrismith	Adj. R^2	0.35	0.32	0.37	0.25	0.19	0.21	0.28	0.36	0.47	0.42
	p-value	< 2.2e-16	2.709e-15	< 2.2e-16	9.297e-12	1.076e-08	1.624e-09	2.718e-13	< 2.2e-16	< 2.2e-16	< 2.2e-16

4.2. Predictive performance of Sentinel-2 covariates

Five experimental scenarios, designed based on the different combinations of Sentinel-2 MSI covariates, were assessed for their performance in retrieving LAI, LC_{ab}, and CCC using RF regression algorithm. The results were confirmed by applying the calibrated models on two test datasets, i.e., one selected through 5-fold cross-validation (*cv*) using the pooled data from two study sites (hereafter, cross-validation test dataset) and an independent held-out (i.e., unseen) dataset from Harrismith (hereafter, independent held-out test dataset). Generally, the results (**Table 8-4**) indicate that, among the compared experimental scenarios, the *SB + Angles* and *SVIs + Angles* scenarios provided the most robust retrievals of all the BVs considered here, except for CCC, where *SB + Angles* and *Avar* provided better results than other scenarios. The LAI retrieval models using *SB + Angles* and *SVIs + Angles* yielded excellent predictive performance with RMSE of 0.50 and 0.43 m^2 m^{-2}. Moreover, they explained the most significant variability of LAI within the two test datasets, i.e., 69% and 67% for cross-validation test data and independent held-out (i.e., unseen) data, respectively. These performances were closer to *Avar*, which achieved RMSEs of 0.52 and 0.55 m^2 m^{-2} and R^2 of 0.66 and 0.67, respectively. In contract, the predictive performance of input scenarios where the per-pixel view and illumination geometry covariates were excluded, i.e., *SB* and *SVIs*, were relatively inferior, achieving RMSE > 0.60 m^2 m^{-2} and R^2 < 0.60) for both test datasets.

Interestingly, the best predictive performance for LC_{ab} was achieved with *SVIs + Angles* with RMSE of 6.57 µg cm^{-2} (R^2: 0.76) and 6.16 µg cm^{-2} (R^2: 0.61), respectively. The next better predictive performance was achieved by *SVIs* and *Avar*, while *SB + Angles* and *SB* were slightly worse. Similar to LAI models, the

CCC retrieval models built with input scenarios that incorporated per-pixel view and illumination geometry covariates, i.e., *SB + Angles, SVIs + Angles*, and *Avar*, were better than those that used scenarios which did not contain any geometric variables, i.e., *SB* and *SVIs*.

Overall, the predictive performance of the BV models was consistent between the two test datasets, demonstrating the contribution of per-pixel view and illumination geometry covariates in improving the robustness of BV retrievals. However, contrary to LAI results, the *SVIs*-LC_{ab} model outperformed the scenarios incorporating per-pixel view and illumination geometry covariates, i.e., *SB + Angles* and *Avar*. As demonstrated by *SB + Angles*, *SVIs + Angles*, and *Avar* results, incorporating pixel-wise geometric information led to robust LAI, LC_{ab}, and CCC predictions using RF. It can also be deduced that carefully chosen red-edge vegetation indices have high information content considering *SVIs*' next best performance after *SVIs + Angles* in the retrieval of LC_{ab}. In contrast, spectral reflectance bands were relatively superior in retrieving LAI and CCC. All input variable scenarios achieved RRMSE of <5%, thus within the GCOS limit of 10% (GCOS, 2009).

Table 8-4. The performance of various Sentinel-2 input variables in retrieving various crop parameters using Random Forest (RF) with two test datasets, i.e., cross-validation data pooled from the two sites and independent test data from Harrismith (in brackets). The **bold** values indicate the best (i.e., lowest) RMSE achieved by the input variable scenario for each biophysical/biochemical variable (BV).

		SB	*SVIs*	*SB + Angles*	*SVIs + Angles*	*Avar*
LAI	R^2	0.59 (0.49)	0.52 (0.58)	0.69 (0.66)	0.67 (0.67)	0.66 (0.67)
	RMSE	0.57 (0.65)	0.62 (0.62)	**0.50** (0.55)	0.51 (**0.53**)	0.52 (0.55)
	RRMSE	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)
	MAE	0.41 (0.48)	0.44 (0.44)	0.34 (0.37)	0.36 (0.36)	0.36 (0.37)
	%Bias	0.00 (-0.03)	0.00 (-0.04)	0.00 (-0.03)	0.00 (-0.03)	0.00 (-0.04)
LC_{ab}	R^2	0.71 (0.49)	0.74 (0.59)	0.72 (0.49)	0.76 (0.61)	0.74 (0.58)
	RMSE	7.16 (7.18)	6.75 (6.31)	7.01 (7.06)	**6.57** (**6.16**)	6.69 (6.43)
	RRMSE	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)
	MAE	5.68 (5.57)	5.30 (4.88)	5.56 (5.32)	5.24 (4.75)	5.57 (4.95)
	%Bias	- 0.01 (0.00)	-0.01 (-0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
CCC	R^2	0.69 (0.69)	0.67 (0.67)	0.72 (0.69)	0.72 (0.72)	0.72 (0.72)
	RMSE	33.48 (26.48)	34.52 (27.86)	**32.07** (26.63)	32.37 (25.98)	32.36 (**24.83**)
	RRMSE	0.02 (0.04)	0.02 (0.04)	0.02 (0.04)	0.02 (0.04)	0.02 (0.03)
	MAE	23.31 (20.96)	23.88 (19.38)	22.60 (20.16)	22.39 (18.19)	22.68 (18.12)
	%Bias	0.00 (0.00)	-0.01 (-0.06)	0.00 (-0.02)	-0.01 (-0.04)	-0.01 (-0.03)

As shown in **Figure 8-5**, all the experimental scenarios resulted in underestimations, to some extent, of all the BVs considered. The LAI seems to saturate at ~5 $m^2 m^{-2}$ with a %Bias of up to 4% with the independent test data. The retrievals of LC_{ab} saturates at ~50 µg cm^{-2}. Finally, using the cross-validation test dataset, the CCC saturates at ~250 µg cm^{-2}, while the same occurs at relatively low values, i.e., ~200 µg cm^{-2} with the independent held-out test dataset. It should be noted, however, that the range was lower in the latter test dataset. As shown in **Table 8–4**, the %Bias was slightly worse in models which did not incorporate per-pixel view and illumination geometry covariates, i.e., up to 6%, across all BVs. Overall, the results from the two test datasets indicate that incorporating pixel-wise sensor view and illumination geometries reduced the % Bias for LAI, LC_{ab} and CCC.

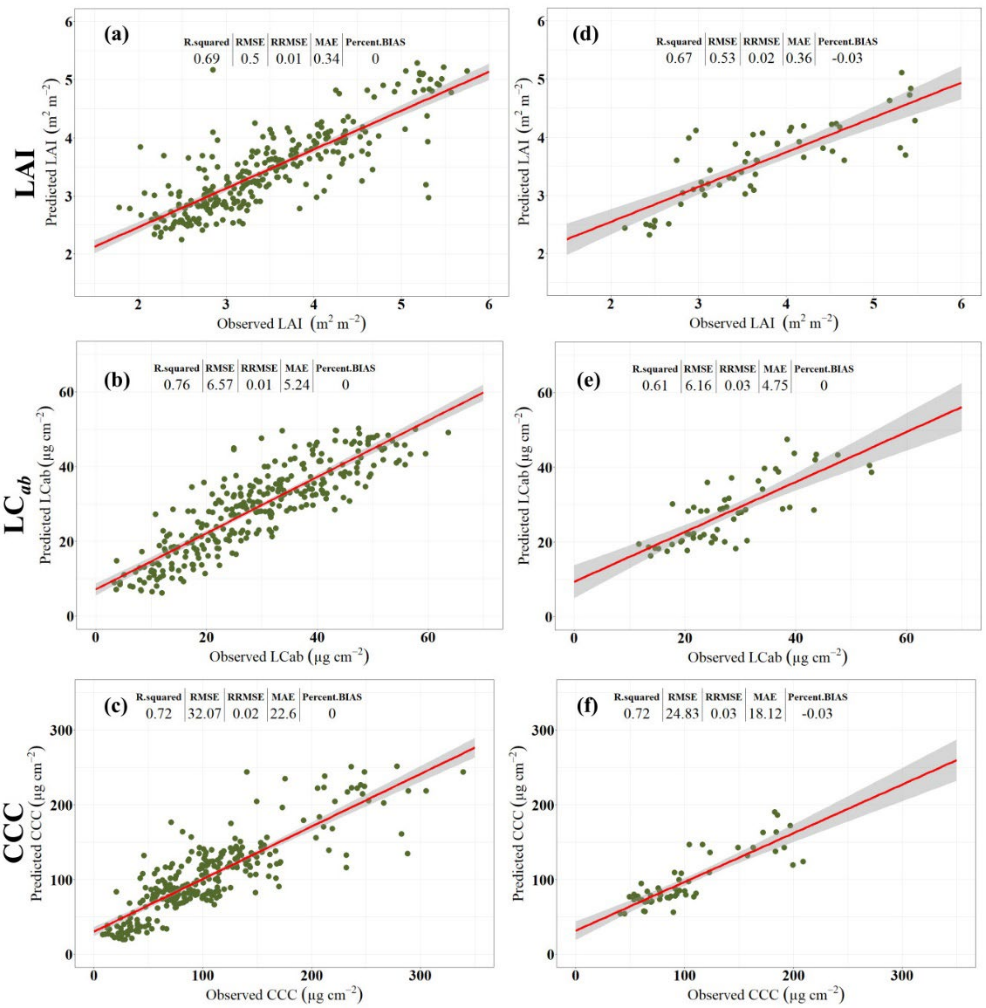


Figure 8-5. Scatterplots for the best Random Forest (RF) models for retrieving various BVs using various Sentinel-2 covariates, cross-validation test dataset pooled from Bothaville and Harrismith (i.e., a–c) and an independent held-out (i.e. 30%) dataset from Harrismith (i.e., d–f). *SB + Angles* achieved the best LAI and CCC predictive accuracy with cross-validation test datasets (i.e., a and c), while *SVIs + Angles* resulted in superior predictive accuracy for LC_{ab} with both test datasets (i.e., b and e), and for LAI with the Independent held-out (i.e., 30%) dataset (i.e., d). *Avar* achieved superior predictive accuracy for CCC with the independent held-out dataset (i.e., f).

The spatial distribution maps of crop biophysical and biochemical variables are presented in **Figure 8-6**. The maps were derived with the best RF models, which incorporated geometric variables, i.e., *SB + Angle*s and *SVIs + Angles* (**Table 8–4**)

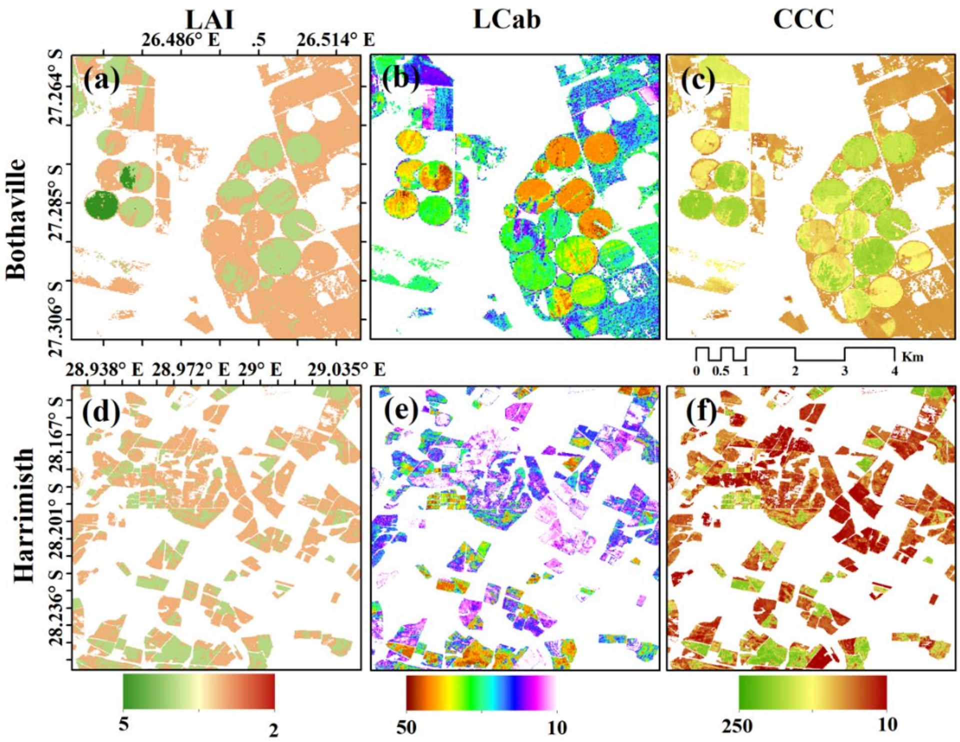


Figure 8-6. Spatial distribution of crop biophysical and biochemical variables for Bothaville (a–c) and Harrismith (d–f). The LAI and CCC maps in (a) and (c) were achieved with *SB + Angles*, while LC_{ab} in (b) and (e) *SVIs + Angles*. *Avar* was used to map CCC in (f).

4.3. Variable importance

We evaluated the contribution of each input variable to the BV retrieval models with the RF variable importance measure, i.e., the %IncMSE. For brevity, we exclude the results for *SB* and *SVIs* which did not incorporate sensor view and illumination geometries as covariates because our interest was on the contribution of geometric covariates. Nevertheless, the results for *SB* were reported in our previous work (Kganyago et al., 2021). Generally, all evaluated input variable scenarios (i.e., that incorporated geometric covariates in the modelling of BVs) show the considerable importance (or influence) of the per-pixel illumination and viewing geometries on crop BV retrieval (**Figure 8-7**). The RF variable importance results for LAI (**Figure 8-7**a, d, and g) show that VZA and SZA were the topmost important input variables, consistently showing a high %IncMSE of >20% when using the *SB + Angles* and *SVIs + Angles*, while they had an average %IncMSE of ≥10%<20% when using *Avar*. The other geometric covariate, i.e., RAA, had an average %IncMSE (i.e., ≥10% <20%) in all the scenarios, along B3, B8A, and B5, and SWIR bands B11 and B12 when using *SB + Angles*, CI_{RE_740}, MTCI, NDI45, MCARI, IRECI, and $RERI_{705}$ when using *SVIs + Angles*, and CI_{green}, CI_{RE_740}, B3, MTCI, B8A, and B12 when using *Avar*. The visible bands, B2 and B4, were among the highly influential variables in the LAI model using *SB + Angles* (i.e., %IncMSE >20%), while B2 was among the top three influential variables in the LAI model using *Avar*.

In **Figure 8-7**b, e, and h, the contribution of geometric covariates is also evident, where when using *SB + Angles* as input variables, VZA was the only geometric variable featured in the top 10 most important variables. When using *SVI + Angles* and *Avar*, only SZA (among other geometric covariates) was the topmost influential variable. The radiometric variables, B5, B11, B8A, and B4, were more influential in the *SB + Angles*-LC_{ab} models, while MCARI, $RERI_{705}$, NDI45, REIP, and S2REP were among the highly influential among the *SVI + Angles* variables. For *Avar*, B11, $RERI_{705}$, MCARI, B5, NDI45, and REIP had the most influence on the LC_{ab} model. Similar to LAI, the RF variable importance results (**Figure 8-7**c, f, and i) show that VZA and SZA seem to play a pivotal role in CCC retrieval, where they were consistently featured among the most important variables in all experimental scenarios, *SB + Angles*, *SVI + Angles* and *Avar*. In addition to these geometric covariates, B5, B3, B11, and B12 were the topmost variables when using *SB + Angles*. However, MTCI, S2REP, $RERI_{705}$, and REIP ranked topmost in *SVI + Angles*, while B5, B3, S2REP, REIP, and MTCI were topmost in *Avar*.

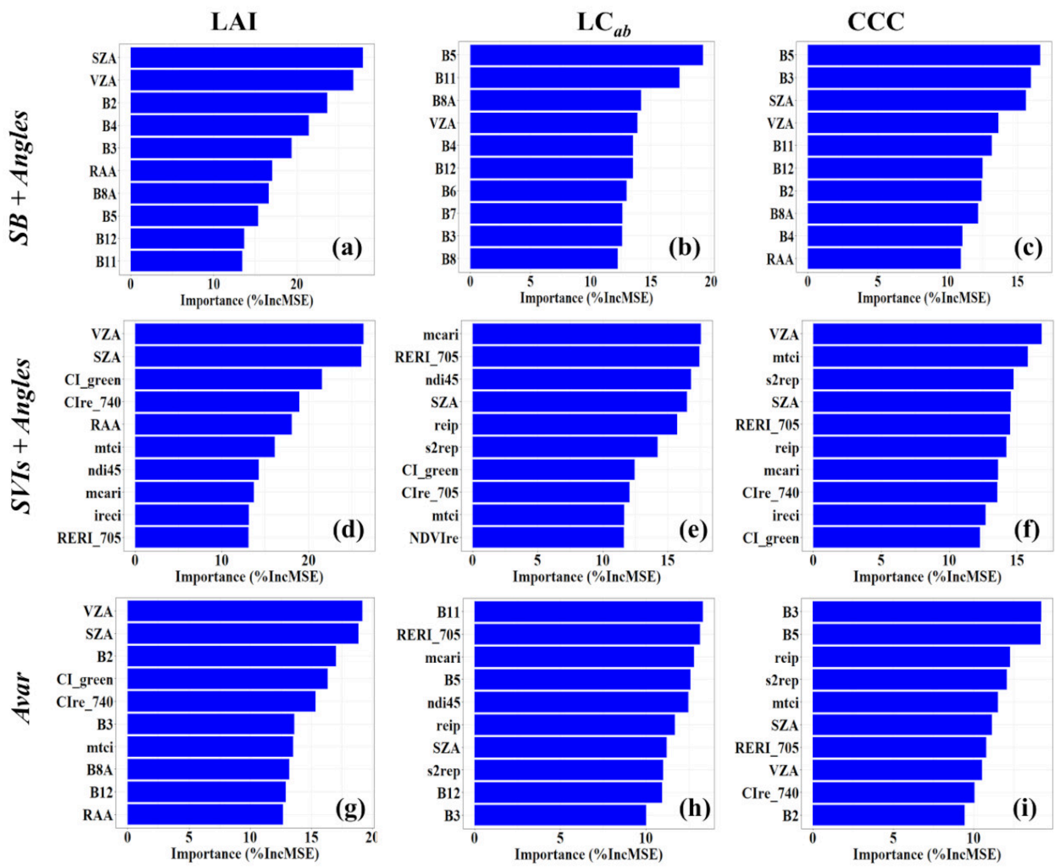


Figure 8-7. Top 10 important variables (high to low) for retrieving LAI, LC$_{ab}$, and CCC using each experimental scenario identified with Random Forest (RF) and 5-fold cross-validation with 100 iterations.

5. Discussion

Understanding the contribution of view and illumination geometry to the accuracy of retrieving crop BVs is critical for reducing the retrieval uncertainty of these BVs by accounting for geometric sources of variability in the reflectance data. As Jacquemoud et al. (2009) note, the view and illumination geometry profoundly impact the canopy reflectance. Therefore, incorporating geometric variables provides an opportunity to exploit additional structural information content or to minimise their effects where it causes noise (Roy et al., 2017, Verrelst et al., 2008). Previous studies focused on the effects of varying illumination conditions resulting from topography (Martín-Ortega et al., 2020), correction of surface anisotropy effects for inter-sensor comparisons(Roy et al., 2017) and global effect of either view or illumination geometry on vegetation reflectance (Pocewicz et al., 2007, Verrelst et al., 2008, Ranson, 1986). However, these studies were done without considering the within-scene, per-pixel view and illumination geometry contribution to the retrieval of BVs.

To address this gap, we evaluated various input variable scenarios (**Table 8-2**) for crop BVs retrieval using the RF regression algorithm. These experimental scenarios were designed to elucidate the contribution of view and illumination geometries to the crop BVs retrieval accuracy. Since Sentinel-2 has a relatively wide FOV, i.e., ~20.6° (i.e., ~290 km), compared to relatively narrow FOV sensors such as Landsat-8 OLI, i.e., ~15° (~190 km), it is prone to within-scene view and illumination geometry variations (**Figure 8-3**), with potential to affect BVs retrieval. To highlight view and illumination geometry contribution to model performance practically, we assessed the predictive power of all input variables per experimental scenario using RF variable importance (%IncMSE). Furthermore, the adjusted R^2 of multiple linear regression was used to determine the amount of spectral reflectance variation explained by geometric covariates (**Table 8-4**).

5.1. Contribution of geometric covariates to BVs retrieval accuracy

The results showed that the experimental scenarios incorporating per-pixel geometric covariates, i.e., SZA, VZA, and RAA, yielded generally robust retrievals of LAI, LC_{ab}, and CCC. For example, adding geometric covariates to the standard input variables for retrieval of LAI using cross-validation test data, i.e., *SB* and *SVIs*, showed improvements in R^2 of 10% and 15%, respectively. At the same time, RMSEs improved by 0.07 m^2 m^{-2} and 0.11 m^2 m^{-2}, respectively. These improvements also manifest in the derived spatial distribution maps, which show with-field and between-field spatial variations in LAI when geometric variables are excluded (i.e., *SB* and *SVIs*) and included (i.e., *SB + Angles* and *SVIs + Angles*) in the model (**Figure 8-S1** and **Figure 8-S2**). The independent held-out test dataset showed improvements in R^2 by 17% and 9% and RMSEs by 0.1 m^2 m^{-2} and 0.09 m^2 m^{-2}. However, relatively few improvements were realised for CCC, where *SB + Angles* achieved improvements of 3% in explained variability and RMSE improvements by only 1.41 µg cm^{-2} using a cross-validation test dataset. In contrast, the independent held-out dataset showed no improvements (instead, a decline in RMSE by 0.15 µg cm^{-2} was observed). Using *SVIs + Angles*, 5% improvements in explained variability were achieved using both test datasets, respectively, while RMSEs improved by 2.15 μg cm^{-2} and 1.88 μg cm^{-2} for the cross-validation and independent held-out test datasets, respectively. For LC_{ab}, the improvements in R^2 and RMSE were marginal and possibly trivial. When the geometric covariates were added to *SB* (i.e., *SB* + *Angles*), only a 1% increase in R^2 and RMSE increase by 0.15 μg cm^{-2} and 0.12 μg cm^{-2} were achieved with both test datasets, respectively. Similarly, *SVIs* + *Angles* led to a 2% improvement in explained variability and RMSE improvements by 0.18 μg cm^{-2} and 0.15 μg cm^{-2} for the cross-validation and independent held-out test datasets, respectively. LC_{ab} (measured at the leaf level) is usually difficult to estimate from canopy reflectance, which contains the influence of the background, plant structural, and biochemical traits. That is why the variability introduced by geometric covariates was not crucial for LC_{ab} estimation in this study. Consequently, spatial distribution maps derived with standard variables, i.e., *SB* and *SVIs* show no apparent differences (**Figure 8-S1** and **Figure 8-S2**).

Indeed, the *SVIs* offer a simple and rapid assessment of crop health, with several studies showing their relationship with and accuracy in retrieving crop BVs (Jay et al., 2017, Main et al., 2011, Haboudane et al., 2008). In retrieving the LAI, CCC, and canopy nitrogen content of sugar beet crops, Jay et al. (2017) found that the SVIs perform equivalent to or slightly better than PROSAIL inversion. It is, therefore, not surprising that the *SVIs* such as MTCI, S2REIP, REIP, NDI45, MCARI and CI_{RE_740} were consistently ranked among the top variables when using *Avar* across all BVs. Also, the *SVIs* scenario was superior to *SB*, *SB* + *Angles*, and *Avar* in retrieving LC_{ab}. Vegetation indices such as MTCI, RVI, and NDI45 were previously found to be highly correlated to LAI and chlorophyll content (Viña et al., 2011). In agreement, Shah et al. (2019) found that MTCI is one of the best-performing SVIs with an R^2 of 0.86 and an RMSE of 6.07 μg cm^{-2} in estimating wheat LC_{ab} based on its sensitivity to wide-ranging chlorophyll content. Similarly, Clevers and Gitelson (2013) observed that the MTCI is an accurate estimator of CCC and nitrogen (N) content based on its linear relationship with these BVs. The results in this study are consistent with these previous studies and others (Main et al., 2011), using cross-validation and independent held-out test datasets. Moreover, *SVIs* selected in this study yielded comparative results to *SB*, except in exceptional cases, i.e., the *SVIs*-LC$_{ab}$ model, where the former was superior. Their excellent performance can be attributed to their relatively low sensitivity to background soils, and residual atmospheric effects, which affect most VIS/NIR indices (Liu et al., 2012). However, their relatively low performance relative to *SVIs + Angles* signifies their sensitivity to per-pixel view and illumination geometry, as consistently reported previously (Verrelst et al., 2008, Jay et al., 2017). For example, Verrelst et al. (2008) found that broadband and narrow-band greenness indices exhibit significant sensitivity to view geometry. In the current study, when per-pixel view and illumination geometric covariates were incorporated (i.e., *SVIs + Angles*), the variability explained improved from 52% (58%), 74% (59%) and 67% (67%) when using *SVIs* only to 67% (67%), 76% (61%) and 72% (72%) for LAI, LC$_{ab}$, and CCC, respectively, using cross-validation test (independent held-out test) datasets. Moreover, when per-pixel view and illumination geometric covariates were incorporated into *SB* (i.e., *SB + Angles*), the variability explained improved from 59% (49%), 71% (49%), and 69% (69%) to 69% (66%), 72% (49%), 72% (69%) for LAI, LC$_{ab}$, and CCC, respectively, using cross-validation test (independent held-out test) datasets. These results signify that the per-pixel view and illumination geometry affect the upwelling radiance from crop canopies as determined by the interaction of various leaf and canopy traits. Such effects were not accounted for during atmospheric correction and thus were propagated to surface reflectance.

Therefore, incorporating pixel-wise view and illumination geometry into the retrieval of BVs is reasonable since it is one of the factors influencing the canopy reflectance. In this study, its inclusion into *SB* and *SVIs* ensured that the variability in reflectance caused by the view and illumination geometry is accounted for, thus leading to improved predictive accuracy relative to using the *SB* or *SVIs* only. As shown by the multiple linear regression analysis (**Table 8-3**), more than 30% of the variability in all Sentinel-2 spectral bands was explained by the per-pixel variation in view and illumination geometries at a 5% significant level ($p < 2.2e\text{-}16$). A previous study (Verrelst et al., 2008) found that topographic and illumination geometry accounted for about 13% of the variability of SVIs. Similarly, Rahman et al. (1999) found that the canopy reflectance of rice and wheat varies significantly with view geometry, where the change in view geometry resulted in different spectral responses owing to the exposed soils and vegetation for each geometry. Nagol et al. (2015) found that the effect of narrow VZA (i.e., ±7°) in Landsat data resulted in non-negligible variations in reflectance, i.e., 20%.

5.2. The predictive power of the geometric covariates

The additional predictive power provided by the geometric covariates was also shown by the RF variable importance results, where one or more view and illumination geometry covariates were among highly informative variables in all biophysical and biochemical models (**Figure 8-7**). Among the geometric covariates, VZA, and SZA were the most influential in retrieving LAI and CCC. In particular, all scenarios showed the high influence (ranked by their %IncMSE) of SZA and VZA for the retrieval of LAI and CCC. Other scenarios, i.e., LC_{ab} retrieval with *SB + Angles*, showed that VZA was the only geometric variable featured in the top 10 important variables (**Figure 8-7**). In contrast, the retrieval of LC_{ab} using *SVIs + Angles* and *Avar* showed SZA as the only geometric variable with a strong influence on the model. The findings in this study correspond to Verrelst et al. (2008) that the spectral response of plant canopies is significantly affected by the sensor view geometry, affecting the accuracy of retrieved BVs.

Moreover, the findings not only ascertain previous studies that demonstrated solar illumination effects on SVIs (Galvao et al., 2011) but also indicate that per-pixel SZA can contribute essential additional information that improves the predictive power of SVIs, thus reducing the uncertainty in BV retrieval. SZA contains per-pixel geometric information of the incident radiation. Critically, within-scene SZA changes with the crop growth stages (i.e., time of year); hence ignoring its effects may lead to inconsistent BV retrieval and subsequent incorrect interpretations. For example, Galvao et al. (2011) found that intra-annual variability in enhanced vegetation index (EVI) from MODIS and nadir Hyperion was regulated by illumination effects than actual changes in LAI. In the context of Precision Agriculture and crop monitoring, such false alarms, i.e., caused by geometric variability, have profound implications. It may lead to incorrect crop condition diagnosis that can lead to wasteful use of scarce and expensive farm inputs (e.g., irrigation water, pesticides, and fertilisers) and poor yields as the stressed crops are not timely and adequately remedied. This can potentially lead to loss of trust in remotely sensed crop BVs for agronomic applications.

As one of the factors affecting reflectance variability (**Table 8-3**), per-pixel geometric information has been ignored in retrieving BVs with optical satellite images. This may be a consequence of the absence of spatially explicit view and illumination information by heritage missions such as Landsat and AVHRR that largely drove the development of BV retrieval techniques. As the results in this study have shown, the per-pixel geometric variables explain a reasonable amount of variability in satellite-based canopy reflectance; hence should be incorporated into MLRA models to improve the accuracy and reliability of crop BVs. Therefore, it is reasonable to deduce that spectral bands or vegetation indices combined with geometric covariates offer the most significant prospects of reliable prescription maps for Precision Agriculture technologies, such as Variable Rate Application (VRA) of fertiliser and pesticides, and to aid agronomic decisions. Although SNAP biophysical processor, based on PROSAIL LUT and ANN incorporate these geometric variables (Weiss & Baret, 2020), validation studies (Kganyago et al., 2020, Xie et al., 2019, Djamai et al., 2019) have shown that the product is associated with significant uncertainties which render its site-specific agronomic application questionable. Therefore, based on the results here and in line with Combal et al. (2003), it can be assumed that locally parameterising the PROSAIL LUT (using prior information) may reduce high uncertainty in BV retrieval and show the contribution of the geometric covariates.

Therefore, models with *SB* or red-edge *SVIs* and pixel-wise geometric covariates are more promising for LAI and CCC retrieval (**Figures 8-S1** and **Figure 8-S2**). This is attractive for Precision Agriculture applications since these geometric covariates come standard with Sentinel-2 products and can be conveniently incorporated into BV retrieval models to account for per-pixel view and illumination geometry effects and achieve unprecedented accuracies and temporal consistency for relatively wide FOV sensors such as Sentinel-2.

5.3. Limitations of the study

Despite the promising results, the main limitation of this study was that the applicability of MLRA models constructed with experimental data is limited by the representativeness of such data; thus, the results obtained here may be specific to crop type, region, and phenology. Therefore, further experiments in different geographic locations and periods are required to ascertain the contribution of per-pixel geometric covariates. Moreover, simulated data using RTMs is recommended for future works to make the models widely applicable and assess the influence of geometric variables on BVs. Nonetheless, the multiple linear regression analysis showed a non-negligible influence on surface reflectance of the considered crop types, while RF variable importance indicated that they are influential to the model performance and accuracy of retrieving LAI, LC_{ab} and CCC. Overall, this study found that per-pixel view and illumination geometric covariates improve MLRA crop BVs retrieval. Future studies should explore the optimisation of Sentinel-2 feature space (including derived vegetation indices using feature selection to avoid potential collinearity of predictors and improve the practical prospects of combining geometric and radiometric variables in retrieving crop BVs.

6. Conclusion

This study evaluated the influence of Sentinel-2 pixel-wise sensor view and solar illumination geometry on the accuracy of estimating crop biophysical and biochemical variables relevant for Precision Agriculture and crop monitoring, i.e., LAI, LC_{ab}, and CCC. Generally, the results showed that the experimental scenarios incorporating per-pixel geometric covariates, i.e., SZA, VZA, and RAA, yielded robust retrievals of LAI and CCC, while there was a negligible improvement (i.e., R^2: 1%–2%) in LC_{ab} retrieval accuracy. In particular, standard input variables, such as spectral bands and spectral vegetation indices, to biophysical retrieval models, combined with geometric covariates resulted in the best (i.e., lowest) RMSE for all considered crop BVs owing to high information content in the spectral bands and red-edge indices, and residual variability explained by the pixel-wise view and illumination geometry. The results showed that this variability exceeded 30% in all Sentinel-2 bands and sites. Moreover, variable importance assessment showed that VZA and SZA were among the topmost influential variables in retrieving all BVs. The spatially explicit view and illumination information provided with Sentinel-2 can be used as additional information to account for variability resulting from sensor and sun geometries in BV retrieval with MLRAs. This study showed that ignoring the influence of sensor and sun geometry may lead to inconsistencies in BV retrieval as the SZA changes along the cropping season, and VZA varies across the scene. Moreover, their influence on reflectance is considerable, i.e., >30%; thus, estimated BVs can be incorrectly interpreted, as shown by Galvao et al. (2011). In Precision Agriculture, it may lead to an inaccurate crop health diagnosis, risking wasteful use of irrigation water, pesticides, and fertilisers, poor yields, and loss of trust in remotely sensed BVs for agronomic applications. Based on the results of this study, we recommend incorporating per-pixel view and illumination geometry in estimating canopy-level BVs such as LAI and CCC, especially when using wide-view sensors such as Sentinel-2. However, further tests are required to reaffirm the contribution of view and illumination geometry towards improving the crop BVs that support Precision Agriculture and aid agronomic decisions.

Appendix 8-1. Supplementary Material

The spatial distribution maps of biophysical and biochemical variables (BVs) derived from standard radiometric input variables, i.e., spectral bands (*SB*) and spectral vegetation indices (*SVIs*), and radiometric variables combined with view and illumination geometry are presented in **Figure 8-S1** and **Figure 8-S2**. As shown, there is some differences in the spatial distributions of the BVs. For example, LAI values in **Figure 8-S1**a and d vary considerably when comparing the same fields between the two maps. In one of the fields (see red boxes), *SB* (**Figure 8-S1**a) exhibits higher LAI values (i.e., around 5 m^2 m^{-2}), while the same field show relatively low LAI values (i.e., around 4 m^2 m^{-2}) in *SB + Angles*. The differences can also be observed in the blue box, where the LAI values are smaller and higher on some fields. These differences can be attributed to the saturation of LAI values around ~5 m^2 m^{-2} when using spectral bands (*SB*) only. The spatial variations of CCC can also be seen in **Figure 8**-S1c and f, where *SB + Angles* exhibits relatively high CCC values. The inclusion of view and illumination geometric variables provided additional structural information that distinguished the dissimilarities in LAI and CCC values between and within fields. Similarly, **Figure 8-S2**a/c and d/f show some spatial variations of LAI and CCC between experimental scenarios, i.e., *SVI* and *SVI + Angles*, thus indicating that the influence of view and illumination geometry. The red-edge SVIs chosen in the current study (**Table 8-2**) are highly correlated to crop BVs, and have relatively low sensitivity to background soils, and residual atmospheric effects (Liu et al., 2012) as discussed in Section 5.1, but their accuracy is greatly enhanced, and %Bias is reduced when incorporating geometric variables. LC_{ab} maps derived from *SB* and *SB + Angles* and *SVI* and *SVI + Angles* have indistinguishable spatial variability (**Figure 8-S1** and **Figure 8-S2**b and e). This is consistent with the findings of this study, that found that trivial improvements (i.e., R^2 of 1% to 2%) when incorporating geometric variables.

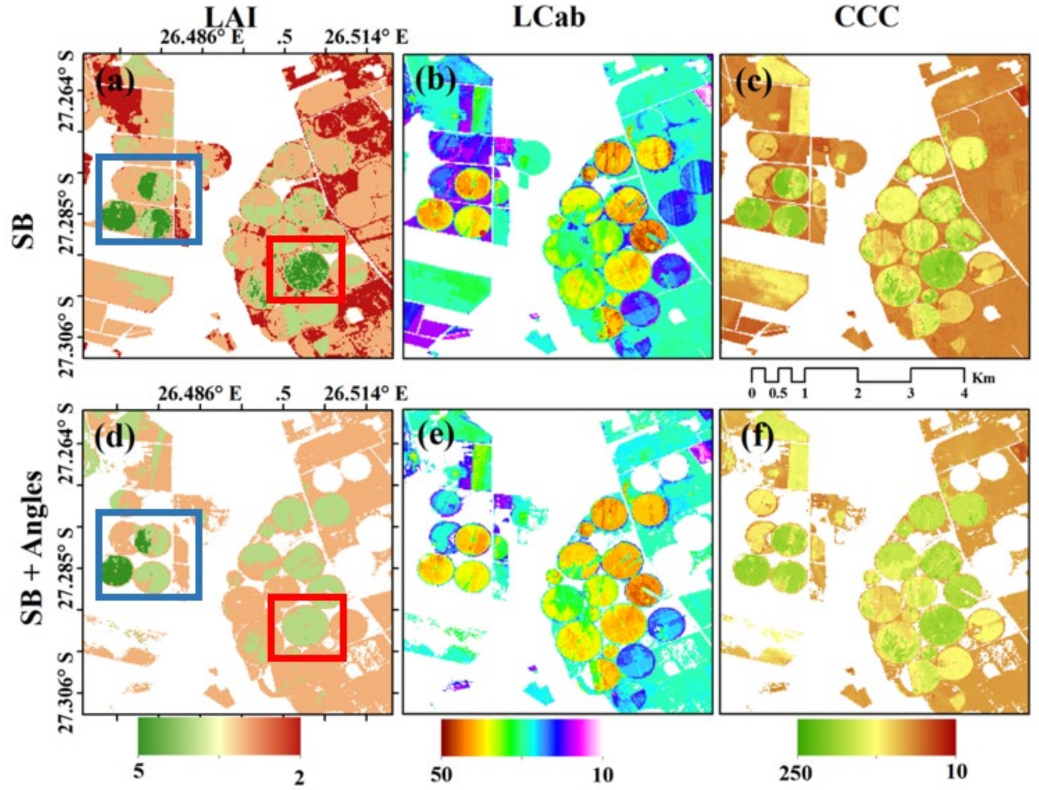


Figure 8-S1. Comparison of the spatial distributions of LAI, LC_{ab}, and CCC derived from *SB* (a–c) and *SB + Angles* (d–f).

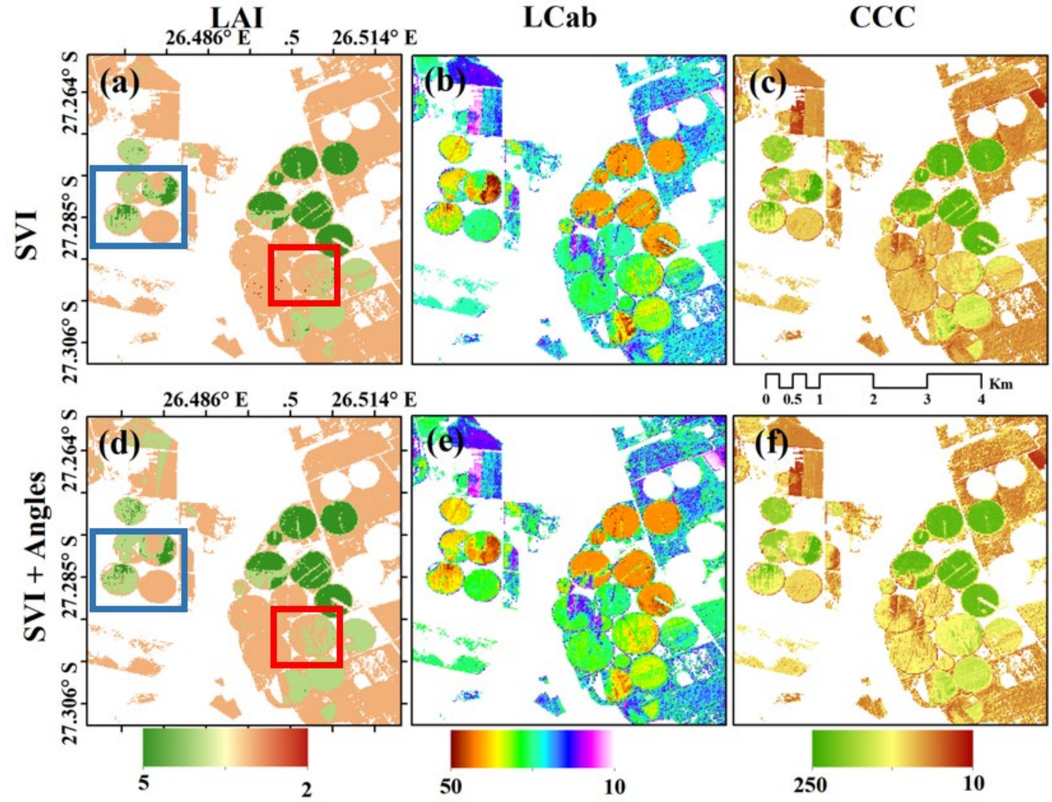


Figure 8-S2. Comparison of the spatial distributions of LAI, LC$_{ab}$, and CCC derived from *SVIs* (a–c) and *SVIs + Angles* (d–f).

CHAPTER 9
OPTIMISING FEATURE SPACE FOR IMPROVED CROP BIOPHYSICAL AND BIOCHEMICAL VARIABLES RETRIEVAL USING THE NOVEL SPECTRAL TRIAD FEATURE SELECTION ALGORITHM ON SENTINEL-2 COVARIATES

This chapter is based on the following manuscript:

Kganyago, M., Adjorlolo, C., Mhangara, P. (**Submitted**). Optimising Sentinel-2 Feature Space for Improved Crop Biophysical and Biochemical Variables Retrieval Using the Novel Spectral Triad Feature Selection Algorithm. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.* [Manuscript ID: JSTARS-2022-01690]

Abstract

Machine Learning Regression Algorithms (MLRAs) are renowned for their capability to learn complex and non-linear associations between the response and predictor variables. However, studies have shown that feature subset selection is more beneficial, yielding high accuracy and low uncertainties in retrieving biophysical and biochemical variables. Feature subset selection techniques are often applied with highly dimensional and correlated hyperspectral data, but are seldom used with the multi-spectral dataset. Instead, previous studies utilising multi-spectral data have mainly applied the entire feature space. The availability of quasi-hyperspectral sensors, e.g., Sentinel-2, presents new challenges where two or more variables may be collinear and impact the performance of MLRA. This study presents a novel Spectral Triad feature selection technique based on music theory and compares it to the entire MSI feature space and Random Forest-Recursive Feature Elimination (RF-RFE). The optimal subsets were evaluated with Random Forest for retrieving Leaf Area Index (LAI), Leaf Chlorophyll *a* + *b* (LC_{ab}) and Canopy Chlorophyll Content (CCC) in a semi-arid agricultural landscape. The results indicated that the proposed STfs algorithm obtained equivalent or better (i.e., by 1%–3%) retrieval results for LAI (R^2_{cv} of 66%, $RMSE_{cv}$ of 0.53 $m^2\ m^{-2}$), LC_{ab} (R^2_{cv}: 74%, $RMSE_{cv}$: 7.09 $\mu g\ cm^{-2}$) and CCC (R^2_{cv}: 77%, $RMSE_{cv}$: 33.69 $\mu g\ cm^{-2}$), using only five, seven and seven variables, respectively, when compared to RF-RFE and entire MSI feature space. Overall, the proposed STfs algorithm has great potential to optimise the spectral feature space of quasi-hyperspectral sensors for rapid crop biophysical and biochemical parameter retrieval.

1. Introduction

The retrieval of biophysical and biochemical variables (BVs) over agricultural areas has become a critical task in recent years based on the need to optimise yield and farm inputs and promote the sustainability of agroecological landscapes. The BVs such as Canopy Chlorophyll Content (CCC), Leaf Area Index (LAI), and Leaf Chlorophyll $a + b$ (LC_{ab}) have been identified as the critical plant growth indicators, essential for supporting Precision Agriculture technologies for fertiliser and pesticides application and irrigation management. LAI, defined as one-side leaf area per unit area (i.e., $m^2\,m^{-2}$), is a significant plant structural parameter indicative of the different growth conditions and phenology of the crops, while chlorophyll content at leaf and canopy levels is essential to detect crop stress from biotic and abiotic stressors, and gross primary productivity (Gitelson et al., 2014) and is highly correlated to nitrogen content. N content is one of the most limiting factors in plants and it is critical for crop growth, health, and yield (Daughtry et al., 2000, Gitelson et al., 2003). Remotely sensed parameters of plant structural properties and N content such as LAI, LC_{ab}, and CCC reduce the need for destructive, laborious, and costly direct estimation and lab-based methods. They are thus crucial for optimising irrigation and N fertilisation for site-specific management and yield (Vincini et al., 2016, Jia et al., 2013).

The approaches to date for the retrieval of BVs can be divided into four broad categories, i.e., spectral vegetation indices (*SVIs*), Radiative Transfer Models (RTMs), Machine Learning Regression Algorithms (MLRAs) and hybrid techniques integrating the Look-Up-Tables (LUTs) generated with RTMs with MLRAs (Frampton et al., 2013). Among these, hybrid techniques have the greatest potential for wide application as the RTM-simulated LUTs can represent varying acquisition conditions and crop structural, biochemical, and physiological conditions while MLRAs can estimate the complex and non-linear relations between the response and predictor variables. However, the RTMs require substantial site-specific data (Verrelst et al., 2012) and their universality claim has not been yet fully explored, as studies assessing the transferability of RTMs are often in similar climatic environments (e.g., Mediterranean climates). Operational techniques such as Artificial Neural Networks (ANN) and PROSAIL-generated

LUTs have further been found to perform poorly in semi-arid regions (Kganyago et al., 2020a). Studies (Verrelst et al., 2015, Verrelst et al., 2012, Kganyago et al., 2021) establish that simpler tree-based and kernel-based MLRAs can perform as effectively as complex algorithms (e.g., ANN). Therefore, the power of MLRAs can be combined with real data consisting of various informative variables.

The optimisation of the feature space, i.e., selection of informative feature subset, for BV retrieval is critical to reducing errors and uncertainties, thus increasing trust in remotely sensed BVs for farm decisions and site-specific management. Feature subset selection is commonly applied in studies utilising hyperspectral data that are characterised by many thousands of narrow and adjacent spectral bands. Some studies (Verrelst et al., 2016, Verrelst et al., 2011) have shown that only four to nine bands, selected through feature selection techniques, are sufficient to achieve robust retrievals of BVs. However, hyperspectral data are limited to small areas and are costly to acquire. In the current study, we argue that feature subset selection is also relevant for the retrieval of crop BVs from multi-spectral data owing to the high dimensionality of various possible covariates that can be derived such as spectral bands, SVIs, and textural measures. Moreover, the launch of quasi-hyperspectral sensors such as Sentinel-3, Worldview-2 and its successors, and Sentinel-2 Multi-Spectral Instrument (MSI)—characterised by many multi-spectral broad and narrow bands that are strategically positioned for vegetation characterisation (Fitzgerald et al., 2010, Delegido et al., 2013, Dimitrov et al., 2019)—may introduce dimensionality and collinearity, which has been shown to affect MLRA performance (Yu & Liu 2004). Moreover, redundant variables in the dataset introduce additional computational burden to an already costly process of tuning hyperparameters, and training, and testing the machine learning models.

Generally, the optimisation of variables for retrieval of crop BVs has not been extensively explored, especially as this relates to quasi-hyperspectral sensors. In fact, previous studies (Kganyago et al., 2021) either applied for the entire feature space (i.e., visible, near-infrared, and shortwave infrared) from such sensors or combined all spectral bands with derived covariates such as SVIs and textural measures, which are often arbitrarily selected or based on prior knowledge (Sibanda et al., 2017). For example, in estimating grassland LAI, Li et al. (2017a) integrated Landsat Operational Land Imager (OLI) bands with seven vegetation indices and found improvements in accuracy once the variables were optimised, while Ramoelo et al. (2015b) utilised WorldView-2 spectral bands and vegetation indices as input to a RF model in estimating grass quality indicators, i.e., leaf N and above-ground biomass, to R^2 of 89% and 84%, respectively. However, the selection of covariates based on prior knowledge is beset with subjectivity and uncertainty owing to the complex and interacting environmental and climatic factors that affect biophysical and biochemical traits in vegetation canopies, thus the canopy reflectance. In agricultural environments, the selection of relevant covariates using prior knowledge is further complicated by the varying agricultural management practices, phenology, crop types, and limiting factors such as soil nutrients and water. Therefore, there is demand for statistical data-based feature selection techniques capable of associating each (or a set of) covariates to the BVs of interest and assign a rank (e.g., importance) based on linear and non-linear relationships among covariates or between covariates and BVs. The statistical data-based approaches, commonly classified as filter-based (e.g., analysis of variance), wrapper-based (e.g., Recursive Feature elimination-Support Vector Machine), and embedded (e.g., sparse Partial Least Squares) algorithms, are attractive for optimising crop BVs for site-specific management in various environments and field conditions, and are more objective since they determine the few optimal variables based on the information content of the covariates, while also eliminating collinearity and reducing training time. It is, therefore, essential to evaluate existing and develop new approaches to optimise crop BV retrieval with Sentinel-2 MSI, i.e., a quasi-hyperspectral sensor.

This study aimed to optimise crop BV retrieval using feature selection techniques to identify and select relevant remotely sensed covariates from Sentinel-2 data. To achieve this, we propose a novel Spectral Triad feature selection (STfs) algorithm inspired by music theory. The STfs algorithm is compared to experimental cases where entire MSI feature space is used (i.e., no variable selection) and a popular wrapper-based algorithm, i.e., Recursive Feature Elimination (RFE), coupled with Random Forest (RF-RFE) (Gregorutti et al., 2017) in terms of their accuracy and training times in retrieving LAI, LC_{ab} and CCC. Moreover, we assess the consistency in the selected feature subsets between STfs and RF-RFE in Bothaville, South Africa. The characteristics of the study area are described in our previous studies (Kganyago, 2021; Kganyago et al., 2021; Kganyago et al., 2022).

2. Materials and methods

Figure 9-1 summarises the methods of the study.

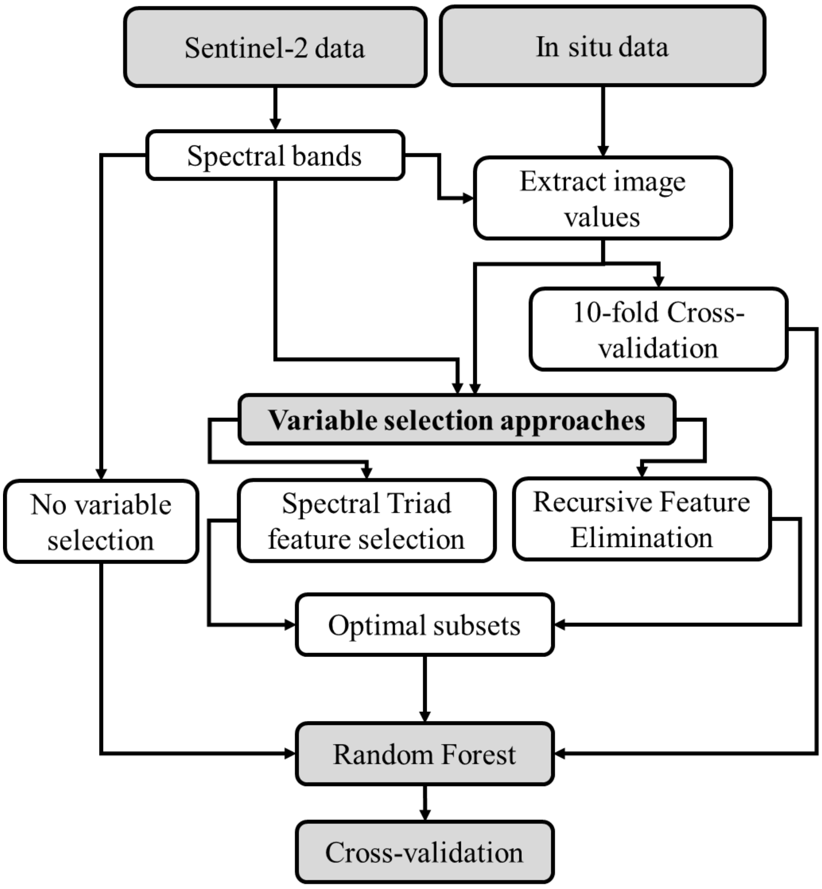


Figure 9-1. An overview of the data and methods used in this study.

2.1. Data

2.1.1. In situ data

The experimental data for this study was collected in Bothaville between 11 and 23 April 2021. The measured crop parameters consisted of LAI and LC_{ab}, using the Plant Canopy Analyzer (Li-Cor 2200c, LI-COR Inc., Lincoln, NE, USA) and Chlorophyll Concentration Meter (MC-100 Apogee Instruments, Inc., Logan, UT, USA), respectively. The Plant Canopy Analyzer was equipped with a view restricting cap of 180° FOV to reduce the influence of the operator and that of neighbouring objects on the LAI measurements. In contrast, LC_{ab} samples were acquired on sun-exposed leaves, where any particular sampling point within the plot comprised a mean of several (i.e., three to five) leaf measurements. For each, 40 m × 40 m plot—selected on a random transect within various crop fields—an average of six to eight random sampling points were measured. The third parameter studied here, i.e., CCC, was obtained as LC_{ab} × LAI (Jacquemoud et al., 2009). The plots were Geo-tagged with a GPS coordinate and a picture using TDC600 (Trimble Inc., Irvine, CA, USA) with an accuracy of 1.5 m. **Table 9-1** shows the descriptive statistics of the measurements.

Table 9-1. Descriptive statistics of the observed LC_{ab} (µg cm^{-2}), and CCC (µg cm^{-2}), and LAI (m^2 m^{-2}).

	n	Min	Mean	Max	SD	Range	Coef. var
LAI	157	1.78	3.43	5.75	0.92	3.97	0.27
LC_{ab}	157	3.32	30.01	63.62	14.92	60.30	0.49
CCC	157	7.87	108.96	339.09	72.35	331.22	66.39

2.1.2. Sentinel-2 covariates

A Sentinel-2 image was collected on 14-04-2021 and was coincidental with fieldwork dates. The atmospherically corrected Level-2A image was retrieved from Sentinel Hub Cloud API (Sinergise Laboratory for geographical information systems, Ltd., Ljubljana, Slovenia). We used 10 m spectral bands located at Band 2:490 nm, Band 3:560 nm, Band 4:665 nm, and Band 8:842 nm, and 20 m bands at Band 5:705 nm, Band 6:740 nm, Band 7:783 nm, Band 8A:865 nm, Band 11:1 610 nm, and Band 12:2 190 nm. To standardise the various resolutions, the 10 m bands were converted to 20 m using nearest neighbour resampling technique in SNAP Toolbox for further analysis.

The selected bands were masked using a crop mask and non-vegetated pixels were masked using NDVI threshold of 0.2 following Kganyago et al. (2021).

2.2. Spectral Triads feature selection (STfs) Algorithm

2.2.1. STfs algorithm

For decades, algorithm development has been inspired by nature with a range of algorithms developed based on the structures of the trees (e.g., Classification and Regression Trees, CART), tree ensembles (i.e., RF), neural activity (e.g., Artificial Neural Networks[ANN]), and genetic mutations (i.e., Genetic Algorithm [GA]). For example, the popular RF algorithm for the classification and regression of remotely sensed images is based on CART but builds many trees instead of one to form a forest. However, GA was designed using theories of natural selection and population genetics mechanisms (Ye 2018), while ANN uses the natural behaviour of animals.

The Spectral Triads feature selection (STfs) algorithm, introduced here, aims to select the most harmonious (i.e., informative) subset of features (e.g., spectral bands) from multi-spectral data. The STfs algorithm is based on music theory, specifically, the major chord formula which consists of the intervals 1-3-5 on a diatonic scale. A chord consists of three or more pitches (or notes) played simultaneously, a diatonic scale is any scale with seven pitches in an octave and consists of five whole steps (or tones) and two half steps (or semitones) in each octave. The major chords, i.e., 1-3-5, consist of the root note (i.e., starting note), a major third (3rd) and the perfect fifth (5th). Therefore, the intervals of the major chords (on a diatonic scale) are an inspiration for the STfs algorithm because their formula remains the same regardless of the root note and ensures a pattern where there is a maximal separation between the half steps (adjacent variables) in multiple octaves. In the context of the STfs algorithm, this interval ensures that each spectral triad contains variables which are adequately separated (i.e., spectrally) from each other, thus avoiding excessive collinearity in the evaluation of the triad. The STfs algorithm uses the major triad chord formula, where 1-3-5 represents the starting variable (root note, *Rn*), and the third and the fifth variables, respectively. The *Rn* refers to the starting variable (an index of the explanatory variables in the dataset), which is 1 by default. The octaves, in the context of STfs, are the search iterations or sequences (progressions) that stop when there are no more unique spectral triads to evaluate or all the explanatory variables in the dataset are evaluated (i.e., equivalent to when all notes on the diatonic scale are played).

For each triad, an evaluation criterion is used to rank the explanatory variables where the best single variable in a triad is selected. The evaluation criteria in the filter mode of STfs (Algorithm 9-1) can be an Entropy-based measure (e.g., Information gain) or ReliefF. While, in wrapper mode, it uses an explainable machine learning algorithm such as Multiple Linear Regression (MLR) analysis or Random Forest variable Importance measure (Percent Increase in Mean Squared Error, %IncMSE). The overview of the algorithm is outlined in **Figure 9-2**.

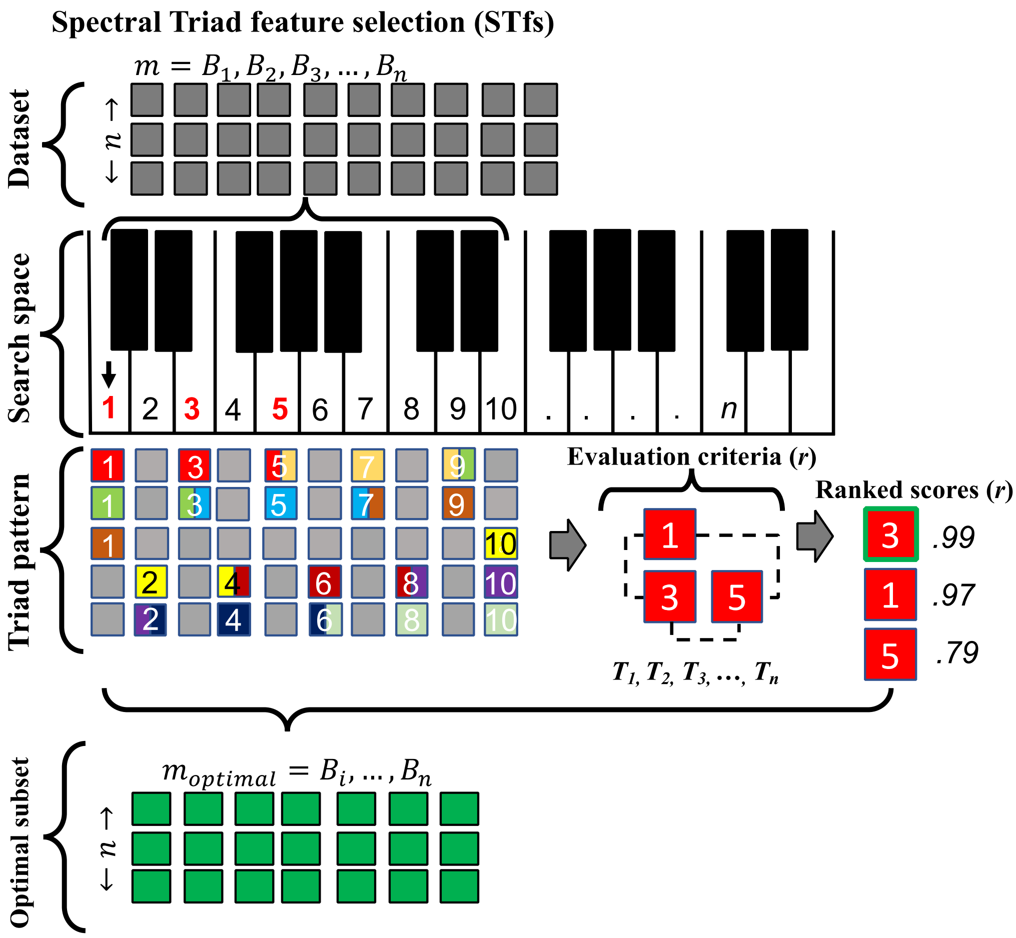


Figure 9-2. An overview of the Spectral Triads feature selection (STfs) algorithm, demonstrating a case of variable subset selection with Sentinel-2's ten spectral bands as input dataset superimposed on ten keys to form a search space for the algorithm and correlation analysis r as an evaluation criteria. All possible spectral triads are evaluated using the search space, with the stopping criterion reached when all triads are evaluated. A final subset consists of all best bands (i.e., with the highest rank scores) from each triad.

The STfs algorithm then returns an optimal subset of features P, i.e., optimal features selected from each triad, which is fewer than the original input features.

Indeed, depending on the dimensionality of the dataset, the number of evaluated spectral triads and selected variables in a triad may still retain high dimensionality, and causes loss in computational efficiency. Therefore, in such cases, an optimal desired number of variables or subset size (*Pd*) can be defined. If this value is different from the default *P* (i.e., all optimal features selected from each triad), the algorithm will be triggered to form and evaluate new triads from the default *P* until *Pd* is achieved. If features in the triad cannot be distinguished using the evaluation metric or the evaluation criteria (see Section 2.2.2), e.g., r, R^2, %IncMSE, or the *p*-value, the entire triad is discarded. The algorithm steps are given in Algorithm 9-1.

Given the training data with Number of covariates n, the STfs routine in filter or wrapper modes (i.e., combined with an explainable learning algorithm such as Random Forest) is outlined below.

Algorithm 1. Spectral Triad feature selection

(1) **for** Number of covariates n **do**

(2) Choose the Root note Rn (from 1 to n of variables),

(3) Based on the Rn, select the third R_3 and fifth R_5 notes (i.e., variables) to form a triad T_1,

(4) Evaluate the relationship of T_1 to the response variable y using ReliefF (2.2.2.1) as an evaluation criterion OR Train the MLR algorithm (2.2.2.2) or Random Forest Model (2.4.1) using T_1 to T_n and the relevant response variable y,

(5) Rank and select the most significant (i.e., influential) variable R_i, i.e., one with the lowest p-value of F-Statistic or highest importance score,

(6) Repeat steps 3 to 5, until all triads T_n are evaluated,

(7) Return all significant or important variables from T_1 to T_n with their triad of origin T_0 and p-value or importance scores,

(8) End for

The STfs algorithm offers several advantages: (1) it provides an objective criterion for choosing the relevant variables, compared to arbitrarily chosen thresholds in filter-based approaches; (2) any evaluation criteria can be used to rank informative variables in each triad, thus allowing the user the flexibility and freedom of choice. For example, the users are not only limited to filter approaches but can employ more robust and explainable machine learning algorithms; and (3) only relevant (i.e., most informative) variables, i.e., in relation to one another and the response variable, are selected to form the final subset for prediction, thus simultaneously dealing with problems of collinearity and dimensionality.

2.2.2. Evaluation criteria

The STfs algorithm can accommodate various evaluation criteria which determine if the algorithm is used in a filter or wrapper mode. For example, Pearson's correlation coefficient r (or its non-parametric alternative, Spearman's rank coefficient ρ), Entropy-based algorithms (such as Information gain) and ReliefF can be used in the filter mode, while an explainable model such as Multiple Linear Regression (or its non-parametric alternative, Generalized Additive Models, GAM) and RF or eXtreme Gradient Boosting (XGBoost) can be used in the wrapper mode. To demonstrate the performance of STfs in filter mode, we used ReliefF algorithm, while in wrapper mode we used MLR and RF as evaluation criteria for each triad. Below is a short description of the evaluation criteria considered here.

2.2.2.1. ReliefF

ReliefF algorithm was developed to overcome the limitations of the original Relief algorithm (Robnik-Šikonja & Kononenko 2003), which was limited to classification problems and could not deal with missing values in the data. Relief and its adaptations aim to estimate the quality of variables according to how well they can distinguish close observations. To achieve this, the original Relief algorithm first randomly selects an observation R_i, then determines two nearest observations (i.e., neighbours), where one is from the same class as R_i, i.e., nearest hit H, and the other is from a different class, i.e., nearest miss M. Then, it estimates and updates the quality W[A] of all variables A depending on their values for R_i, H, and M. This process is repeated for m times, defined by the user. Instead of only two nearest neighbours (i.e., H and M), the extension of Relief, i.e., ReliefF, uses k nearest H and M and the final W[A] of A is determined by averaging the contribution of k. ReliefF was later adapted for regression problems (called RReliefF), where a probability that the predicted values of two observations are different is used instead of whether they belong to the same class for classification problems (Robnik-Šikonja & Kononenko 1997; Robnik-Šikonja & Kononenko

2003). Further details can be found in relevant publications (Kononenko, Robnik-Šikonja & Pompe 1996; Robnik-Šikonja & Kononenko 1997; Robnik-Šikonja & Kononenko 2003). RReliefF was performed in R-Statistic software using the "FSelectorRcpp" package which is a reimplementation of the "FSelector" package, but eliminates dependence on JAVA/WEKA and allows parallel processing and sparse matrix.

2.2.2.2. Multiple linear regression analysis

Multiple Linear Regression (MLR, Equation 1) determines the most influential predictor variables (x) according to the amount that each predictor variable (y) reduces the residual sum of squares.

$$y = m_1 x_1 + m_2 x_2 + \cdots + m_n x_n + b \tag{1}$$

Where y is the response variable (i.e., LAI, LC_{ab}, CCC), and $x_{1 \ldots n}$ are predictor variables, m_1 represent regression coefficients, while b represents the intercept. The regression coefficient, i.e., the partial derivative of the response variable for each predictor variable, measures the linear sensitivity of y to inputs $x_{i \ldots n}$ (Mohanty & Codell, 2002). The coefficient of determination adjusted for the degree of freedom (adjusted R^2), was used to determine the relationship between the response variable and predictor variables in each triad, while the F-statistic *p-value* was used to determine the significance of the relationships at $\alpha = 0.05$ and rank the most significant explanatory variable.

2.3. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a greedy optimisation wrapper-based algorithm, which sequentially removes the least important variables while retaining a few and an optimal subset of variables. Because it uses a machine learning regressor to evaluate each predictor variable for its predictive power in the modelling of a response variable, it falls under the category of wrapper-based variable selection approaches and can be combined with various MLRAs. In this way, it starts with a full set of variables to iteratively build regression models, while progressively eliminating the variables which do not improve the estimation errors of the model at each iteration until all features are exhausted. Consequently, RFE reduces dimensionality and eliminates dependencies between predictor variables.

The performance of each model is assessed using RMSE and the cross-validation resampling technique (Demarchi et al., 2020). Previous studies (Demarchi et al., 2020, Kganyago et al., 2017, Pullanagari et al., 2018) have combined the RFE algorithm with Support Vector Machine (SVM-RFE) and RF-RFE in classification and regression problems. In this study, we used the RF-RFE, where at each iteration, the algorithm eliminates the variables with the least importance on the RF models based on a 10-fold cross-validation strategy.

2.4. Modelling and prediction accuracy assessment

2.4.1. Random Forest

Utilising R-statistics software (Tierney, 2012) and the "Random Forest" package, the retrieval of BVs under consideration here, LAI, LCab, and CCC, was carried out (Liaw & Wiener, 2002). Popular tree-based MLRA called RF uses bagging to create several decision trees repeatedly and independently from a set of random training samples produced by bootstrap resampling from the original dataset (Fawagreh et al., 2014). Approximately 64% of the training data are kept for model building (i.e., in-bag samples), while the 36% remaining samples, i.e., out-of-bag (OOB) samples are used for model accuracy evaluation and for computing variable importance (Gislason et al., 2006). When permuting OOB samples while holding all the other factors constant, the %IncMSE is calculated to determine the importance of each explanatory variable. %IncMSE, a ranking metric for the various variables, is used to evaluate the predictive ability of the variables. For instance, a variable might be considered significant if its exclusion from the model significantly lowers prediction accuracy. In essence, a high importance score for an explanatory variable implies a strong association with the response variable or predictive power (Ramoelo et al., 2015). Using values from 100 to 500 with an interval of 10, and from 1 to p (i.e., number of explanatory variables) with a single interval, the tuning parameters, *ntree* (number of trees) and *mtry* (number of variables) at every split, were determined.

2.4.2. Prediction accuracy assessment

The prediction errors of each subset of features or the entire feature space were assessed with the Coefficient of determination (R^2), Root Mean Squared Error

(RMSE), Mean Absolute Error (MAE), and Relative Root Mean Squared Error (RRMSE). The R^2 is a correlation-based metric that determines how much of the variance of the response variable, such as LAI, is explained by the explanatory variables (e.g., spectral bands). The RMSE and MAE, on the other hand, measure the amount of error between the predictions and observations in the units of the BVs. The RRMSE is an index for comparing different variables or ranges, where the RRMSE <10% is considered Excellent and 10% to 20% is considered Good (Richter et al., 2012). Lastly, a propensity of a model to under- or overestimate a BV is measured by Percentage Bias (%Bias), where 0% is the optimum %Bias, and values near 0% denote a model that is correct (Gara et al., 2019). These accuracy metrics were computed using R-statistics version 4.1.2 (Tierney, 2012), packages "Metrics" (https://cran.r-project.org/web/packages/Metrics/index.html, accessed 18 July 2021) and "Fgmutils" (https://rdrr.io/cran/Fgmutils/, accessed 18 July 2021).

3. Results and Discussion

The launch of a quasi-hyperspectral new generation sensor, i.e., Sentinel-2, characterised by many broad- and narrow-band wavelengths, presents challenges of collinearity to the machine learning algorithms, making them suboptimal in terms of accuracies and computational times. Therefore, this calls for the optimisation of features for specific problems and applications through feature selection to ensure that the optimal subset (i.e., fewer bands that can achieve equivalent or better accuracy for the entire dataset) is used for prediction. In this study, the Spectral Triad feature selection (STfs) algorithm (based on music theory) is proposed, with two variants, i.e., filter-based STfs and wrapper-based STfs. Specifically, in filter-based mode, STfs was used with multiple linear regression and ReliefF algorithms, while in the wrapper-based mode, it was used with a RF algorithm. The results are presented in **Table 9-2** and **Figure 9-4**.

As shown in **Table 9-2** and **Figure 9-4**, the number and location of spectral bands selected by the proposed STfs algorithms varied by BVs, i.e., LAI, LC_{ab}, and CCC, and ranged from five to eight. This finding is consistent with studies by Verrelst, et al. (2016) and Verrelst, et al. (2012) that found that four to nine spectral bands (out of hundreds from hyperspectral sensors) were optimal for retrieving crop parameters. For LAI, STfs-ReliefF had the best performance with R^2_{cv} of 66% and RMSE*cv* of 0.53 $m^2 m^{-2}$ using only five variables, i.e., B3:560 nm, B5:705 nm, B6:740 nm, B8:842 nm, and B11:1 610 nm, respectively. In contrast, the STfs-MLR and STfs-RF had an equivalent performance with R^2_{cv} of 63% and RMSE*cv* of 0.55 $m^2 m^{-2}$. However, the former used more (i.e., seven) spectral bands, (B3:560 nm, B4:665 nm, B5:705 nm, B7:783 nm, B8:842 nm, B11:1 610 nm, and B12:2 190 nm), while the latter used only six spectral bands (i.e., B3:560 nm, B4:665 nm, B5:705 nm, B6:740 nm, B11:1 610 nm, and B12:2 190 nm). In contrast, STfs-MLR was optimal in the retrieval of both leaf-level and canopy-level chlorophyll content, i.e., LC_{ab} (R^2_{cv}: 74%, RMSE*cv*: 7.09 µg cm^{-2}) and CCC (R^2_{cv}: 77%, RMSE*cv*: 33.69 µg cm^{-2}), using only seven spectral bands (i.e., B2:490 nm, B3:560 nm, B4:665 nm, B5:705 nm, B6:740 nm, B8A:865 nm, and B11:1 610 nm and B3:560 nm, B4:665 nm, B5:705 nm, B6:740 nm, B7:783 nm, B8A:865 nm, and B11:1 610 nm). The results are consistent with Kganyago et al. (2021), who found the sparse Partial Least Squares (i.e., embedded feature selection algorithm) to yield better performance over Gradient Boosting Machines in retrieving LC_{ab} by using only seven variables. The spectral bands selected by RF-RFE are: B3:560 nm, B5:705 nm, B6:740 nm, B8A:865 nm, and B11:1 610 nm for LAI. While for LCC, the selected bands are: B4:665 nm, B6:740 nm, B7:783 nm, B11:1 610 nm, B12:2 190 nm. For CCC, the selected bands by RF-RFE are: B3:560 nm, B5:705 nm, B6:740 nm, B8A:865 nm, and B12:2 190 nm. Consistent with Kganyago et al. (2021), LAI and CCC had similar bands selected by STfs-Relief (except for B2 selected for CCC but not LAI) and RF-RFE (except for B12 selected for CCC but not LAI) (**Figure 9-4**). This indicates that similar plant traits influence the two crop parameters.

Table 9-2. The performance of optimal variables selected by Spectral Triad feature selection (STfs), Random Forest-Recursive Feature Elimination (RF-RFE) and comparison to entire feature space (i.e., no feature selection).

	Feature selection approaches	***n* Selected features**	**R^2_{cv}**	**RMSE*cv***	**MAE*cv***	**Training time (*s*)**
LAI	STfs-MLR	7	0.63	0.55	0.37	26.39
	STfs-ReliefF	**5**	**0.66**	**0.53**	**0.36**	24.25
	STfs-RF	6	0.63	0.55	0.37	25.67
	RFE-RF	**5**	**0.67**	**0.53**	**0.36**	23.33
	No feature selection	10	0.63	0.56	0.38	29.11

LC_{ab}	**STfs-MLR**	**7**	**0.74**	**7.09**	**5.45**	25.44
	STfs-ReliefF	6	0.73	7.23	5.63	24.09
	STfs-RF	8	0.74	7.13	5.54	26.43
	RFE-RF	**5**	**0.73**	**7.14**	**5.65**	22.63
	No feature selection	10	0.73	7.21	5.65	27.34
CCC	**STfs-MLR**	**7**	**0.77**	**33.69**	**24.01**	27.28
	STfs-ReliefF	7	0.76	34.45	24.26	26.80
	STfs-RF	7	0.74	35.13	25.50	25.89
	RFE-RF	**5**	**0.78**	**32.96**	**23.36**	23.83
	No feature selection	10	0.75	35.30	25.19	29.58

Figure 9-3 shows the Similarities and differences in selected features between STfs and RF-RFE. None of the evaluated feature selection algorithms selected B2:490 nm for LAI, while STfs-MLR and STfs-ReliefF selected it for LC_{ab}, and only STfs-ReliefF for CCC. Sentinel-2 B3:560 nm, B5:705 nm and B11:1 610 nm for LAI, B4:665 nm for LC_{ab}, and B5:705 nm and B6:740 nm for CCC were mutually selected by all algorithms. The similarity of selected bands across different crop parameters may be due to the known co-variation of plant traits within various spectral regions (Ollinger 2011). Generally, the selected bands by STfs were consistent with those selected by a well-established algorithm, i.e., RF-RFE. The finding indicates that the novel STfs can select highly informative variables and is comparable to well-established feature subset selection algorithms. Therefore, STfs is a promising technique critical for optimising the feature space of quasi-hyperspectral data, thus reducing the complexity of the models and improving processing speeds. The fact that the spectral bands chosen by STfs and RF-RFE were dispersed across the whole MSI spectral coverage in **Figure 9-4** is among the most important findings. It highlights the significance of the MSI for various land applications, including Precision Agriculture. Besides, studies reveal that the various plant traits impact the electromagnetic spectrum across the board, not just in certain places. For instance, although leaf pigments like chlorophyll are known to absorb heavily in the blue (400–500 nm) and red-edge (650–700 nm) regions, LAI affects the entire spectrum (Ollinger 2011; Verrelst et al. 2016).

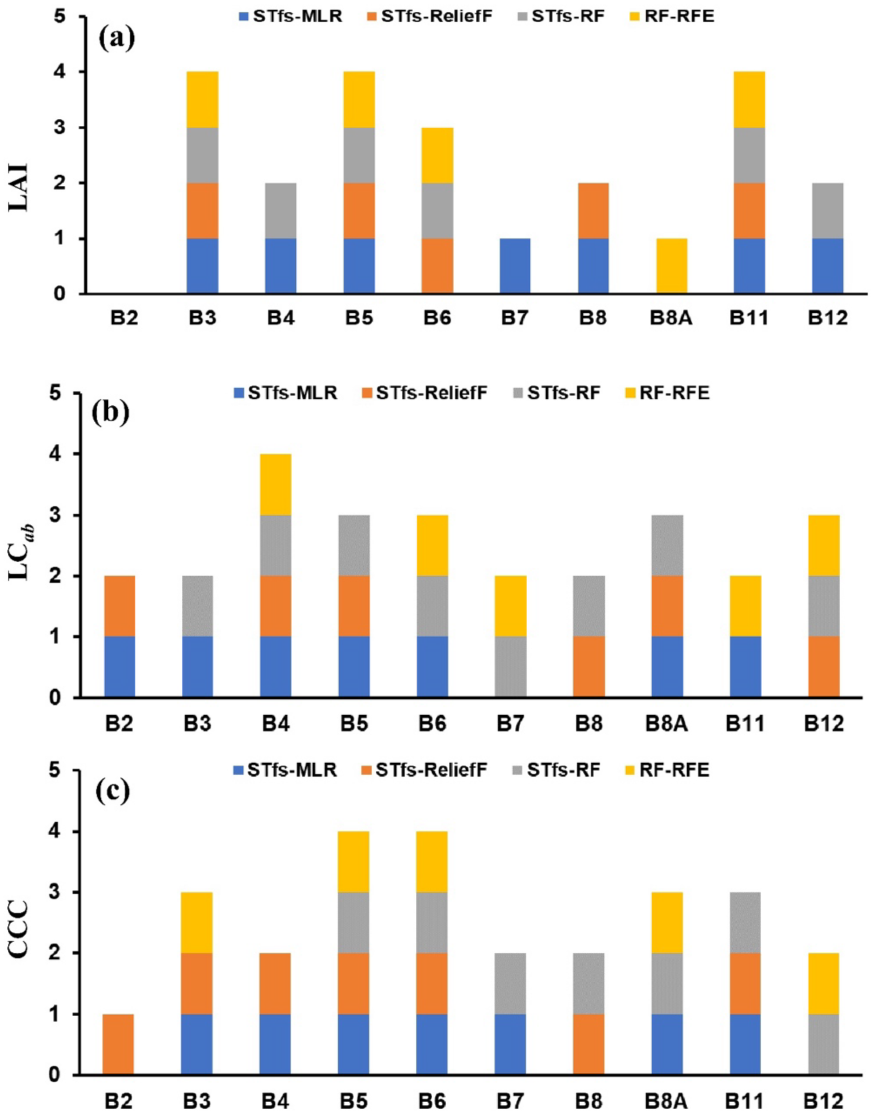


Figure 9-3. Similarities and differences in selected features between the proposed Spectral Triad feature selection (STfs) and Random Forest-Recursive Feature Elimination (RF-RFE) algorithms.

When benchmarked against the entire feature space (i.e., no feature selection), results show that the retrieval accuracy of crop BVs using optimised MSI feature space by STfs algorithms was consistently slightly higher R^2_{cv} by 3%, 1%, and 2% and RMSE*cv* by 0.03 $m^2\ m^{-2}$, 0.12 µg cm^{-2}, 1.61 µg cm^{-2}, for LAI, LC_{ab}, and CCC, respectively. Similarly, the comparison of optimised MSI feature space by STfs algorithms to a well-established wrapper-based feature selection algorithm, i.e., RF-RFE, also shows minute differences in accuracies with R^2_{cv} differences of only 1% across all considered crop BVs. In addition, depending on the crop factors, the RMSE*cv* differences were marginally more favourable to STfs or RF-RFE. Certainly, feature subset selection is beneficial, resulting in better accuracies and lower training times (**Table 9–2**). This observation is consistent with Mutanga et al. (2015) who found that three normalised ratios achieved better RMSE of calibration, (i.e., 0.175 vs 0.187) using 28 NDIs, while five RIs achieved 0.176 vs 0.188 using the entire dataset of 56 RIs. In estimating crude protein and metabolisable energy using airborne hyperspectral imaging, Pullanagari et al. (2018) found that RF-RFE outperformed the full spectrum, achieving R^2_{cv} of 80% vs 66%, and 78% vs 61%, respectively.

In the current study, RF-RFE was consistently faster, which can be attributed to a relatively small subset, i.e., five spectral bands. STfs resulted in better training times when compared to the entire feature space (i.e., No feature selection). The relatively small feature subsets obtained by RF-RFE and STfs are critical for reducing model complexity and the chances of overfitting from redundant features (Pullanagari, Kereszturi & Yule 2016). As shown in **Figure 9-4**, the distributions of predicted vs observed BVs show no apparent differences between entire feature space (i.e., No feature selection), RF-RFE and STfs, except for LAI, where some values around 2 m^2 m^{-2} were severely overestimated as >4 m^2 m^{-2} by the entire feature space (**Figure 9-4**c). These overestimations are also visible in some areas (see red box) on the spatial distribution maps (**Figure 9-5**g). The optimised feature spaces by STfs and RF-RFE in **Figure 9-4**a and c, respectively, were able to reduce these overestimations to below 4 m^2 m^{-2}, thus demonstrating the robustness of retrievals derived from optimal feature subsets. The optimal subsets by STfs and RF-RFE also show better characterisation of low LC_{ab} values (shown in cyan to blue in **Figure 9-5**) than in the entire feature space.

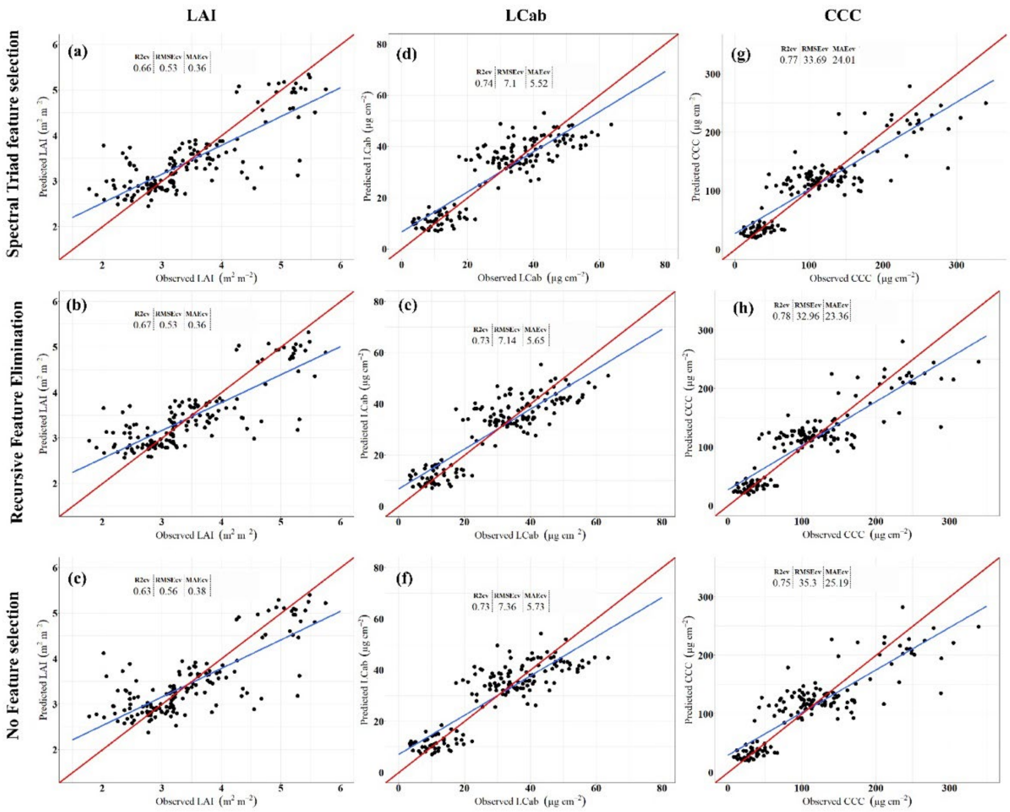


Figure 9-4. Comparison of the scatterplots of crop biophysical and biochemical variables derived from optimised Sentinel-2 MSI subsets selected by various feature selection techniques, i.e., Spectral Triad feature selection (a, d, g), random forest-recursive feature elimination (b, c, h), and no feature selection (c, f, i).

While STfs algorithms selected various spectral bands per crop biophysical and biochemical parameters, RF-RFE consistently selected five spectral bands. This finding can be attributed to the search criteria used by the proposed STfs algorithm versus the RFE algorithm. STfs uses spectral triads, i.e., combinations of three spectral bands at each iteration, and uses the evaluator (i.e., filter-band or machine learning algorithm) to estimate the relationship with the response variable, ranks, and selects a single most contributing spectral band in a triad. In contrast, RFE sequentially removes the least contributing bands in the regression models until there is no possible elimination without causing a loss in accuracy. The remaining spectral bands are considered the optimal subset. In STfs, spectral bands have many chances of being selected because the relatively low-contributing spectral bands in each triad may be a high contributor in another triad, while eliminated spectral bands in RFE are not considered again in the evaluation of the subset. Therefore, STfs is likely to retain a higher number of spectral bands than RFE, as observed in the current study. In highly dimensional data, the desired number of variables or subset size (*Pd*) can be specified by users, in which case, the algorithm will create and evaluate new triads until the *Pd* has been reached.

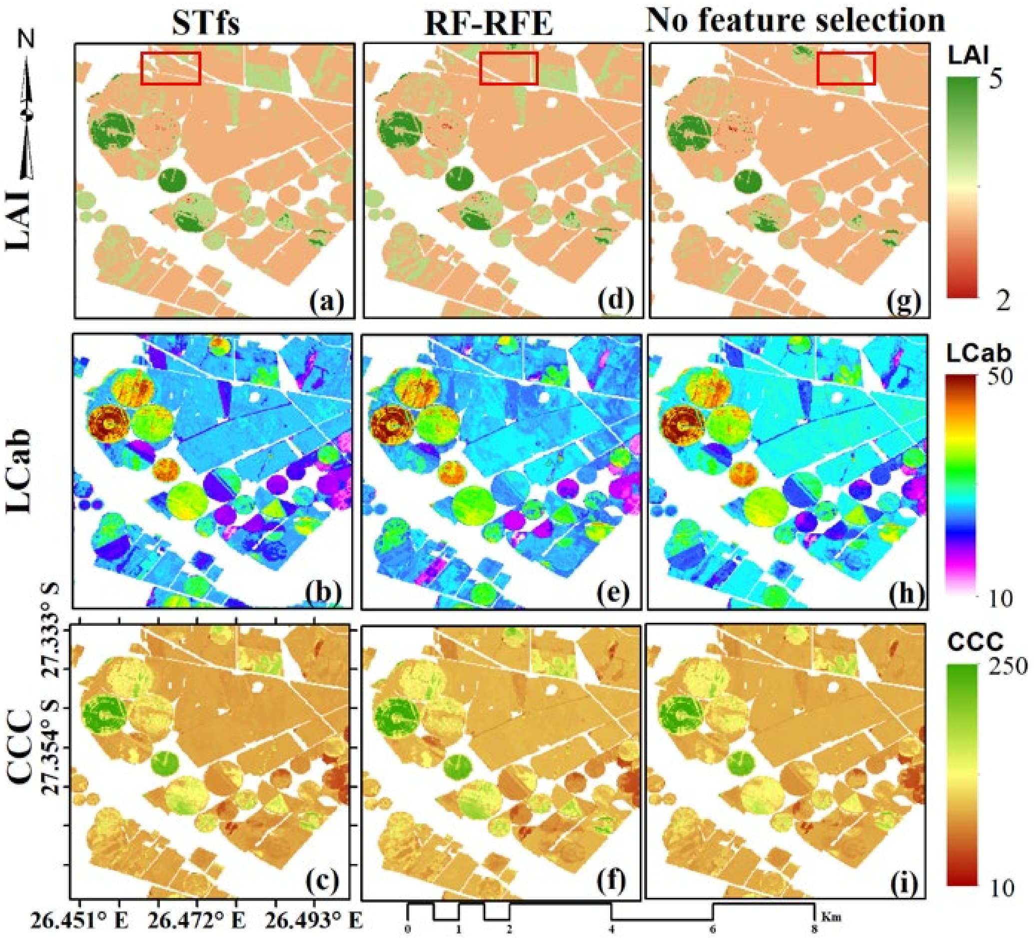


Figure 9-5. Maps of crop biophysical and biochemical variables retrieved using the optimised Sentinel-2 MSI subsets selected by the novel Spectral Triad feature selection (STfs), Random Forest-Recursive Feature Elimination, and entire feature space (i.e., no feature selection).

4. Conclusions

This study proposed the novel Spectral Triad feature selection (STfs) for optimising the spectral feature spaces of quasi-hyperspectral sensors, which are characterised by many typical broadband and strategically located narrow-band wavebands. The performance of the proposed feature selection technique was benchmarked against the well-established RFE coupled with a machine learning algorithm, i.e., RF, and the entire MSI spectral feature space, in retrieving essential biophysical and biochemical variables for agricultural crops. Overall, the results demonstrated that the MSI feature space could be optimised through feature selection, which results in slightly better or equivalent retrieval accuracies to the entire MSI feature space.

Moreover, the spectral bands selected by STfs were consistent with those selected by a well-established algorithm, i.e., RF-RFE. This study showed that the proposed STfs algorithm has great potential to optimise the spectral feature space of quasi-hyperspectral sensors for rapid crop biophysical and biochemical variables retrieval.

CHAPTER 10
SYNTHESIS AND CONCLUSIONS

1. Introduction

Global food security is threatened by a myriad of anthropogenic and natural phenomena. The situation is worse in developing African countries where more than 20% of the population is facing hunger. Meanwhile, there are many other linked challenges, such as high poverty and malnutrition levels, inequality, slow economic growth, and high unemployment rates. Agriculture is the backbone of several economies in southern Africa, contributing 35% towards their GDPs (Gross Domestic Product) and employing and sustaining about 70% to 80% of the population (Mango et al., 2017). Indeed, addressing food security challenges requires a concerted effort with the adoption of integrative agricultural practices and digital technologies to improve our resilience to extreme climate-related conditions (e.g., drought, heatwaves, pests, and diseases). Such practices will enable agricultural intensification and preservation of the agroecological resources (e.g., soils and water) to benefit current and future generations. One of the potentially viable solutions is Precision Agriculture (PA), which comprises various technologies and standards for field-scale management of spatiotemporal variability. It addresses the optimisation of an agricultural system to ensure efficiency and profitability while also considering the sustainability of the agricultural resources by reducing the environmental impact caused by the excessive application of inputs.

Since PA is concerned with managing spatiotemporal variability, technologies that can rapidly and systematically capture the status of agricultural commodities and provide accurate and actionable information to support remedial actions are in demand. However, in its traditional sense, e.g., automatically guided tractors, Variable Rate Application (VRA) vehicles, automatic weather stations, high data volumes, and other field instrumentation, compounded by low education levels among the farming community, poor rural connectivity, and small field sizes (typically around 0.5 Ha), the feasibility of PA (and cost vs benefit) in sub-Saharan African settings become questionable. Therefore, in developing regions, cheaper technologies with relatively low implementation costs that can generate reasonably accurate information of low complexity, are critical.

Remote sensing technology can remotely and repeatedly acquire data over vast areas at high to low spatial and temporal resolution. In particular, high-resolution remote sensing satellites such as Sentinel-2 are among the most promising sources of spatial variability information critical for adopting PA in developing regions because the data are relatively cheap or free of charge. Studies have demonstrated its capability to rapidly assess the spatial and temporal variability of crop development and health and produce spatially explicit information to support VRA technologies and site-specific farm management. A quest to improve remotely sensed key crop production indicators, such as cropland extent, crop yield and growth, and structural parameters (i.e., LAI, chlorophyll, and biomass), has been ongoing for decades. This has mainly been driven by developments in sensor technologies and often yielding new estimation techniques (see Chapter 2 for a detailed review). This thesis addressed major practical and technical challenges in the estimation of crop parameters from remotely sensed data, namely:

(1) Transferability of existing operational tools, such as SNAP Biophysical Processor, to semi-arid African agricultural landscapes;
(2) Radiometric errors under partly cloudy conditions, consistency of atmospheric correction techniques, their effect on the accuracy of crop parameters;
(3) Performance of contemporary machine learning algorithms, their spatial transferability, and sensitivity to various variable combinations in semi-arid areas; and
(4) Optimisation of Sentinel-2 feature space with a novel feature selection algorithm to improve prospects for accurate crop parameter retrieval.

This chapter synthesises the findings and makes conclusions.

2. Existing Models for Crop Biophysical and Biochemical Retrieval and Image Radiometric Quality

It is critical to assess the state-of-the-art operational BV retrieval models to determine their portability to new areas, such as African semi-arid agricultural landscapes, and to inform further development. In Chapter 3, the performance of a pre-trained Artificial Neural Networks (ANN) retrieval model (with simulated spectra from Radiative Transfer Model, i.e., PROSAIL) was assessed for its accuracy in the retrieval of LAI in a semi-arid African landscape using Bottom-of-Atmosphere (BOA) derived from Sen2Cor and Top-of-Atmosphere (TOA) reflectance and two Sentinel-2 resolutions, i.e., 10 m and 20 m. In essence, ambiguities about the impact of data processing level and spatial resolution on derived biophysical parameters were identified. This is a crucial undertaking since inter-comparison and consistency of biophysical parameters such as LAI are crucial for accurate and effective agricultural monitoring. Furthermore, users and experts keen to use the products and ensure ongoing development, respectively, can benefit from evaluating uncertainty in remotely sensed products. However, the validation of existing models and their comparability with well-established products has received little research attention (Bochenek et al., 2017, Campos-Taberner et al., 2018), and such studies are particularly scarce in sub-Saharan Africa.

The results showed that LAI retrieved using a pre-trained ANN model (also called SNAP Biophysical Processor) correlates moderately well with field-measured LAI in agricultural landscapes. However, for PA applications, the error rates are very significant, i.e., RMSE > 2 m^2 m^{-2}. Error assessment by crop type revealed that retrieved LAI values from SNAP were underestimated, and the accuracy varied by crop type. The accuracy of SNAP-derived LAI is influenced by the image resolution and processing levels. Nonetheless, on the one hand, the accuracy variation between the two Sentinel-2 resolutions is insignificant. On the other hand, across resolutions, there were noticeable differences between LAI obtained from BOA and TOA. These discrepancies can be ascribed to residual errors in some input bands following the atmospheric correction.

The LAI obtained from SNAP was discovered to be compatible with global LAI products, i.e., MODIS (MCD15A3H) and PROBA-V LAI. This indicates the potential of the product for extensive crop monitoring where within-field variation is not a concern. Based on the results, the following recommendations can be made:

- Owing to large errors when SNAP-derived LAI was validated against field-measured LAI data, it is not suitable for supporting site-specific management decisions.

- Therefore, there is a need to further improve biophysical retrieval models to achieve better accuracies and to support Precision Agriculture.
- Future studies should consider quantifying the impact of atmospheric correction techniques on crop parameter retrieval accuracy.

The quality of the satellite images is compromised by the atmospheric signal attenuation from aerosols (such as aerosol optical depth and atmospheric water content) and clouds. Many atmospheric correction (AC) techniques have been proposed since the availability of Sentinel-2 MSI data, capable of correcting atmospheric and adjacency effects and masking cloudy pixels. However, the effectiveness of various AC techniques for reducing atmospheric signal attenuation and improving image quality had not been extensively explored. Despite many validation and inter-comparison efforts (Sola et al., 2018), these techniques had not been validated in semi-arid African agricultural landscapes. Moreover, the effect of residual errors after atmospheric correction on biophysical and biochemical parameters relevant to Precision Agriculture was seldom been investigated.

Clouds essentially reduce the number of observations needed to enable quick agricultural management choices, making optical satellite images useless because they arrive with a greater than 5-day delay. This is typically the case during the rainy season, where some scenes may contain over 90% of cloud cover, thus entirely obscuring the landscape. Therefore, using clear-sky pixels in partly cloudy situations is a typical approach to maximise the number of usable observations for quantitative analysis. This is accomplished using algorithms for cloud screening and removal, followed by pixel compositing and gap filling (Frantz, et al., 2017; Zhu et al., 2018; Griffiths et al., 2019, Qiu et al., 2019). However, prior to this study, relatively little research on the radiometric reliability of those cloud-free pixels had been done. The erroneous expectation that the AC techniques would be equally effective in partly cloudy and cloud-free settings could lead to uneven radiometric measurements and inaccurate crop parameters, among other problems.

Chapter 4 sought to address the research gaps identified above and to implement research recommendations from Chapter 2 and Chapter 3 by assessing the effectiveness of the state-of-the-art AC techniques for Sentinel-2 data, i.e., Sen2Cor and *i*Cor, in deriving SR under partly cloudy conditions. This was done by comparing the derived SR with in situ surface reflectance, and further by evaluating the radiometric consistency of the SR products derived from the two AC approaches and the uncorrected TOA reflectance. Moreover, the impact of AC techniques on the retrieval of essential crop parameters was assessed. The findings can be summarised as follows:

- Both AC techniques had a poor fit with the reference SR in the VIS and SWIR bands, thus likely to negatively impact retrieval techniques that depend on these bands, such as assessment of structural parameters (i.e., LAI), canopy water and nitrogen content.
- The fit was relatively high in the red-edge and NIR bands in Sen2Cor and *i*Cor, thus indicating that these bands can be used to achieve reliable crop parameter estimates, especially in partly cloudy scenes, ~60%.
- Further validation in different cloud cover conditions is needed in future studies to understand the magnitude of the effects of clouds on SR derived from various AC techniques.
- Sen2Cor and *i*Cor demonstrated a nearly 1:1 resemblance to TOA data in most bands, suggesting that the AC approach had little impact and may be disregarded when obtaining crop parameters (such as LAI) directly from TOA.
- Crop parameters related to chlorophyll can benefit from AC.

3. Machine Learning Regression Algorithms and Their Transferability

Machine Learning Regression Algorithms (MLRAs) have seen increased interest in recent years, with the literature review (Chapter 2) indicating their wide adoption for retrieving crop parameters. It is appropriate to use MLRAs to retrieve critical agricultural crop parameters because of their high accuracy and versatility in processing various features and datasets. In contrast to other retrieval techniques such as spectral vegetation indices and RTM simulations, MLRAs may learn intricate correlations between measured crop attributes and reflectance data from field, aerial and satellite measurements, or RTM simulations. It is critical to determine optimal algorithms for specific situations (e.g., environmental conditions, datasets, plant physiological structures, and conditions) since there is no universal algorithm.

In Chapter 5, we investigated whether various cutting-edge MLRAs could enhance remotely sensed crop parameter retrieval, which would help inform agricultural management decisions and advance Precision Agriculture. This is especially the case in developing regions such as Africa where the operational use of Earth observation for Precision Agriculture is still in its infancy. Specifically, we examined the retrieval accuracy of various MLRAs, namely, sparse Partial Least Squares (sPLS), Random Forest (RF), and Gradient Boosting Machine (GBM) in retrieving LAI, LC_{ab}, and CCC using Sentinel-2 satellite data. Furthermore, we used built-in model agnostics (i.e., variable importance) to determine which spectral bands had the most significant impact on the accuracy of the model. Findings and recommendations are as follows:

- RF performed better than other methods in estimating all crop parameters.
- GBM was superior to sPLS in retrieving two crop parameters, i.e., LAI and CCC, but sPLS was relatively good at predicting LC_{ab}.
- All MLRAs produced acceptable accuracy according to GCOS/GMES, i.e., RRMSE of 10%, and they were on par with studies employing simulated and hyperspectral data as well as more sophisticated MLRAs like NN and GPR (Delloye et al., 2018; Fernandes et al., 2014; Jochem Verrelst et al., 2012; Jochem Verrelst et al., 2015).
- It is advised that fewer (i.e., four to eight spectral bands subsets) be tested in the future based on the superior prediction performance of sPLS over GBM in estimating LC_{ab} using only seven variables. This recommendation is consistent with other research (Jochem Verrelst et al., 2011; J. Verrelst et al., 2016), which discovered that four to nine bands were adequate to produce reliable estimations of crop parameters.
- In agreement with some previous studies (Jochem Verrelst et al., 2011; C. Wu et al., 2008), B5 centred at 705 nm was consistently superior at retrieving all the crop parameters.

- It is interesting to note that the results also indicate some degree of dependence between LAI and CCC, since the variables that had the largest impact on the performance of the model were comparable, namely B5, B11, B3, and B4. This result is in line with other studies that discovered that it is challenging to separate co-dependent relationships between structural and leaf characteristics, showing that these bands have ample information content and predictive ability (Delloye et al., 2018; Ollinger, 2011).
- By effectively utilising all Sentinel-2 MSI bands, RF gave the most reliable predictive accuracies for all the crop parameters, making it a strong candidate for operationalisation. This finding is crucial for the development of future biophysical and biochemical products utilising one algorithm, increasing the likelihood and efficiency of developing precise prescription maps for VRA technologies.

Based on the promising results from Chapter 5, it was critical to examine the spatial transferability of advanced MLRAs models such as eXtreme Gradient Boosting (XGBoost). This examination (i.e., the spatial transferability of retrieval techniques) is necessary to (1) lessen the requirements for substantial training datasets, (2) lower the computational costs associated with calibrating MLRAs, and (3) increase prediction accuracy in data-scarce areas by applying the information gained from data-rich areas. This is because MLRAs are sensitive to the representativeness of the experimental data (in terms of crop types and phenology) and other external conditions (such as atmospheric and sensor acquisition configurations). Chapter 6 presented a novel spatial transfer scheme, where some amount of the samples from the destination site can be added to account for new (i.e., unseen) variations and increase the spatial transferability. Specifically, we evaluated the impact of varying the amounts (i.e., 25%, 50%, and 75%) of available training samples in the destination (or Target) site $\mathcal{D}_T$, where the $\mathcal{D}_T$ consisted of Bothaville and Harrismith sites in the Free State province, South Africa. Additionally, it was assessed how accurately the trained (i.e., Source site $\mathcal{D}_S$) model influenced the retrieval accuracy in the destination (i.e., Target) site $\mathcal{D}_T$. The main results and conclusions of Chapter 6 are as follows:

- The transferability of LAI models was dependent, to some extent, on the characteristics of the Source site $\mathcal{D}_S$ (i.e., crop types and plant structural properties). For example, the multi-crop LAI models from Bothaville could be transferred with only 25% additional samples needed in the $\mathcal{D}_T$ to retrieve LAI.
- To obtain reliable retrievals of LAI, a model from Harrismith $\mathcal{D}_S$ had to first be re-trained using substantially more training data in the $\mathcal{D}_T$.
- The issues affecting the spatial transferability of LC_{ab} and CCC, respectively, are related to: their variability within and between fields and crop types; dynamic natural processes; and because different types of field equipment were utilised during the study (such as SPAD and MC-100). Nonetheless, it was discovered that the data range of the $\mathcal{D}_S$ models mattered as the transferability of the model was hindered by unobserved low or high crop parameter ranges.
- Additionally, when the trained and unseen data ranges diverge, the accuracy of $\mathcal{D}_S$ model influences transferability less.
- The findings of this study outperform those of earlier related studies, offering hope for reliable crop parameter retrievals without the need for many extra samples at the $\mathcal{D}_T$.

4. Optimising Sentinel-2 Feature Space for Improved Retrieval Accuracy

The launch of quasi-hyperspectral sensors such as Worldview-2 and MERIS attracted significant attention in the literature. Various researchers have been inquisitive about the contribution of the strategically positioned spectral bands towards the classification accuracy and uncertainty of retrieving various biophysical and biochemical parameters such as N content, Biomass, and LAI in various environments. From 2016, the availability of Sentinel-2 MSI —consisting of two spatial resolutions, each with a unique spectral configuration—led to further research questions such as whether it is sufficient to use VNIR-10 m bands or the

RE, NIR, and SWIR bands which have 20 m resolution in retrieving biochemical and biophysical parameters.

Until now, there has been a lack of consensus on the choice of spectral bands for retrieving biochemical and biophysical parameters. Where some studies utilise the entire feature space of Sentinel-2, others are in favour of using only a few necessary bands for obtaining high retrieval accuracies. Specific examples are Clevers et al. (2017) and Delegido et al. (2013), where the former revealed that SVIs based on 10 m bands, (i.e., VNIR), are superior for retrieving potato crop biochemical and biophysical parameters, while the latter demonstrated that excluding 20 m RE bands led to systematic errors in the retrieval of multi-crop parameters. However, SWIR bands are increasingly singled out as the important variables in retrieval models based on MLRAs (Chrysafis et al., 2020; Kganyago et al., 2021; Verrelst et al., 2015). To demystify these inconsistencies, Chapter 7 evaluated the individual predictive accuracy of Sentinel-2 spectral configurations, i.e., S2-10 m and S2-20 m, in retrieving important crop parameters for Precision Agriculture. The former is characterised by broad VNIR bands and the latter by RE, NIR, and SWIR bands. Three MLRAs were used, i.e., XGBoost, RF, and GPR, thus allowing us to determine the sensitivity of each MLRA to the two configurations.

The key findings from Chapter 7 are as follows:

- the S2-20 m is highly robust for retrieving LAI with low uncertainties, when compared to S2-10 configuration,
- the predictive accuracies of S2-10 m and S2-20 m, in retrieving LC_{ab} and CCC, depends on the site characteristics, i.e., crop types or phenology,
- as a result, any of the two configurations can be utilised independently because none of them significantly differ from one another and the benchmark dataset, i.e., all resampled bands,
- RF, XGBoost, and GPR perform different according to the various configurations and study sites. Therefore, various MLRAs should be explored prior to selecting the one that is best for the crop parameters and conditions of a certain site.

Apart from the lack of agreement regarding the two Sentinel-2 configuration in crop parameter retrieval, literature review revealed that radiometric measurements from satellites exhibit non-negligible effects from view and illumination geometry, termed Bidirectional Reflectance Distribution Function (BRDF). The BRDF effects may introduce ambiguity in retrieving biophysical and biochemical parameters, hence techniques for its removal are well documented in literature (Roy et al., 2017, Yin et al., 2020). At the same time, other studies (Kneubühler et al., 2008) show that geometry-induced spectral variation may provide additional structural information content, because of the dependence of reflected radiance on the geometric configurations of the sun and sensor. To support this claim, we can draw results from several studies that used multi-angular data (Verrelst et al., 2008, Duveiller et al., 2015, Vierling et al., 1997). These studies showed that multi-angular data were useful for characterising structural variation of many layers of vegetation that affect LAI.

Interestingly, the commonly used multi-spectral sensors such as Landsat and Sentinel-2 were shown to exhibit within-scene angular effects owing to their wide Field-of-View (FOV) (Nagol et al., 2015, Roy et al., 2017). For each of the 12 detectors of Sentinel-2, there is considerable pixel-wise surface anisotropy caused by varying within-scene (~295 km) view and illumination geometry (Gascon et al., 2017, Roy et al., 2017). Fortunately, Sentinel-2 data comes with pixel-wise and band-wise views and illumination geometry layers, hence providing some prospects for clarifying whether acquisition configurations provide additional information or noise to the estimation of crop parameters. To assess the contribution of per-pixel view and illumination geometry to the accuracy of retrieving LAI, LC_{ab} and CCC, in Chapter 8, we designed and evaluated the performance of five experimental scenarios where per-pixel geometric covariates, i.e., View and Sun Zenith Angles (VZA and SZA, respectively), and Relative Azimuth Angle (RAA), are excluded and then included in popular radiometric variables, i.e., spectral bands (*SB*) and spectral vegetation indices (*SVIs*), respectively. Moreover, using an explainable MLRA, i.e., RF, we were able to identify the most influential variables to the model accuracy, considering different experimental scenarios. Interpretable MLRAs such as RF chosen for this analysis are critical for revealing novel causal links and insights between a predictor and response and enabling the detection and diagnosis of biases in the models and associated input data (Azodi et al., 2020).

It should be noted that the Sentinel-2 Land biophysical processor (SL2P) (available in SNAP), uses Sentinel-2's geometric information as input to the biophysical retrieval process (Weiss and Baret, 2020). However, SL2P uses ANN for inversion, which is known for its high complexity, many confounding tuning parameters, and its opaqueness. Consequently, it does not provide model agnostics, which may reveal the influence of per-pixel geometric variables on the estimations. Our analysis in Chapter 8 sought to address the following research questions: (1) how much variability in crop canopy reflectance is explained by the per-pixel view and illumination geometry variations? (2) What is the contribution of incorporating per-pixel view and illumination geometry to the accuracy of retrieving crop BVs using MLRAs? (3) Are geometric variables influential in improving modelling accuracy when coupled with commonly used radiometric input variables, i.e., spectral bands and SVIs? To the best of our knowledge, no studies were found that addressed these pertinent questions.

The key findings from Chapter 8 are as follows:

- Per-pixel geometric covariates, i.e., SZA, VZA, and RAA, improve retrievals of canopy parameters, i.e., LAI and CCC, while improvement in LC_{ab} retrieval accuracy is negligible.
- Spectral variability exceeding 30% in all Sentinel-2 bands is explained by the pixel-wise view and illumination geometry. This finding is consistent with many studies (Ranson, 1986, Middleton, 1991, Pocewicz et al., 2007, Verrelst et al., 2008, Ma et al., 2020) that showed that view and illumination geometry have a considerable effect on radiometric and phenology metrics.
- Ignoring the variability caused by view and illumination geometry in wide FOV sensors may lead to incorrectly interpreted estimations. Such estimations have implications for inaccurate crop health diagnosis and loss of trust in remotely sensed biophysical and biochemical parameters for agronomic applications.

- VZA and SZA are among the most influential variables in all biophysical and biochemical parameter models; thus, they can be used as additional information to account for spectral variability resulting from sensor and sun geometries in retrieving crop parameters with MLRAs.

The results found in the experiments performed in Chapter 7 and Chapter 8 are critical for informing future retrieval of essential crop parameters from the two Sentinel-2 configurations and by incorporating other covariates such as geometric variables. However, the results also underscore that there may be multiple possible subsets for deriving biophysical and biochemical parameters from the evaluated combinations of covariates. This may, therefore, suggest that two or more variables may be redundant, thus introducing additional computational burden to an already costly process of tuning hyperparameters, and training and testing the MLRAs. As shown by previous studies, feature dimensionality reduction and subset selection are more beneficial, yielding high accuracies and low uncertainties in biophysical and biochemical variables retrieval. Moreover, one of the critical recommendations emanating from Chapter 8 is the optimisation of Sentinel-2 feature space using feature selection to avoid potential collinearity of predictors and improve the operationalisation prospects of retrieving crop parameters. However, in Chapter 6 it was deemed that to improve spatial transferability, predictors should be optimised for different crop parameters according to their relative contribution and quality.

Meanwhile, previous studies (Verrelst et al., 2016, Verrelst et al., 2011) found that only four to nine bands may be sufficient to achieve robust retrievals of BVs. While some studies used arbitrarily selected variables or were based on prior knowledge, we argue that the selection of covariates based on prior knowledge is beset with subjectivity and uncertainty. Our rationale for this argument is that many complex and interacting environmental and climatic factors affect biophysical and biochemical traits in various vegetation canopies. This is further complicated by the varying agricultural management practices, phenology, crop types, and limiting factors such as soil nutrients and water; this makes knowledge-based feature selection for every possible scenario erroneous. Statistical, data-based feature selection techniques are greatly needed based on their capability to associate each or a set of covariates to the crop parameter of interest and assign a rank (e.g., importance) based on linear and non-linear relationships among covariates (or between covariates and crop parameters). It is, therefore, critical to develop new approaches for optimising feature spaces of quasi-hyperspectral sensors.

Therefore, in Chapter 9, we proposed the novel Spectral Triad feature selection (STfs) algorithm inspired by music theory, and compared its consistency in crop parameter retrieval accuracy and selected variables to a well-established feature selection algorithm, i.e., RF-RFE and all covariates (thus with no feature selection).

The key findings are as follows:

- The MSI feature space can be optimised through feature selection, to a slightly better or equivalent retrieval accuracies to the entire MSI feature space.
- STfs is consistently slightly better than the entire feature space in the retrieval of LAI, LC_{ab}, and CCC.
- When compared to RF-RFE, optimised feature space by STfs shows minute differences in accuracies across all considered crop parameters.
- Selected features (i.e., bands) relatively similar between STfs and RF-RFE, but the latter selects a considerably lower number of features, i.e., five.

5. Conclusion

Achieving food security for all is among the world’s critical challenges today. This is especially the case amid frequent, prolonged, and severe climate-related disasters and socio-economic development pressures. Consequently, the UN SDGs set out to coordinate and track efforts by different countries. The SDG-2 aims to improve the resilience of agricultural systems to climate change and increase their productivity and sustainability. However, some of the efforts were deterred by the COVID-19 pandemic, necessitating increased adoption of technologies and innovative management practices such as Precision Agriculture. The latter promises higher productivity and returns without compromising environmental sustainability. Assessment of spatial variability at the field level by remote sensing of essential crop parameters offers prospects for timely detection and diagnosis of crop stress, optimises agronomic inputs and farm decisions, and ensures high yields. Good agricultural productivity would increase availability, affordability, and access to food, thus helping to eliminate hunger in >20% of the sub-Saharan population. The overall essence of the current study was to characterise the field variability in crop parameters relevant to Precision Agriculture using Sentinel-2 data and robust MLRAs.

In achieving this aim, existing operational retrieval tools were validated, and their errors (i.e., uncertainties) in a semi-arid landscape were quantified for the first time. The analysis is critical to inform customers and experts interested in the usage of the product and future development of crop parameters, respectively. Moreover, technical challenges related to atmospheric correction were addressed. Then, the study also compared the performance of many robust machine learning algorithms with various input variables, concluding that tree-based models (Random Forest and eXtreme Gradient Boosting) and kernel-based (Gaussian Process Regression) were optimal for specific crop parameters and were sensitive to various input variables.

The study established that Sentinel-2 VNIR bands at 10 m and RE-SWIR bands at 20 m can be used separately without data fusion or resampling processes, which may cause unnecessary delays to critical farm management decisions. Moreover, adding within-scene (i.e., per-pixel) geometric covariates to common radiometric variables (i.e., spectral bands and vegetation indices) improves the accuracy of retrieving essential crop parameters for Precision Agriculture. Furthermore, the study explored the spatial transferability of crop parameter retrieval models. It showed that the model transferability could be improved by adding a few samples from the target site to account for variability in the new (unseen) site. Lastly, a novel feature selection technique, i.e., Spectral Triad feature selection, was proposed to deal with possibly many optimal subsets within Sentinel-2 feature space for specific parameters. This optimised feature space by the new algorithm performed better than the entire Sentinel-2 MSI feature space and was consistent with the well-established algorithm, i.e., Recursive Feature Elimination. Our results were within the limits of below 10% accuracy defined by GMES.

Future works should:

- Perform further validations in different cloud cover conditions to clarify further the magnitude of the effects of clouds on SR derived from various AC techniques,
- Improve spatial transferability of generic and crop-specific biophysical and biochemical models by exploring hybrid techniques,
- Perform further tests in other areas to reaffirm the contribution of view and illumination geometry to the retrieval accuracy of crop parameters. Sensitivity tests of each crop parameter to the acquisition configurations may also be essential, especially for wide FOV sensors such as Sentinel-2, and
- Compare MLRAs to determine which is best for a given site conditions and crop parameters.

The contribution of this study is in improving the accuracy of crop parameters used for Precision Agriculture using a new generation space-bourne sensor, i.e., Sentinel-2 MSI, and robust machine learning algorithms. Specifically, the study addressed a range of technical challenges limiting the applicability of remote sensing for Precision Agriculture such as (1) unfavourable atmospheric conditions such as cloud cover and (2) radiometric accuracy and consistency effects on biophysical and biochemical parameters. Moreover, a range of machine learning algorithms was employed with a variation of Sentinel-2 variables to improve the capability of characterising field spatial variability of biophysical and biochemical parameters of agricultural crops.

REFERENCES

Abdel-Rahman, E.M., Mutanga, O., Odindi, J., Adam, E., Odindo, A. & Ismail, R. 2014. A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. *Computers and Electronics in Agriculture*. 106:11–19. DOI: 10.1016/j.compag .2014.05.001

Abdulridha, J., Ampatzidis, Y., Kakarla, S.C. & Roberts, P. 2019. Detection of target spot and bacterial spot diseases in tomato using UAV-based and benchtop-based hyperspectral imaging techniques. *Precision Agriculture*. (0123456789). DOI: 10.1007/s11119-019-09703-4

Adam, F., Mönks, M., Esch, T. & Datcu, M. 2018. Cloud Removal in High Resolution Multispectral Satellite Imagery: Comparing Three Approaches. DOI: 10.3390/ecrs-2-05166

Adelabu, S.; Mutanga, O.; Adam, E. Testing the reliability and stability of the internal accuracy assessment of random forest for classifying tree defoliation levels using different validation methods. *Geocarto International* 2015, 30, 810-821.

Adjorlolo, C., Cho, M.A., Mutanga, O. & Ismail, R. 2012. Optimizing spectral resolutions for the classification of C3 and C4 grass species, using wavelengths of known absorption features. *Journal of Applied Remote Sensing*. 6(1). DOI: 10.1117/1.jrs.6.063560

Alexandridis, T.K., Cherif, I., Kalogeropoulos, C., Monachou, S., Eskridge, K. & Silleos, N. 2013. Rapid error assessment for quantitative estimations from Landsat 7 gap-filled images. *Remote Sensing Letters*. 4(9):920–928. DOI: 10.1080/2150704X.2013.815380

Alexandridis, T.K., Ovakoglou, G. & Clevers, J.G.P.W. 2019. Relationship between MODIS EVI and LAI across time and space. *Geocarto International*. (January). DOI: 10.1080 /10106049.2019.1573928

Alexandridis, T.K., Ovakoglou, G. & Clevers, J.G.P.W. 2020. Relationship between MODIS EVI and LAI across time and space. *Geocarto International*. 35(13):1385–1399. DOI: 10.1080/10106049.2019.1573928

Alexandridis, T.K., Ovakoglou, G., Cherif, I., Giménez, M.G., Laneve, G., Kasampalis, D., Moshou, D., Kartsios, S., et al. 2020. Designing AfriCultuReS services to support food security in Africa. *Transactions in GIS*. 1–29. DOI: 10.1111/tgis.12684

Amin, E., Verrelst, J., Rivera-Caicedo, J.P., Pipia, L., Ruiz-Verdú, A. & Moreno, J. 2021. Prototyping Sentinel-2 green LAI and brown LAI products for cropland monitoring. *Remote Sensing of Environment*. 255(November 2020). DOI: 10.1016/j.rse.2020.112168

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A. & Friedl, L. 2017. Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-Spatial Information Science*. 20(2):77–96. DOI: 10.1080/10095020.2017.1333230

Apolo-Apolo, O.E., Pérez-Ruiz, M., Martínez-Guanter, J. & Egea, G. 2020. A mixed data-based deep neural network to estimate leaf area index in wheat breeding trials. *Agronomy*. 10(2). DOI: 10.3390/agronomy10020175

Atzberger, C. & Richter, K. 2012. Spatially constrained inversion of radiative transfer models for improved LAI mapping from future Sentinel-2 imagery. *Remote Sensing of Environment*. 120:208–218. DOI: 10.1016/j.rse.2011.10.035

Atzberger, C. 2004. Object-based retrieval of biophysical canopy variables using artificial neural nets and radiative transfer models. *Remote Sensing of Environment*. 93(1–2). DOI: 10.1016/j.rse.2004.06.016

Atzberger, C. 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. *Remote Sensing*. 5(2):949–981. DOI: 10.3390/rs5020949

Atzberger, C., Darvishzadeh, R., Immitzer, M., Schlerf, M., Skidmore, A. & le Maire, G. 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. DOI: 10.1016/j.jag.2015.01.009

Ayumi, V. 2017. Pose-based human action recognition with Extreme Gradient Boosting. *Proceedings - 14th IEEE Student Conference on Research and Development: Advancing Technology for Humanity, SCOReD 2016*. DOI: 10.1109/SCORED.2016 .7810099

Azodi, C.B., Tang, J. & Shiu, S.H. 2020. Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends in Genetics*. 36(6):442–455. DOI: 10.1016/j.tig .2020.03.005

Bahrami, H., Homayouni, S., Safari, A., Mirzaei, S., Mahdianpari, M. & Reisi-Gahrouei, O. 2021. Deep learning-based estimation of crop biophysical parameters using multi-source and multi-temporal remote sensing observations. *Agronomy*. 11(7). DOI: 10.3390/agronomy11071363

Bajwa, S.G., Mishra, A.R. & Norman, R.J. 2010. Canopy reflectance response to plant nitrogen accumulation in rice. *Precision Agriculture*. 11(5):488–506. DOI: 10.1007/s11119-009-9142-0

Barbosa, A., Trevisan, R., Hovakimyan, N. & Martin, N.F. 2020. Modeling yield response to crop management using convolutional neural networks. *Computers and Electronics in Agriculture*. 170(February):105197. DOI: 10.1016/j.compag.2019.105197

Baret, F., Houlès, V. & Guérif, M. 2007. Quantification of plant stress using remote sensing observations and crop models: The case of nitrogen management. *Journal of Experimental Botany*. 58(4). DOI: 10.1093/jxb/erl231

Baret, F., Weiss, M., Verger, A. & Smets, B. 2013. ATBD for LAI, FAPAR and FCOVER from PROBA-V Products at 300 m Resolution (GEOV3). 61

Battude, M., al Bitar, A., Morin, D., Cros, J., Huc, M., Sicre, C.M., le Dantec, V. & Demarez, V. 2016. Estimating maize biomass and yield over large areas using high spatial and temporal resolution Sentinel-2 like remote sensing data. *Remote Sensing of Environment*. 184:668–681

Bayat, B., van der Tol, C. & Verhoef, W. 2020. Retrieval of land surface properties from an annual time series of Landsat TOA radiances during a drought episode using coupled radiative transfer models. *Remote Sensing of Environment*. 238(March 2018):110917. DOI: 10.1016/j.rse.2018.09.030

Belgiu, M. & Csillik, O. 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sensing of Environment*. 204(January 2017):509–523. DOI: 10.1016/j.rse.2017.10.005

Bellvert, J., Mata, M., Vallverdú, X., Paris, C. & Marsal, J. 2020. Optimizing precision irrigation of a vineyard to improve water use efficiency and profitability by using a decision-oriented vine water consumption model. *Precision Agriculture*. (0123456789). DOI: 10.1007/s11119-020-09718-2

Beltran, J.C., Valdez, P. & Naval, P. 2019. Predicting Protein-Protein Interactions based on Biological Information using Extreme Gradient Boosting. *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2019*. DOI: 10.1109/CIBCB.2019.8791241

Bergstra, J. & Bengio, Y. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*. 13.

Berk, A.; Anderson, G.; Acharya, P.; Bernstein, L.; Muratov, L.; Lee, J.; Fox, M.; Adler-Golden, S. MODTRAN5: 2006 update, SPIE Conference Title: Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XII. In Proceedings of Proc. SPIE.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. & Jones, Z.M. 2016. Mlr: Machine learning in R. *Journal of Machine Learning Research*. 17.

Blackburn, G.A. 1998. Quantifying Chlorophylls and Caroteniods at Leaf and Canopy Scales. *Remote Sensing of Environment*. 66(3). DOI: 10.1016/s0034-4257(98)00059-5

Blackburn, G.A. 2007. Hyperspectral remote sensing of plant pigments. *Journal of Experimental Botany*. 58(4). DOI: 10.1093/jxb/erl123

Bochenek, Z., Dąbrowska-Zielińska, K., Gurdak, R., Niro, F., Bartold, M. & Grzybowski, P. 2017. Validation of the LAI biophysical product derived from Sentinel-2 and Proba-V images for winter wheat in western Poland. *Geoinformation Issues*. 9(1(9)):15–26. Available: http://www.igik.edu.pl/upload/File/wydawnictwa/GI9BochenekZ.pdf

Boegh, E., Houborg, R., Bienkowski, J., Braban, C.F., Dalgaard, T., van Dijk, N., Dragosits, U., Holmes, E., et al. 2013. Remote sensing of LAI, chlorophyll and leaf nitrogen pools of crop- and grasslands in five European landscapes. *Biogeosciences*. 10(10):6279–6307. DOI: 10.5194/bg-10-6279-2013

Boegh, E., Soegaard, H., Broge, N., Schelde, K., Thomsen, A., Hasager, C.B. & Jensen, N.O. 2002. Airborne multispectral data for quantifying leaf area index, nitrogen concentration, and photosynthetic efficiency in agriculture. *Remote Sensing of Environment*. 81(2–3):179–193. DOI: 10.1016/S0034-4257(01)00342-X

Bolton, D.K. & Friedl, M.A. 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*. 173. DOI: 10.1016/j.agrformet.2013.01.007

Boyer, C.N., Wade Brorsen, B., Solie, J.B. & Raun, W.R. 2011. Profitability of variable rate nitrogen application in wheat production. *Precision Agriculture*. 12(4):473–487. DOI: 10.1007/s11119-010-9190-5

Bréda, N.J.J. 2003. Ground-based measurements of leaf area index: A review of methods, instruments and current controversies. *Journal of Experimental Botany*. 54(392):2403–2417. DOI: 10.1093/jxb/erg263

Breiman, L. 2019. Random forests. *Random Forests*. 1–122. DOI: 10.1201/9780429469275-8

Brown, L.A., Fernandes, R., Djamai, N., Meier, C., Gobron, N., Morris, H., Canisius, F., Bai, G., et al. 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(February):71–87. DOI: 10.1016 /j.isprsjprs.2021.02.020

Brown, M.E., Pinzón, J.E., Didan, K., Morisette, J.T. & Tucker, C.J. 2006. Evaluation of the consistency of Long-term NDVI time series derived from AVHRR, SPOT-vegetation, SeaWiFS, MODIS, and landsat ETM+ sensors. *IEEE Transactions on Geoscience and Remote Sensing*. 44(7). DOI: 10.1109/TGRS.2005.860205

Bsaibes, A., Courault, D., Baret, F., Weiss, M., Olioso, A., Jacob, F., Hagolle, O., Marloie, O., et al. 2009. Albedo and LAI estimates from FORMOSAT-2 data for crop monitoring. *Remote Sensing of Environment*. 113(4):716–729. DOI: 10.1016/j.rse.2008.11.014

Cahalan, R.F., Oreopoulos, L., Wen, G., Marshak, A., Tsay, S.C. & DeFelice, T. 2001. Cloud characterization and clear-sky correction from Landsat-7. *Remote Sensing of Environment*. 78(1–2):83–98. DOI: 10.1016/S0034-4257(01)00251-6

Caicedo, J.P.R., Verrelst, J., Munoz-Mari, J., Moreno, J. & Camps-Valls, G. 2014. Toward a semiautomatic machine learning retrieval of biophysical parameters. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 7(4). DOI: 10.1109/JSTARS.2014.2298752

Camacho, F., Cernicharo, J., Lacaze, R., Baret, F. & Weiss, M. 2013. GEOV1: LAI, FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. Part 2: Validation and intercomparison with reference products. *Remote Sensing of Environment*. 137:310–329. DOI: 10.1016/j.rse.2013.02.030

Camacho, F., Fuster, B., Li, W., Weiss, M., Ganguly, S., Lacaze, R. & Baret, F. 2021. Crop specific algorithms trained over ground measurements provide the best performance for GAI and fAPAR estimates from Landsat-8 observations. *Remote Sensing of Environment*. 260. DOI: 10.1016/j.rse.2021.112453

Campos-Taberner, M., García-Haro, F.J., Busetto, L., Ranghetti, L., Martínez, B., Gilabert, M.A., Camps-Valls, G., Camacho, F., et al. 2018. A critical comparison of remote sensing Leaf Area Index estimates over rice-cultivated areas: From Sentinel-2 and Landsat-7/8 to MODIS, GEOV1 and EUMETSAT polar system. *Remote Sensing*. 10(5). DOI: 10.3390/rs10050763

Campos-Taberner, M., García-Haro, F.J., Camps-Valls, G., Grau-Muedra, G., Nutini, F., Crema, A. & Boschetti, M. 2016. Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring. *Remote Sensing of Environment*. 187:102–118. DOI: 10.1016/j.rse.2016.10.009

Campos-Taberner, M., Moreno-Martínez, Á., García-Haro, F.J., Camps-Valls, G., Robinson, N.P., Kattge, J. & Running, S.W. 2018. Global estimation of biophysical variables from Google Earth Engine platform. *Remote Sensing*. 10(8):1–17. DOI: 10.3390/rs10081167

Camps-Valls, G., Gómez-Chova, L., Muñoz-Marí, J., Vila-Francés, J., Amorós, J., del Valle-Tascon, S. & Calpe-Maravilla, J. 2009. Biophysical parameter estimation with adaptive Gaussian processes. In *International Geoscience and Remote Sensing Symposium (IGARSS)*. V. 4. DOI: 10.1109/IGARSS.2009.5417372

Camps-Valls, G., Sejdinovic, D., Runge, J. & Reichstein, M. 2019. A perspective on Gaussian processes for Earth observation. *National Science Review*. 6(4):616–618. DOI: 10.1093/nsr/nwz028

Charles, H., Godfray, H. & Garnett, T. 2014. Food security and sustainable intensification. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 369(1639). DOI: 10.1098/rstb.2012.0273

Chemura, A., Mutanga, O. & Odindi, J. 2017. Empirical Modeling of Leaf Chlorophyll Content in Coffee (*Coffea arabica*) Plantations with Sentinel-2 MSI Data: Effects of Spectral Settings, Spatial Resolution, and Crop Canopy Cover. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 10(12). DOI: 10.1109/JSTARS .2017.2750325

Chen, J.M., Pavlic, G., Brown, L., Cihlar, J., Leblanc, S.G., White, H.P., Hall, R.J., Peddle, D.R., et al. 2002. Derivation and validation of Canada-wide coarse-resolution leaf area index maps using high-resolution satellite imagery and ground measurements. *Remote Sensing of Environment*. 80(1):165–184. DOI: 10.1016/S0034-4257(01)00300-5

Chen, S., Wang, S., Shukla, M.K., Wu, D., Guo, X., Li, D. & Du, T. 2020. Delineation of management zones and optimization of irrigation scheduling to improve irrigation water productivity and revenue in a farmland of Northwest China. *Precision Agriculture*. 21(3):655–677. DOI: 10.1007/s11119-019-09688-0

Chen, T. & Guestrin, C. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 13-17-Augu: *The Journal of the Association of Physicians of India.* 785–794. DOI: 10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., et al. 2019. *Package "xgboost"*. https://xgboost.readthedocs.io/en/stable/index.html

Cho, M.A., Ramoelo, A. & Dziba, L. 2017. Response of land surface phenology to variation in tree cover during green-up and senescence periods in the semi-arid savanna of Southern Africa. *Remote Sensing*. 9(7). DOI: 10.3390/rs9070689

Chrysafis, I., Korakis, G., Kyriazopoulos, A.P. & Mallinis, G. 2020. Retrieval of leaf area index using sentinel-2 imagery in a mixed mediterranean forest area. *ISPRS International Journal of Geo-Information*. 9(11). DOI: 10.3390/ijgi9110622

Chun, H. & Keles, S. 2007. Sparse partial least squares for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society, Series B*. 72(1):3–25.

Chun, H. & Keleş, S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. Roy. Stat. Soc. Ser. B. (Stat. Method.) 2010, 72, 3-25.

Chung, D.; Chun, H.; Keles, S. An Introduction to the 'spls' Package. Version 1.0 ed.; CRAN: University of Wisconsin, Madison, WI 53706., 2012. https://cran.r-project.org/web/packages/spls/vignettes/spls-example.pdf, Accessed: 15 June 2021.

Ciganda, V., Gitelson, A. & Schepers, J. 2008. Vertical profile and temporal variation of chlorophyll in maize canopy: Quantitative "crop vigor" indicator by means of reflectance-based techniques. *Agronomy Journal*. 100(5). DOI: 10.2134/agronj 2007.0322

Cihlar, J., Ly, H., Li, Z., Chen, J., Pokrant, H. & Huang, F. 1997. Multitemporal, multichannel AVHRR data sets for land biosphere studies - Artifacts and corrections. *Remote Sensing of Environment*. 60(1):35–57. DOI: 10.1016/S0034-4257(96)00137-X

Clark, M.; Tilman, D. Comparative analysis of environmental impacts of agricultural production systems, agricultural input efficiency, and food choice. Environ. Res. Lett. 2017, 12, 064016. DOI: 10.1088/1748-9326/aa6cd5

Clark, R.N. & Roush, T.L. 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research*. 89(B7). DOI: 10.1029 /JB089iB07p06329

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.C., Skakun, S. v. & Justice, C. 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*. 219(August):145–161. DOI: 10.1016/j.rse.2018 .09.002

Claverie, M., Matthews, J.L., Vermote, E.F. & Justice, C.O. 2016. A 30+ year AVHRR LAI and FAPAR climate data record: Algorithm description and validation. *Remote Sensing*. 8(3). DOI: 10.3390/rs8030263

Clevers, J.G.P.W. & Gitelson, A.A. 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(1):344–351. DOI: 10.1016/j.jag.2012.10.008

Clevers, J.G.P.W., Kooistra, L. & Van den Brande, M.M.M. 2017. Using Sentinel-2 data for retrieving LAI and leaf and canopy chlorophyll content of a potato crop. *Remote Sensing*. 9(5):1–15. DOI: 10.3390/rs9050405

Collins, W. 1978. Remote Sensing of Crop Type and Maturity. *Photogrammetric Engineering and Remote Sensing*. 44(1): 43–55.

Colomina, I. & Molina, P. 2014. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*. 92:79–97. DOI: 10.1016/j.isprsjprs.2014.02.013

Combal, B., Baret, F., Weiss, M., Trubuil, A., Macé, D., Pragnère, A., Myneni, R., Knyazikhin, Y., et al. 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). DOI: 10.1016/S0034-4257(02)00035-4

Corti, M., Cavalli, D., Cabassi, G., Vigoni, A., Degano, L. & Marino Gallina, P. 2019. Application of a low-cost camera on a UAV to estimate maize nitrogen-related variables. *Precision Agriculture*. 20(4):675–696. DOI: 10.1007/s11119-018-9609-y

Croft, H., Arabian, J., Chen, J.M., Shang, J. & Liu, J. 2020. Mapping within-field leaf chlorophyll content in agricultural crops for nitrogen management using Landsat-8 imagery. *Precision Agriculture*. 21(0123456789):856–880. DOI: 10.1007/s11119-019-09698-y

CUI, B., ZHAO, Q. jun, HUANG, W. jiang, SONG, X. YU, YE, H. CHUN & ZHOU, X. FENG. 2019. Leaf chlorophyll content retrieval of wheat by simulated RapidEye, Sentinel-2 and EnMAP data. *Journal of Integrative Agriculture*. 18(6):1230–1245. DOI: 10.1016/S2095-3119(18)62093-3.

Curran, P.J. 2001. Imaging spectrometry for ecological applications. *ITC Journal*. 3(4). DOI: 10.1016/S0303-2434(01)85037-6

Darvishzadeh, R., Skidmore, A., Schlerf, M. & Atzberger, C. 2008. Inversion of a radiative transfer model for estimating vegetation LAI and chlorophyll in a heterogeneous grassland. *Remote Sensing of Environment*. 112(5). DOI: 10.1016/j.rse.2007.12.003

Dash, J., and P. J. Curran. 2005. “The MERIS Terrestrial Chlorophyll Index.” International Journal of Remote Sens-ing 25: 5403–5413.

Daughtry, C. S. T., C. L. Walthall, M. S. Kim, E. Brown, and J. E. McMurtrey. 2000. "Estimating Corn Leaf Chloro-phyll Concentration from Leaf and Canopy Reflectance." Remote Sensing Sensing of Environment 74: 229–239.

Davi, H., Barbaroux, C., Francois, C. & Dufrêne, E. 2009. The fundamental role of reserves and hydraulic constraints in predicting LAI and carbon allocation in forests. *Agricultural and Forest Meteorology*. 149(2). DOI: 10.1016/j.agrformet.2008.08.014

de Castro, A.I., Jurado-Expósito, M., Peña-Barragán, J.M. & López-Granados, F. 2012. Airborne multi-spectral imagery for mapping cruciferous weeds in cereal and legume crops. *Precision Agriculture*. 13(3):302–321. DOI: 10.1007/s11119-011-9247-0

De Keukelaere, L., Sterckx, S., Adriaensen, S., Knaeps, E., Reusen, I., Giardino, C., Bresciani, M., Hunter, P., et al. 2018. Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. *European Journal of Remote Sensing*. 51(1):525–542. DOI: 10.1080/22797254.2018.1457937

Delegido, J., Verrelst, J., Alonso, L. & Moreno, J. 2011. Evaluation of sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content. *Sensors*. 11(7):7063–7081. DOI: 10.3390/s110707063

Delegido, J., Verrelst, J., Meza, C.M., Rivera, J.P., Alonso, L. & Moreno, J. 2013. A red-edge spectral index for remote sensing estimation of green LAI over agroecosystems. *European Journal of Agronomy*. 46:42–52. DOI: 10.1016/j.eja.2012.12.001

Delegido, J., Verrelst, J., Rivera, J.P., Ruiz-Verdú, A. & Moreno, J. 2015. Brown and green LAI mapping through spectral indices. *International Journal of Applied Earth Observation and Geoinformation*. 35(PB):350–358. DOI: 10.1016/j.jag.2014.10.001

Delloye, C., Weiss, M. & Defourny, P. 2018. Retrieval of the canopy chlorophyll content from Sentinel-2 spectral bands to estimate nitrogen uptake in intensive winter wheat cropping systems. *Remote Sensing of Environment*. 216(July):245–261. DOI: 10.1016/j.rse.2018 .06.037

Dhakar, R., Sehgal, V.K., Chakraborty, D., Sahoo, R.N. & Mukherjee, J. 2019. Field scale wheat LAI retrieval from multispectral Sentinel 2A-MSI and LandSat 8-OLI imagery: effect of atmospheric correction, image resolutions and inversion techniques. *Geocarto International*. 0(0):1–21. DOI: 10.1080/10106049.2019.1687591

Dimitrov, P., Kamenova, I., Roumenina, E., Filchev, L., Ilieva, I., Jelev, G., Gikov, A., Banov, M., et al. 2019. Estimation of biophysical and biochemical variables of winter wheat through sentinel-2 vegetation indices. *Bulgarian Journal of Agricultural Science*. 25(5):819–832.

Djamai, N. & Fernandes, R. 2018. Comparison of SNAP-derived Sentinel-2A L2A product to ESA product over Europe. *Remote Sensing*. 10(6). DOI: 10.3390/rs10060926

Djamai, N., Fernandes, R., Weiss, M., McNairn, H. & Goïta, K. 2019. Validation of the Sentinel Simplified Level 2 Product Prototype Processor (SL2P) for mapping cropland biophysical variables using Sentinel-2/MSI and Landsat-8/OLI data. *Remote Sensing of Environment*. 225(April):416–430. DOI: 10.1016/j.rse.2019.03.020

Dong, T., Liu, J., Shang, J., Qian, B., Ma, B., Kovacs, J.M., Walters, D., Jiao, X., et al. 2019. Assessment of red-edge vegetation indices for crop leaf area index estimation. *Remote Sensing of Environment*. 222(May 2018):133–143. DOI: 10.1016/j.rse.2018.12.032

Dong, T., Shang, J., Chen, J.M., Liu, J., Qian, B., Ma, B., Morrison, M.J., Zhang, C., et al. 2019. Assessment of portable chlorophyll meters for measuring crop leaf chlorophyll concentration. *Remote Sensing*. 11(22). DOI: 10.3390/rs11222706

dos Santos, A.F., da Silva, R.P., Zerbato, C., de Menezes, P.C., Kazama, E.H., Paixão, C.S.S. & Voltarelli, M.A. 2019. Use of real-time extend GNSS for planting and inverting peanuts. *Precision Agriculture*. 20(4). DOI: 10.1007/s11119-018-9616-z

Doxani, G., Vermote, E., Roger, J.C., Gascon, F., Adriaensen, S., Frantz, D., Hagolle, O., Hollstein, A., et al. 2018. Atmospheric correction inter-comparison exercise. *Remote Sensing*. 10(2):1–18. DOI: 10.3390/rs10020352

Drusch, M., del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., et al. 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*. 120. DOI: 10.1016/j.rse.2011.11.026

Du, P., Samat, A., Waske, B., Liu, S. & Li, Z. 2015. Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS Journal of Photogrammetry and Remote Sensing*. 105:38–53.

Durbha, S.S., King, R.L. & Younan, N.H. 2007. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. *Remote Sensing of Environment*. 107(1–2):348–361. DOI: 10.1016/j.rse.2006.09.031

Duveiller, G., R. Lopez-Lozano, and A. Cescatti. 2015. "Exploiting the Multi-Angularity of the MODIS Temporal Signal to Identify Spatially Homogeneous Vegetation Cover: A Demonstration for Agricultural Monitoring Applications." Remote sensing of environment 166: 61–77. doi:10.1016/j. rse.2015.06.001

Erisman, J.W., Sutton, M.A., Galloway, J., Klimont, Z. & Winiwarter, W. 2008. How a century of ammonia synthesis changed the world. *Nature Geoscience*. 1(10). DOI: 10.1038/ngeo325

Estévez, J., Salinero-Delgado, M., Berger, K., Pipia, L., Rivera-Caicedo, J.P., Wocher, M., Reyes-Muñoz, P., Tagliabue, G., et al. 2022. Gaussian processes retrieval of crop traits in Google Earth Engine based on Sentinel-2 top-of-atmosphere data. *Remote Sensing of Environment*. 273. DOI: 10.1016/j.rse.2022.112958

Estévez, J., Vicent, J., Rivera-Caicedo, J.P., Morcillo-Pallarés, P., Vuolo, F., Sabater, N., Camps-Valls, G., Moreno, J., et al. 2020. Gaussian processes retrieval of LAI from Sentinel-2 top-of-atmosphere radiance data. *ISPRS Journal of Photogrammetry and Remote Sensing*. 167(June):289–304. DOI: 10.1016/j.isprsjprs.2020.07.004

Fang, H., Wei, S. & Liang, S. 2012. Validation of MODIS and CYCLOPES LAI products using global field measurement data. *Remote Sensing of Environment*. 119:43–54. DOI: 10.1016/j.rse.2011.12.006

Fang, H., Zhang, Y., Wei, S., Li, W., Ye, Y., Sun, T. & Liu, W. 2019. Validation of global moderate resolution leaf area index (LAI) products over croplands in northeastern Chinas. *Remote Sensing of Environment*. 233(August):111377. DOI: 10.1016/j.rse.2019.111377

Fawagreh, K.; Gaber, M.M.; Elyan, E. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering*: An Open Access Journal 2014, 2, 602-609.

Fensholt, R., Rasmussen, K., Nielsen, T.T. & Mbow, C. 2009. Evaluation of earth observation based long term vegetation trends - Intercomparing NDVI time series trend analysis consistency of Sahel from AVHRR GIMMS, Terra MODIS and SPOT VGT data. *Remote Sensing of Environment*. 113(9). DOI: 10.1016/j.rse.2009.04.004

Féret, J.B., Gitelson, A.A., Noble, S.D. & Jacquemoud, S. 2017. PROSPECT-D: Towards modeling leaf optical properties through a complete lifecycle. *Remote Sensing of Environment*. 193:204–215. DOI: 10.1016/j.rse.2017.03.004

Fernandes, R., Weiss, M., Camacho, F., Berthelot, B., Baret, F. & Duca, R. 2014. Development and assessment of leaf area index algorithms for the Sentinel-2 multispectral imager. *International Geoscience and Remote Sensing Symposium (IGARSS)*. 3922–3925. DOI: 10.1109/IGARSS.2014.6947342

Fitzgerald, G., Rodriguez, D. & O'Leary, G. 2010. Measuring and predicting canopy nitrogen nutrition in wheat using a spectral index-The canopy chlorophyll content index (CCCI). *Field Crops Research*. 116(3). DOI: 10.1016/j.fcr.2010.01.010

Frampton, W.J., Dash, J., Watmough, G. & Milton, E.J. 2013. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing*. 82:83–92. DOI: 10.1016/j.isprsjprs.2013.04.007

Frantz, D. 2019. FORCE-Landsat + Sentinel-2 analysis ready data and beyond. *Remote Sensing*. 11(9). DOI: 10.3390/rs11091124

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. 29(5). DOI: 10.1214/aos/1013203451

Fu, Y., Yang, G., Wang, J. & Feng, H. 2013. A comparative analysis of spectral vegetation indices to estimate crop leaf area index. *Intelligent Automation and Soft Computing*. 19(3):315–326. DOI: 10.1080/10798587.2013.824176

Galvao, L. S., J. R. dos Santos, D. A. Roberts, F. M. Breunig, M. Toomey, and Y. M. de Moura. 2011. "On Intra-Annual EVI Variability in the Dry Season of Tropical Forest: A Case Study with MODIS and Hyperspectral Data." Remote Sensing of Environment 115 (9): 2350–2359. doi:10.1016/j.rse.2011.04.035

Gao, F., Anderson, M.C., Zhang, X., Yang, Z., Alfieri, J.G., Kustas, W.P., Mueller, R., Johnson, D.M., et al. 2017. Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery. *Remote Sensing of Environment*. 188:9–25. DOI: 10.1016/j.rse.2016.11.004

Gao, F., Masek, J., Schwaller, M. & Hall, F. 2006. On the blending of the landsat and MODIS surface reflectance: Predicting daily landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*. 44(8). DOI: 10.1109/TGRS.2006.872081

Gara, T.W., Darvishzadeh, R., Skidmore, A.K., Wang, T. & Heurich, M. 2019. Accurate modelling of canopy traits from seasonal Sentinel-2 imagery based on the vertical distribution of leaf traits. *ISPRS Journal of Photogrammetry and Remote Sensing*. 157(April):108–123. DOI: 10.1016/j.isprsjprs.2019.09.005

García-Haro, F.J., Campos-Taberner, M., Muñoz-Marí, J., Laparra, V., Camacho, F., Sánchez-Zapero, J. & Camps-Valls, G. 2018. Derivation of global vegetation biophysical parameters from EUMETSAT Polar System. *ISPRS Journal of Photogrammetry and Remote Sensing*. 139:57–74. DOI: 10.1016/j.isprsjprs.2018.03.005

Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., Lonjou, V., Lafrance, B., et al. 2017. Copernicus Sentinel-2A calibration and products validation status. *Remote Sensing*. 9(6). DOI: 10.3390/rs9060584.

GCOS. (2011). Systematic Observation Requirements for Satellite-based Products for Climate. In Supplemental details to the satellite-based component of the Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC: 2011 update (2011 update ed., Vol. GCOS - No. 154, pp. 138). Geneva, Switzerland: World Meteorological Organization (WMO). https://library.wmo.int/index.php?lvl=notice_display&id=12907

Gislason, P.O., Benediktsson, J.A. & Sveinsson, J.R. 2006. Random forests for land cover classification. In *Pattern Recognition Letters*. V. 27. DOI: 10.1016/j.patrec.2005.08.011

Gitelson, A.A., Gritz, Y. & Merzlyak, M.N. 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):271–282. DOI: 10.1078/0176-1617-00887

Gitelson, A.A., Peng, Y. & Huemmrich, K.F. 2014. Relationship between fraction of radiation absorbed by photosynthesizing maize and soybean canopies and NDVI from remotely sensed data taken at close range and from MODIS 250 m resolution data. *Remote Sensing of Environment*. 147:108–120. DOI: 10.1016/j.rse.2014.02.014

Gitelson, A.A., Viña, A., Ciganda, V., Rundquist, D.C. & Arkebauer, T.J. 2005. Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters*. 32(8):1–4. DOI: 10.1029/2005GL022688

Giuliani, G., Chatenoux, B., de Bono, A., Rodila, D., Richard, J.P., Allenbach, K., Dao, H. & Peduzzi, P. 2017. Building an Earth Observations Data Cube: lessons learned from the

Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD). *Big Earth Data*. 1(1–2):100–117. DOI: 10.1080/20964471.2017.1398903

Golhani, K., Balasundram, S.K., Vadamalai, G. & Pradhan, B. 2018. A review of neural networks in plant disease detection using hyperspectral data Airborne Imaging Spectrometer for Applications. *Information Processing in Agriculture*. 5(3):354–371. DOI: 10.1016/j.inpa.2018.05.002

Gower, S.T., Kucharik, C.J. & Norman, J.M. 1999. Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial ecosystems. *Remote sensing of environment*. 70(1):29–51.

Gregorutti, B., Michel, B. & Saint-Pierre, P. 2017. Correlation and variable importance in random forests. *Statistics and Computing*. 27(3):659–678.

Griffiths, P., Nendel, C. & Hostert, P. 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. *Remote Sensing of Environment*. 220(November 2018):135–151. DOI: 10.1016/j.rse.2018.10.031

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z. & Yang, X. 2013. Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. *International Journal of Remote Sensing*. 34(14):5166–5186

Gupta, A., Gusain, K. & Popli, B. 2016. Verifying the value and veracity of extreme gradient boosted decision trees on a variety of datasets. In *11th International Conference on Industrial and Information Systems, ICIIS 2016 - Conference Proceedings*. V. 2018-January. DOI: 10.1109/ICIINFS.2016.8262984

Guyot, G., and F. Baret. 1988. Utilisation de la haute re´solution spectrale pour suivre l'e´tat des couverts ve´ge´taux [Utilisation of high spectral resolution for monitoring vege- tation condition]. In T.-D.-G. enne and J. J. Hunt (eds.), Proceedings of 4th international colloquim spectral signa- tures of objects in remote sensing, proceedings of the con- ference held 18–22 January 1988 in Aussois (Modane), France (pp. 279–286). ESA SP-287. European Space Agency

Haboudane, D., Tremblay, N., Miller, J.R. & Vigneault, P. 2008. Remote estimation of crop chlorophyll content using spectral indices derived from hyperspectral data. In *IEEE Transactions on Geoscience and Remote Sensing*. V. 46. DOI: 10.1109/TGRS .2007.904836

Hagolle, O., Huc, M., Pascual, D.V. & Dedieu, G. 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. *Remote Sensing of Environment*. 114(8). DOI: 10.1016/j.rse.2010.03.002

Hagolle, O., Huc, M., Pascual, D.V. & Dedieu, G. 2015. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, LandSat, VENμS and Sentinel-2 images. *Remote Sensing*. 7(3):2668–2691. DOI: 10.3390/rs70302668

Herrmann, I., Karnieli, A., Bonfil, D. J., Cohen Y., & Alchanatis, V. (2010) SWIR-based spectral indices for assessing nitrogen content in potato fields, International Journal of Remote Sensing, 31:19, 5127-5143, DOI: 10.1080/01431160903283892

Herrmann, I., Karnieli, A., Bonfil, D.J., Cohen, Y. & Alchanatis, V. 2010. SWIR-based spectral indices for assessing nitrogen content in potato fields. *International Journal of Remote Sensing*. 31(19). DOI: 10.1080/01431160903283892

Herrmann, I., Pimstein, A., Karnieli, A., Cohen, Y., Alchanatis, V. & Bonfil, D.J. 2011. LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands. *Remote Sensing of Environment*. 115(8):2141–2151. DOI: 10.1016/j.rse.2011.04.018

Hilker, T., Coops, N.C., Coggins, S.B., Wulder, M.A., Brown, M., Black, T.A., Nesic, Z. & Lessard, D. 2009. Detection of foliage conditions and disturbance from multi-angular high spectral resolution remote sensing. *Remote Sensing of Environment*. 113(2):421–434. DOI: 10.1016/j.rse.2008.10.003

Hilker, T., Lyapustin, A., Hall, F.G., Wang, Y., Coops, N.C., Drolet, G. & Black, T.A. 2009. An assessment of photosynthetic light use efficiency from space: Modeling the atmospheric and directional impacts on PRI reflectance. *Remote Sensing of Environment*. 113(11):2463–2475. DOI: 10.1016/j.rse.2009.07.012

Hively, W.D., Shermeyer, J., Lamb, B.T., Daughtry, C.T., Quemada, M. & Keppler, J. 2019. Mapping crop residue by combining landsat and worldview-3 satellite imagery. *Remote Sensing*. 11(16). DOI: 10.3390/rs11161857

Houborg, R. & Boegh, E. 2008. Mapping leaf chlorophyll and leaf area index using inverse and forward canopy reflectance modeling and SPOT reflectance data. *Remote Sensing of Environment*. 112(1):186–202. DOI: 10.1016/j.rse.2007.04.012

Houborg, R. & McCabe, M.F. 2018. A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning. *ISPRS Journal of Photogrammetry and Remote Sensing*. 135:173–188. DOI: 10.1016/j.isprsjprs.2017.10.004

Houborg, R., Fisher, J.B. & Skidmore, A.K. 2015a. Advances in remote sensing of vegetation function and traits. *International Journal of Applied Earth Observation and Geoinformation*. 43:1–6. DOI: 10.1016/j.jag.2015.06.001

Houborg, R., McCabe, M., Cescatti, A., Gao, F., Schull, M. & Gitelson, A. 2015b. Joint leaf chlorophyll content and leaf area index retrieval from Landsat data using a regularized model inversion system (REGFLEC). *Remote Sensing of Environment*. 159:203–221. DOI: 10.1016/j.rse.2014.12.008

Huang, S., Miao, Y., Yuan, F., Gnyp, M.L., Yao, Y., Cao, Q., Wang, H., Lenz-Wiedemann, V.I.S., et al. 2017. Potential of RapidEye and WorldView-2 satellite data for improving rice nitrogen status monitoring at different growth stages. *Remote Sensing*. 9(3). DOI: 10.3390/rs9030227

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X. & Ferreira, L.G. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*. 83(1–2). DOI: 10.1016/S0034-4257(02)00096-2

Huete, A., Didan, K., Van Leeuwen, W., Miura, T. & Glenn, E. 2010. MODIS vegetation indices. In *Land remote sensing and global environmental change*. Springer. 579–602

Hughes, G. On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf Theory 1968, 14, 55-63.

Ilori, C.O., Pahlevan, N. & Knudby, A. 2019. Analyzing performances of different atmospheric correction techniques for Landsat 8: Application for coastal remote sensing. *Remote Sensing*. 11(4):1–20. DOI: 10.3390/rs11040469

Jacquemoud, S., Baret, F., Andrieu, B., Danson, F.M. & Jaggard, K. 1995. Extraction of vegetation biophysical parameters by inversion of the PROSPECT + SAIL models on sugar beet canopy reflectance data. Application to TM and AVIRIS sensors. *Remote Sensing of Environment*. 52(3). DOI: 10.1016/0034-4257(95)00018-V

Jacquemoud, S., Verhoef, W., Baret, F., Bacour, C., Zarco-Tejada, P.J., Asner, G.P., François, C. & Ustin, S.L. 2009. PROSPECT + SAIL models: A review of use for vegetation characterization. *Remote Sensing of Environment*. 113(SUPPL. 1). DOI: 10.1016/j.rse.2008.01.026.

Jay, S., F. Maupas, R. Bendoula, and N. Gorretta. 2017. “Retrieving LAI, Chlorophyll and Nitrogen Contents in Sugar Beet Crops from Multi-Angular Optical Remote Sensing: Comparison of Vegetation Indices and PROSAIL Inversion for Field Phenotyping.” Field Crops Research 210: 33–46. doi:10.1016/j.fcr.2017.05.005

Jensen, J.R. 1983. Biophysical Remote Sensing. *Annals of the Association of American Geographers*. 73(1):111–132. DOI: 10.1111/j.1467-8306.1983.tb01399.x

Jensen, R., Mausel, P., Dias, N., Gonser, R., Yang, C., Everitt, J. & Fletcher, R. 2007. Spectral analysis of coastal vegetation and land cover using AISA + hyperspectral data. *Geocarto International*. 22(1). DOI: 10.1080/10106040701204354

Jia, F., Liu, G., Liu, D., Zhang, Y., Fan, W. & Xing, X. 2013. Comparison of different methods for estimating nitrogen concentration in flue-cured tobacco leaves based on hyperspectral reflectance. *Field Crops Research*. 150:108–114. DOI: 10.1016/j.fcr .2013.06.009

Jiang, J., Weiss, M., Liu, F. & Baret, S. 2019. The impact of canopy structure assumption on the retrieval of GAI and Leaf Chlorophyll Content for wheat and maize crops. In *International Geoscience and Remote Sensing Symposium (IGARSS)*. DOI: 10.1109/IGARSS.2019.8899064

Jiang, Z., Huete, A., Didan, K. & Miura, T. 2008. Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*. 112(10):3833–3845. DOI: 10.1016/j.rse.2008.06.006

Johansen, K., Duan, Q., Tu, Y.H., Searle, C., Wu, D., Phinn, S., Robson, A. & McCabe, M.F. 2020. Mapping the condition of macadamia tree crops using multi-spectral UAV and WorldView-3 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*. 165(April):28–40. DOI: 10.1016/j.isprsjprs.2020.04.017

Jolliffe, I.T.; Trendafilov, N.T.; Uddin, M. A modified principal component technique based on the LASSO. *Journal of computational and Graphical Statistics* 2003, 12, 531-547.

Julien, Y. & Sobrino, J.A. 2019. Optimizing and comparing gap-filling techniques using simulated NDVI time series from remotely sensed global data. *International Journal of Applied Earth Observation and Geoinformation*. 76. DOI: 10.1016/j.jag.2018.11.008

Juutinen, S., Virtanen, T., Kondratyev, V., Laurila, T., Linkosalmi, M., Mikola, J., Nyman, J., Räsänen, A., et al. 2017. Spatial variation and seasonal dynamics of leaf-Area index in the arctic tundra-implications for linking ground observations and satellite images. *Environmental Research Letters*. 12(9). DOI: 10.1088/1748-9326/aa7f85

Kganyago, M. & Mhangara, P. 2019. The role of African emerging space agencies in earth observation capacity building for facilitating the implementation and monitoring of the African development agenda: The case of African earth observation program. *ISPRS International Journal of Geo-Information*. 8(7):1–23. DOI: 10.3390/ijgi8070292

Kganyago, M. 2021. Using sentinel-2 observations to assess the consequences of the COVID-19 lockdown on winter cropping in Bothaville and Harrismith , South Africa. *Remote Sensing Letters*. 12(9):827–837. DOI: 10.1080/2150704X.2021.1942582

Kganyago, M., Adjorlolo, C. & Mhangara, P. 2022. Exploring Transferable Techniques to Retrieve Crop Biophysical and Biochemical Variables Using Sentinel-2 Data. *Remote Sensing*. 14(16). DOI: 10.3390/rs14163968

Kganyago, M., Mhangara, P. & Adjorlolo, C. 2021. Estimating Crop Biophysical Parameters Using Machine Learning Algorithms and Sentinel-2 Imagery. *Remote Sensing*. 13(21):1–21.

Kganyago, M., Mhangara, P., Alexandridis, T., Laneve, G., Ovakoglou, G. & Mashiyi, N. 2020a. 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):883–892. DOI: 10.1080/2150704X.2020.1767823

Kganyago, M., Odindi, J., Adjorlolo, C. & Mhangara, P. 2017. Selecting a subset of spectral bands for mapping invasive alien plants: A case of discriminating parthenium hysterophorus using field spectroscopy data. *International Journal of Remote Sensing*. 38(20):5608–5625. DOI: 10.1080/01431161.2017.1343510

Kganyago, M., Odindi, J., Adjorlolo, C. & Mhangara, P. 2018. Evaluating the capability of Landsat 8 OLI and SPOT 6 for discriminating invasive alien species in the African Savanna landscape. *International Journal of Applied Earth Observation and Geoinformation*. 67(December 2017):10–19. DOI: 10.1016/j.jag.2017.12.008

Kganyago, M., Ovakoglou, G., Mhangara, P., Odindi, J., Adjorlolo, C. & Mashiyi, N. 2020b. Validation of atmospheric correction approaches for Sentinel-2 under partly-cloudy conditions in an African agricultural landscape. In *SPIE Proc*. DOI: 10.1117/12 .2572293

Kneubühler, M., B. Koetz, S. Huber, N. E. Zimmermann, and M. E. Schaepman. 2008. "Space-Based Spectrodirectional Measurements for the Improved Estimation of Ecosystem Variables." Canadian Journal of Remote Sensing 34: 192–205

Knyazikhin, Y., Martonchik, J. v., Myneni, R.B., Diner, D.J. & Running, S.W. 1998. Synergistic algorithm for estimating vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from MODIS and MISR data. *Journal of Geophysical Research Atmospheres*. 103(D24). DOI: 10.1029/98JD02462

Kobayashi, N., Tani, H., Wang, X. & Sonobe, R. 2020. Crop classification using spectral indices derived from Sentinel-2A imagery. *Journal of Information and Telecommunication*. 4(1):67–90. DOI: 10.1080/24751839.2019.1694765

Kononenko, I., Robnik-Šikonja, M. & Pompe, U. 1996. ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. *Artificial intelligence: Methodology, Systems, Applications*. 1–15. Available: http://lkm.fri.uni-lj.si/rmarko/papers/kononenko96-aimsa.pdf

Konstantinov, A. v. & Utkin, L. v. 2021. Interpretable machine learning with an ensemble of gradient boosting machines. *Knowledge-Based Systems*. 222:106993. DOI: 10.1016/j.knosys.2021.106993

Lanaras, C., Bioucas-Dias, J., Baltsavias, E. & Schindler, K. 2017. Super-Resolution of Multispectral Multiresolution Images from a Single Sensor. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. V. 2017-July. DOI: 10.1109/CVPRW.2017.194

Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E. & Schindler, K. 2018. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*. 146. DOI: 10.1016/j.isprsjprs.2018 .09.018

Lawley, V., Lewis, M., Clarke, K. & Ostendorf, B. 2016. Site-based and remote sensing methods for monitoring indicators of vegetation condition: An Australian review. *Ecological Indicators*. 60:1273–1283. DOI: 10.1016/j.ecolind.2015.03.021

Lê Cao, K.A., Rossouw, D., Robert-Granié, C. & Besse, P. 2008. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*. 7(1). DOI: 10.2202/1544-6115.1390

Lee, Y.J., Yang, C.M., Chang, K.W. & Shen, Y. 2008. A simple spectral index using reflectance of 735 nm to assess nitrogen status of rice canopy. *Agronomy Journal*. 100(1). DOI: 10.2134/agronj2007.0018

LI, H., CHEN, Z. xin, JIANG, Z. WEI, WU, W. BIN, REN, J. QIANG, LIU, B. & TUYA, H. 2017. Comparative analysis of GF-1, HJ-1, and Landsat-8 data for estimating the leaf area index of winter wheat. *Journal of Integrative Agriculture*. 16(2):266–285. DOI: 10.1016/S2095-3119(15)61293-X

Li, X., Zhang, L., Du, B., Zhang, L. & Shi, Q. 2017. Iterative Reweighting Heterogeneous Transfer Learning Framework for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 10(5). DOI: 10.1109/JSTARS.2016.2646138

Li, X., Zhang, Y., Bao, Y., Luo, J., Jin, X., Xu, X., Song, X. & Yang, G. 2014. Exploring the best hyperspectral features for LAI estimation using partial least squares regression. *Remote Sensing*. 6(7):6221–6241. DOI: 10.3390/rs6076221

Li, Y., Chen, J., Ma, Q., Zhang, H.K. & Liu, J. 2018. Evaluation of Sentinel-2A Surface Reflectance Derived Using Sen2Cor in North America. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 11(6). DOI: 10.1109/JSTARS.2018.2835823.

LI, Z. WANG, XIN, X. PING, TANG, H., YANG, F., CHEN, B. RUI & ZHANG, B. HUI. 2017. Estimating grassland LAI using the Random Forests approach and Landsat imagery in the meadow steppe of Hulunber, China. *Journal of Integrative Agriculture*. 16(2):286–297. DOI: 10.1016/S2095-3119(15)61303-X

Liang, L., Di, L., Zhang, L., Deng, M., Qin, Z., Zhao, S. & Lin, H. 2015. Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sensing of Environment*. 165:123–134. DOI: 10.1016/j.rse.2015.04.032

Liang, L., Yang, M., Deng, K., Zhang, L., Lin, H. & Liu, Z. 2011. A new hyperspectral index for the estimation of Nitrogen contents of wheat canopy. *Shengtai Xuebao/ Acta Ecologica Sinica*. 31(21):6594–6605

Liao, C., Wang, J., Dong, T., Shang, J., Liu, J. & Song, Y. 2019. Using spatio-temporal fusion of Landsat-8 and MODIS data to derive phenology, biomass and yield estimates for corn and soybean. *Science of the Total Environment*. 650:1707–1721. DOI: 10.1016/j.scitotenv.2018.09.308

Liaw, A.; Wiener, M. Classification and regression by randomForest. R news 2002, 2, 18-22.

Liu, Y., Liu, R. & Chen, J.M. 2012. DOI: 10.1029/2012JG002084.

Liu, Z. & Jin, G. 2017. Importance of woody materials for seasonal variation in leaf area index from optical methods in a deciduous needleleaf forest. *Scandinavian Journal of Forest Research*. 32(8):726–736. DOI: 10.1080/02827581.2016.1272713

Loggenberg, K., Strever, A., Greyling, B. & Poona, N. 2018. Modelling water stress in a Shiraz vineyard using hyperspectral imaging and machine learning. *Remote Sensing*. 10(2):1–14. DOI: 10.3390/rs10020202

Lonjou, V., Desjardins, C., Hagolle, O., Petrucci, B., Tremas, T., Dejus, M., Makarau, A. & Auer, S. 2016. MACCS-ATCOR joint algorithm (MAJA). In *Remote Sensing of Clouds and the Atmosphere XXI*. V. 10001. DOI: 10.1117/12.2240935

López-Granados, F., Torres-Sánchez, J., Serrano-Pérez, A., de Castro, A.I., Mesas-Carrascosa, F.J. & Peña, J.M. 2016. Early season weed mapping in sunflower using UAV technology: variability of herbicide treatment maps against weed thresholds. *Precision Agriculture*. 17(2):183–199. DOI: 10.1007/s11119-015-9415-8

López-Serrano, P.M., Corral-Rivas, J.J., Díaz-Varela, R.A., Álvarez-González, J.G. & López-Sánchez, C.A. 2016. Evaluation of radiometric and atmospheric correction algorithms for aboveground forest biomass estimation using landsat 5 TM data. *Remote Sensing*. 8(5). DOI: 10.3390/rs8050369

Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau, E. & Gascon, F. 2016. Sentinel-2 SEN2COR: L2A processor for users. In *European Space Agency, (Special Publication) ESA SP*. V. SP-740

Loveland, T.R. & Dwyer, J.L. 2012. Landsat: Building a strong future. *Remote Sensing of Environment*. 122(October 2000):22–29. DOI: 10.1016/j.rse.2011.09.022

Lu, J., Zhou, M., Gao, Y. & Jiang, H. 2018. Using hyperspectral imaging to discriminate yellow leaf curl disease in tomato leaves. *Precision Agriculture*. 19(3):379–394. DOI: 10.1007/s11119-017-9524-7

Ma, X., A. Huete, N. N. Tran, J. Bi, S. Gao, and Y. Zeng. 2020. “Sun-Angle Effects on Remote-Sensing Phenology Observed and Modelled Using Himawari-8.” Remote Sensing 12 (8): 1339. doi:10.3390/rs12081339

Ma, Y., Liu, S., Song, L., Xu, Z., Liu, Y., Xu, T. & Zhu, Z. 2018. Estimation of daily evapotranspiration and irrigation water efficiency at a Landsat-like scale for an arid irrigation area using multi-source remote sensing data. *Remote Sensing of Environment*. 216. DOI: 10.1016/j.rse.2018.07.019

Main, R., Cho, M.A., Mathieu, R., O’Kennedy, M.M., Ramoelo, A. & Koch, S. 2011. An investigation into robust spectral indices for leaf chlorophyll estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*. 66(6):751–761. DOI: 10.1016/j.isprsjprs .2011.08.001.

Maine, N., Lowenberg-DeBoer, J., Nell, W.T. & Alemu, Z.G. 2010. Impact of variable-rate application of nitrogen on yield and profit: A case study from South Africa. *Precision Agriculture*. 11(5):448–463. DOI: 10.1007/s11119-009-9139-8

Malenovský, Z., Lucieer, A., King, D.H., Turnbull, J.D. & Robinson, S.A. 2017. Unmanned aircraft system advances health mapping of fragile polar vegetation. *Methods in Ecology and Evolution*. 8(12):1842–1857. DOI: 10.1111/2041-210X.12833

Manandhar, A., Zhu, H., Ozkan, E. & Shah, A. 2020. Techno-economic impacts of using a laser-guided variable-rate spraying system to retrofit conventional constant-rate sprayers. *Precision Agriculture*. (0123456789). DOI: 10.1007/s11119-020-09712-8

Mango, N., Siziba, S. & Makate, C. 2017. The impact of adoption of conservation agriculture on smallholder farmers' food security in semi-arid zones of southern Africa. *Agriculture and Food Security*. 6(1). DOI: 10.1186/s40066-017-0109-5

Mansaray, L.R., Kanu, A.S., Yang, L. & Huang, J. 2022. Dynamic modelling of rice leaf area index with quad-source optical imagery and machine learning regression models. *Geocarto International*. 37(3). DOI: 10.1080/10106049.2020.1745299

Marcello, J., Eugenio, F., Perdomo, U. & Medina, A. 2016. Assessment of atmospheric algorithms to retrieve vegetation in natural protected areas using multispectral high resolution imagery. *Sensors (Switzerland)*. 16(10):1–18. DOI: 10.3390/s16101624

Markwell, J., Osterman, J.C. & Mitchell, J.L. 1995. Calibration of the Minolta SPAD-502 leaf chlorophyll meter. *Photosynthesis Research*. 46(3):467–472. DOI: 10.1007/BF 00032301

Marshak, A. Wen, G. Coakley Jr J. A. et al., "A simple model for the cloud adjacency effect and the apparent bluing of aerosols near clouds," Journal of Geophysical Research: Atmospheres, 113(D14), (2008).

Martins, V.S., Barbosa, C.C.F., de Carvalho, L.A.S., Jorge, D.S.F., Lobo, F. de L. & de Moraes Novo, E.M.L. 2017. Assessment of atmospheric correction methods for sentinel-2 MSI images applied to Amazon floodplain lakes. *Remote Sensing*. 9(4). DOI: 10.3390 /rs9040322

Mason, P. & Reading, B. 2007. Systematic Observation Requirements for Satellite-based Products for Climate. *23rd Conference on IIPS*.

Matasci, G., Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W. & Zald, H.S.J. 2018. Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and lidar plots. *Remote Sensing of Environment*. 209(December 2017):90–106. DOI: 10.1016/j.rse.2017.12.020

Meron, M., Tsipris, J., Orlov, V., Alchanatis, V. & Cohen, Y. 2010. Crop water stress mapping for site-specific irrigation by thermal imagery and artificial reference surfaces. *Precision Agriculture*. 11(2):148–162. DOI: 10.1007/s11119-009-9153-x

Micijevic, E., Haque, Md.O., Scaramuzza, P., Storey, J., Anderson, C. & Markham, B. 2019. Landsat 9 pre-launch sensor characterization and comparison with Landsat 8 results. DOI: 10.1117/12.2533102

Middleton, E. M. 1991. "Solar Zenith Angle Effects on Vegetation Indices in Tallgrass Prairie." Remote sensing of environment 38 (1): 45–62. Doi:10.1016/0034-4257(91)90071-D

Mirzaie, M., Darvishzadeh, R., Shakiba, A., Matkan, A.A., Atzberger, C. & Skidmore, A. 2014. Comparative analysis of different uni- and multi-variate methods for estimation of vegetation water content using hyper-spectral measurements. *International Journal of Applied Earth Observation and Geoinformation*. 26(1):1–11. DOI: 10.1016/j.jag .2013.04.004

Miura, T., Huete, A.R., Yoshioka, H. & Holben, B.N. 2001. An error and sensitivity analysis of atmospheric resistant vegetation indices derived from dark target-based atmospheric correction. *Remote Sensing of Environment*. 78(3):284–298. DOI: 10.1016/S0034-4257(01)00223-1

Mohanty, S., and R. Codell. 2002. "Sensitivity Analysis Methods for Identifying Influential Parameters in a Problem with a Large Number of Random Variables." WIT Transactions on Modelling and Simulation 31

Monaghan, J.M., Daccache, A., Vickers, L.H., Hess, T.M., Weatherhead, E.K., Grove, I.G. & Knox, J.W. 2013. More "crop per drop": Constraints and opportunities for precision irrigation in European agriculture. *Journal of the Science of Food and Agriculture*. 93(5):977–980. DOI: 10.1002/jsfa.6051

Moreira, C., Chou, Y.-L., Velmurugan, M., Ouyang, C., Sindhgatta, R. & Bruza, P. 2021. LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems*. (April):113561. DOI: 10.1016/j.dss. 2021.113561

Mulla, D.J. 2013. Twenty five years of remote sensing in precision agriculture. *Biosystems Engineering*. 114(4).

Mutanga, O. & Skidmore, A.K. 2004. Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing*. 25(19). DOI: 10.1080/01431160310001654923

Mutanga, O. & Skidmore, A.K. 2007. Red edge shift and biochemical content in grass canopies. *ISPRS Journal of Photogrammetry and Remote Sensing*. 62(1). DOI: 10.1016/j.isprsjpr s.2007.02.001

Mutanga, O., Adam, E. & Cho, M.A. 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(1):399–406. DOI: 10.1016/j.jag.2012.03.012

Mutanga, O., Adam, E., Adjorlolo, C. & Abdel-Rahman E.M. 2015. Evaluating the robustness of models developed from field spectral data in predicting African grass foliar nitrogen concentration using WorldView-2 image as an independent test dataset. *International Journal of Applied Earth Observation and Geoinformation*. 34(1):178–187. DOI: 10.1016/j.jag.2014.08.008

Mutanga, O., Dube, T. & Galal, O. 2017. Remote Sensing Applications : Society and Environment Remote sensing of crop health for food security in Africa : Potentials and constraints. 8(July):231–239.

Myneni, R. 2012. MODIS LAI/FPAR Product User's Guide [Online]. lpdaac.usgs.gov: USGS LP DAAC. Available: https://lpdaac.usgs.gov/sites/default/files/public/modis/docs/MODIS-LAI-FPAR-User-Guide.pdf.

Myneni, R.B. & Williams, D.L. 1994. On the relationship between FAPAR and NDVI. *Remote Sensing of Environment*. 49(3). DOI: 10.1016/0034-4257(94)90016-7

Myneni, R.B., Hoffman, S., Knyazikhin, Y., Privette, J.L., Glassy, J., Tian, Y., Wang, Y., Song, X., et al. 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sensing of Environment*. 83(1–2):214–231. DOI: 10.1016/S0034-4257(02)00074-3

Nagol, J. R., J. O. Sexton, D. -H. Kim, A. Anand, D. Morton, E. Vermote, and J. R. Townshend. 2015. "Bidirectional Effects in Landsat Reflectance Estimates: Is There a Problem to Solve?" Isprs Journal of Photogrammetry and Remote Sensing 103: 129– 135. doi:10.1016/j.isprsjprs.2014.09.006

Nasir, R., Khan, M.J., Arshad, M. & Khurshid, K. 2019. Convolutional neural network based regression for leaf water content estimation. In *2019 2nd International Conference on Latest Trends in Electrical Engineering and Computing Technologies, INTELLECT 2019*. DOI: 10.1109/INTELLECT47034.2019.8954985

Ndlovu, H.S., Odindi, J., Sibanda, M., Mutanga, O., Clulow, A., Chimonyo, V.G.P. & Mabhaudhi, T. 2021. A comparative estimation of maize leaf water content using machine learning techniques and unmanned aerial vehicle (Uav)-based proximal and remotely sensed data. *Remote Sensing*. 13(20). DOI: 10.3390/rs13204091

Nellemann, C. 2009. *The environmental food crisis: the environment's role in averting future food crises: a UNEP rapid response assessment*. UNEP/Earthprint.

Nguy-Robertson, A., Gitelson, A., Peng, Y., Viña, A., Arkebauer, T. & Rundquist, D. 2012. Green leaf area index estimation in maize and soybean: Combining vegetation indices to achieve maximal sensitivity. *Agronomy Journal*. 104(5):1336–1347. DOI: 10.2134 /agronj2012.0065

Nicodemus, F. E., J. C. Richmond, J. J. Hsia, I. Ginsberg, and T. Limperis. 1977. Geometrical Considerations and Nomenclature for Reflectance. Washington D.C., USA: U.S. Department of Commerce, National Bureau of Standards.

Novelli, F. & Vuolo, F. 2019. Assimilation of sentinel-2 leaf area index data into a physically-based crop growth model for yield estimation. *Agronomy*. 9(5). DOI: 10.3390 /agronomy9050255

Ohde, T. 2013. Spectral effects of atmospheric dust and clouds on estimation of chlorophyll-a concentration from radiation measurements. *Remote Sensing Letters*. 4(8):774–782. DOI: 10.1080/2150704X.2013.798707

Okun, O.; Priisalu, H. Random forest for gene expression based cancer classification: overlooked issues. *In Proceedings of Iberian conference on pattern recognition and image analysis*; pp. 483-490.

Ollinger, S. v. 2011. Sources of variability in canopy reflectance and the convergent properties of plants. *New Phytologist*. 189(2):375–394. DOI: 10.1111/j.1469-8137.2010.03536.x

Orynbaikyzy, A., Gessner, U. & Conrad, C. 2022. Spatial Transferability of Random Forest Models for Crop Type Classification Using Sentinel-1 and Sentinel-2. *Remote Sensing*. 14(6). DOI: 10.3390/rs14061493

Padilla, F.M., Gallardo, M., Peña-Fleitas, M.T., de Souza, R. & Thompson, R.B. 2018. Proximal optical sensors for nitrogen management of vegetable crops: A review. *Sensors (Switzerland)*. 18(7):1–23. DOI: 10.3390/s18072083

Pan, S. J. and Yang, Q. "A Survey on Transfer Learning," in *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

Parry, C., Blonquist, J.M. & Bugbee, B. 2014. In situ measurement of leaf chlorophyll concentration: Analysis of the optical/absolute relationship. *Plant Cell and Environment*. 37(11). DOI: 10.1111/pce.12324

Pasqualotto, N., D'Urso, G., Bolognesi, S.F., Belfiore, O.R., van Wittenberghe, S., Delegido, J., Pezzola, A., Winschel, C., et al. 2019a. Retrieval of evapotranspiration from sentinel-2: Comparison of vegetation indices, semi-empirical models and SNAP biophysical processor approach. *Agronomy*. 9(10). DOI: 10.3390/agronomy9100663

Pasqualotto, N., Delegido, J., van Wittenberghe, S., Rinaldi, M. & Moreno, J. 2019b. Multi-crop green LAI estimation with a new simple sentinel-2 LAI index (SeLI). *Sensors (Switzerland)*. 19(4). DOI: 10.3390/s19040904

Pathy, A., Meher, S. & P, B. 2020. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. *Algal Research*. 50. DOI: 10.1016/j.algal.2020.102006

Peng, Y. & Gitelson, A.A. 2011. Application of chlorophyll-related vegetation indices for remote estimation of maize productivity. *Agricultural and Forest Meteorology*. 151(9):1267–1276. DOI: 10.1016/j.agrformet.2011.05.005

Peng, Y. & Gitelson, A.A. 2012. Remote Sensing of Environment Remote estimation of gross primary productivity in soybean and maize based on total crop chlorophyll content. 117:440–448. DOI: 10.1016/j.rse.2011.10.021

Peralta, N.R., Costa, J.L., Balzarini, M., Castro Franco, M., Córdoba, M. & Bullock, D. 2015. Delineation of management zones to improve nitrogen management of wheat. *Computers and Electronics in Agriculture*. 110. DOI: 10.1016/j.compag.2014.10.017

Pereira-Sandoval, M., Ruescas, A., Urrego, P., Ruiz-Verdú, A., Delegido, J., Tenjo, C., Soria-Perpinyà, X., Vicente, E., et al. 2019. Evaluation of atmospheric correction algorithms over spanish inland waters for sentinel-2 multi spectral imagery data. *Remote Sensing*. 11(12):1–23. DOI: 10.3390/rs11121469

Persello, C. & Bruzzone, L. 2012. Active learning for domain adaptation in the supervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*. 50(11 PART1). DOI: 10.1109/TGRS.2012.2192740

Persello, C. & Bruzzone, L. 2014. Active and semisupervised learning for the classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*. 52(11). DOI: 10.1109/TGRS.2014.2305805

Pinty, B. & Verstraete, M.M. 1992. GEMI: a non-linear index to monitor global vegetation from satellites. *Vegetatio*. 101(1). DOI: 10.1007/BF00031911

Pipia, L., Amin, E., Belda, S., Salinero-Delgado, M. & Verrelst, J. 2021. Green lai mapping and cloud gap-filling using gaussian process regression in google earth engine. *Remote Sensing*. 13(3). DOI: 10.3390/rs13030403

Pipia, L., Muñoz-Marí, J., Amin, E., Belda, S., Camps-Valls, G. & Verrelst, J. 2019. Fusing optical and SAR time series for LAI gap fillingwith multioutput Gaussian processes. *Remote Sensing of Environment*. 235(September):111452. DOI: 10.1016/j.rse.2019.111452

Pocewicz, A., L. A. VIerling, L. B. Lentile, and R. Smith. 2007. "View Angle Effects on Relationships Between MISR Vegetation Indices and Leaf Area Index in a Recently Burned Ponderosa Pine Forest." Remote Sensing of Environment 107 (1–2): 322–333. doi:10.1016/j.rse.2006.06.019

Pritchard, R., Alexandridis, T., Amponsah, M., Ben Khatra, N., Brockington, D., Chiconela, T., Ortuño Castillo, J., Garba, I., et al. 2022. Developing capacity for impactful use of Earth Observation data: Lessons from the AfriCultuReS project. *Environmental Development*. 42(September 2021). DOI: 10.1016/j.envdev.2021.100695

Pullanagari, R.R., Kereszturi, G. & Yule, I. 2018. Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression. *Remote Sensing*. 10(7). DOI: 10.3390/rs10071117

Pullanagari, R.R., Kereszturi, G. & Yule, I.J. 2016. Mapping of macro and micro nutrients of mixed pastures using airborne AisaFENIX hyperspectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*. 117:1–10. DOI: 10.1016/j.isprsjprs.2016.03.010

Qiu, S., Zhu, Z. & He, B. 2019. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sensing of Environment*. 231(August 2018):111205. DOI: 10.1016/j.rse.2019.05.024

Rahman, H., D. Quadir, A. Zahedul Islam, and S. Dutta. 1999. "Viewing Angle Effect on the Remote Sensing Monitoring of Wheat and Rice Crops." Geocarto international 14 (1): 74–79. doi:10.1080/10106049908542097

Ramoelo, A., Cho, M.A., Mathieu, R., Madonsela, S., van de Kerchove, R., Kaszta, Z. & Wolff, E. 2015a. 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. DOI: 10.1016/j.jag .2014.12.010

Ramoelo, A., Cho, M.A., Mathieu, R., Madonsela, S., van de Kerchove, R., Kaszta, Z. & Wolff, E. 2015b. 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. DOI: 10.1016/j.jag.2014.12.010

Ramoelo, A., Skidmore, A.K., Cho, M.A., Schlerf, M., Mathieu, R. & Heitkönig, I.M.A. 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(1). DOI: 10.1016/j.jag.2012.05.009

Ramoelo, A., Skidmore, A.K., Schlerf, M., Mathieu, R. & Heitkönig, I.M.A. 2011. ISPRS Journal of Photogrammetry and Remote Sensing Water-removed spectra increase the retrieval accuracy when estimating savanna grass nitrogen and phosphorus concentrations. *ISPRS Journal of Photogrammetry and Remote Sensing*. 66(4):408–417. DOI: 10.1016/j.isprsjprs.2011.01.008

Rasmussen, C.E. Gaussian processes in machine learning. *In Proceedings of Summer school on machine learning*; pp. 63-71.

Rasul, A., Ibrahim, S., Onojeghuo, A.R. & Balzter, H. 2020. A trend analysis of leaf area index and land surface temperature and their relationship from global to local scale. *Land*. 9(10):1–17. DOI: 10.3390/land9100388

Rautiainen, M., Heiskanen, J. & Korhonen, L. 2012. Seasonal changes in canopy leaf area index and MODIS vegetation products for a boreal forest site in central Finland. *Boreal Environment Research*. 17(1):72–84.

Richter, K., Hank, T.B., Vuolo, F., Mauser, W. & D'Urso, G. 2012. Optimal exploitation of the sentinel-2 spectral capabilities for crop leaf area index mapping. *Remote Sensing*. 4(3). DOI: 10.3390/rs4030561

Richter, K.; Hank, T.B.; Mauser, W.; Atzberger, C. Derivation of biophysical variables from Earth observation data: Validation and statistical measures*. J. Appl. Remote Sens*. 2012, 6, 063557.

Rivera-Caicedo, J.P., Verrelst, J., Muñoz-Marí, J., Camps-Valls, G. & Moreno, J. 2017. Hyperspectral dimensionality reduction for biophysical variable statistical retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*. 132:88–101. DOI: 10.1016/j.isprsjprs.2017.08.012

Robnik-Šikonja, M. & Kononenko, I. 1997. An adaptation of {R}elief for attribute estimation in regression. *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)*. 5:296–304.

Robnik-Šikonja, M. & Kononenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*. 53(1):23–69.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sánchez, J.P. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*. 67(1):93–104. DOI: 10.1016/j.isprsjprs.2011.11.002

Roosjen, P. P., B. Brede, J. M. Suomalainen, H. M. Bartholomeus, L. Kooistra, and J. G. Clevers. 2018. "Improved Estimation of Leaf Area Index and Leaf Chlorophyll Content of a Potato Crop Using Multi-Angle Spectral Data–Potential of Unmanned Aerial Vehicle Imagery." International Journal of Applied Earth Observation and Geoinformation 66: 14–26. doi:10.1016/j.jag. 2017.10.012

Rouse, J. J., 1973. Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation. https://ntrs.nasa.gov/api/citations/19750020419/downloads/19750020419.pdf

Roy, D.P., Li, J., Zhang, H.K., Yan, L., Huang, H. & Li, Z. 2017. Examination of Sentinel-2A multi-spectral instrument (MSI) reflectance anisotropy and the suitability of a general method to normalize MSI reflectance to nadir BRDF adjusted reflectance. *Remote Sensing of Environment*. 199:25–38. DOI: 10.1016/j.rse.2017.06.019.

Ruder, Sebastian. 2017. "Transfer Learning - Machine Learning's Next Frontier." https://www.ruder.io/transfer-learning/

Ruescas, A.B., Pereira-Sandoval, M., Tenjo, C., Ruiz-Verdú, A., Steinmetz, F. & de Keukelaere, L. 2016. Sentinel-2 atmospheric correction inter-comparison over two lakes in Spain and Peru-Bolivia. In *Proceedings of the Colour and Light in the Ocean from Earth Observation (CLEO) Workshop, Frascati, Italy*. 6–8.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. 1986. Learning representations by back-propagating errors. *Nature*. 323(6088). DOI: 10.1038/323533a0

Sakamoto, T. 2020. Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*. 160(July 2019):208–228. DOI: 10.1016/j.isprsjprs.2019.12.012

Schaaf, C.B.; Gao, F.; Strahler, A.H.; Lucht, W.; Li, X.; Tsang, T.; Strugnell, N.C.; Zhang, X.; Jin, Y.; Muller, J.-P. First operational BRDF, albedo nadir reflectance products from MODIS. Remote sensing of Environment 2002, 83, 135-148.

Schmidt, K.S. & Skidmore, A.K. 2003. Spectral discrimination of vegetation types in a coastal wetland. *Remote Sensing of Environment*. 85(1):92–108. DOI: 10.1016/S0034-4257(02)00196-7

Segarra, J., Buchaillot, M.L., Araus, J.L. & Kefauver, S.C. 2020. DOI: 10.3390/agronomy10050641

Shah, S.H., Angel, Y., Houborg, R., Ali, S. & McCabe, M.F. 2019. A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat. *Remote Sensing*. 11(8):920. DOI: 10.3390/rs11080920

Sharma, R. C., K. Kajiwara, and Y. Honda. 2013. "Estimation of Forest Canopy Structural Parameters Using Kernel-Driven Bi- Directional Reflectance Model Based Multi-

Angular Vegetation Indices.” *Isprs Journal of Photogrammetry and Remote Sensing* 78: 50–57. doi:10.1016/j.isprsjprs.2012.12.006

Shawe-Taylor, J.; Cristianini, N. Kernel methods for pattern analysis; Cambridge university press: 2004.

Shelestov, A., Kolotii, A., Skakun, S., Baruth, B., Lozano, R.L. & Yailymov, B. 2017. Biophysical parameters mapping within the SPOT-5 take 5 initiative. *European Journal of Remote Sensing*. 50(1):300–309. DOI: 10.1080/22797254.2017.1324743

Shoko, C. & Mutanga, O. 2017. Int J Appl Earth Obs Geoinformation Seasonal discrimination of C3 and C4 grasses functional types : An evaluation of the prospects of varying spectral con fi gurations of new generation sensors. 62(April):47–55.

Sibanda, M., Mutanga, O., Dube, T., Mothapo, M.C. & Mafongoya, P.L. 2019. Remote sensing equivalent water thickness of grass treated with different fertiliser regimes using resample HyspIRI and EnMAP data. *Physics and Chemistry of the Earth*. 112(May):246–254. DOI: 10.1016/j.pce.2018.12.003

Sibanda, M., Mutanga, O., Dube, T., S Vundla, T. & L Mafongoya, P. 2019. Estimating LAI and mapping canopy storage capacity for hydrological applications in wattle infested ecosystems using Sentinel-2 MSI derived red edge bands. *GIScience and Remote Sensing*. 56(1):68–86. DOI: 10.1080/15481603.2018.1492213

Sibanda, M., Mutanga, O., Rouget, M. & Kumar, L. 2017. Estimating biomass of native grass grown under complex management treatments using worldview-3 spectral derivatives. *Remote Sensing*. 9(1). DOI: 10.3390/rs9010055

Silva Junior, C.A. da, Teodoro, L.P.R., Teodoro, P.E., Baio, F.H.R., Pantaleão, A. de A., Capristo-Silva, G.F., Facco, C.U., Oliveira-Júnior, J.F. de, et al. 2020. Simulating multispectral MSI bandsets (Sentinel-2) from hyperspectral observations via spectroradiometer for identifying soybean cultivars. *Remote Sensing Applications: Society and Environment*. 19. DOI: 10.1016/j.rsase.2020.100328

Snee, R.D. 1977. Validation of Regression Models: Methods and Examples. *Technometrics*. 19(4). DOI: 10.1080/00401706.1977.10489581

Sola, I., García-Martín, A., Sandonís-Pozo, L., Álvarez-Mozos, J., Pérez-Cabello, F., González-Audícana, M. & Montorio Llovería, R. 2018. Assessment of atmospheric correction methods for Sentinel-2 images in Mediterranean landscapes. *International Journal of Applied Earth Observation and Geoinformation*. 73(May):63–76. DOI: 10.1016/j.jag .2018.05.020

Souza, V. de A., Roberti, D.R., Ruhoff, A.L., Zimmer, T., Adamatti, D.S., de Gonçalves, L.G.G., Diaz, M.B., Alves, R. de C.M., et al. 2019. Evaluation of MOD16 algorithm over irrigated rice paddy using flux tower measurements in Southern Brazil. *Water (Switzerland)*. 11(9). DOI: 10.3390/w11091911

Sprintsin, M., Karnieli, A., Berliner, P., Rotenberg, E., Yakir, D. & Cohen, S. 2007. The effect of spatial resolution on the accuracy of leaf area index estimation for a forest planted in the desert transition zone. *Remote Sensing of Environment*. 109(4):416–428. DOI: 10.1016/j.rse.2007.01.020

Stamatiadis, S., Schepers, J.S., Evangelou, E., Glampedakis, A., Glampedakis, M., Dercas, N., Tsadilas, C., Tserlikakis, N., et al. 2020. Variable-rate application of high spatial resolution can improve cotton N-use efficiency and profitability. *Precision Agriculture*. 21(3):695–712. DOI: 10.1007/s11119-019-09690-6

Sterckx, S. Knaeps, E. Adriaensen, S. et al., "OPERA: An atmospheric correction for land and water." 2-5. *Sentinel-3 for Science Workshop*, Proceedings of a workshop held 2-5 June, 2015 in Venice, Italy. Edited by L. Ouwehand. ESA SP-734, ISBN 978-92-9221-298-8, id.7.

http://odnature.naturalsciences.be/downloads/publications/Sterckxetal_2015S3Forscience_finalv2-1.pdf

Sun, C., Sun, Z., Liu, T., Guo, D., Mu, S., Yang, H., Ju, W. & Li, J. 2013. The validation of a model estimating the leaf area index of grasslands in southern China. *Rangeland Journal*. 35(3):245–250. DOI: 10.1071/RJ12025

Szantoi, Z. & Strobl, P. 2019. Copernicus Sentinel-2 Calibration and Validation. *European Journal of Remote Sensing*. 52(1):253–255. DOI: 10.1080/22797254.2019.1582840

Tarnavsky, E., Garrigues, S. & Brown, M.E. 2008. Multiscale geostatistical analysis of AVHRR, SPOT-VGT, and MODIS global NDVI products. *Remote Sensing of Environment*. 112(2). DOI: 10.1016/j.rse.2007.05.008

Tavakoli, H. & Gebbers, R. 2019. Assessing Nitrogen and water status of winter wheat using a digital camera. *Computers and Electronics in Agriculture*. 157(January 2018):558–567. DOI: 10.1016/j.compag.2019.01.030

Tian, Y., Zhang, Y., Knyazikhin, Y., Myneni, R.B., Glassy, J.M., Dedieu, G. & Running, S.W. 2000. Prototyping of MODIS LAI and FPAR algorithm with LASUR and LANDSAT data. *IEEE Transactions on Geoscience and Remote Sensing*. 38(5 II). DOI: 10.1109 /36.868894

Tian, Y.C., Gu, K.J., Chu, X., Yao, X., Cao, W.X. & Zhu, Y. 2013. Comparison of different hyperspectral vegetation indices for canopy leaf nitrogen concentration estimation in rice. *Plant and Soil*. 376(1):193–209. DOI: 10.1007/s11104-013-1937-0

Tillack, A., Clasen, A., Kleinschmit, B. & Förster, M. 2014. Estimation of the seasonal leaf area index in an alluvial forest using high-resolution satellite-based vegetation indices. *Remote Sensing of Environment*. 141:52–63. DOI: 10.1016/j.rse.2013.10.018

Tong, A. & He, Y. 2013. Comparative analysis of SPOT, Landsat, MODIS, and AVHRR normalized difference vegetation index data on the estimation of leaf area index in a mixed grassland ecosystem. *Journal of Applied Remote Sensing*. 7(1). DOI: 10.1117 /1.jrs.7.073599

Tong, X.Y., Xia, G.S., Lu, Q., Shen, H., Li, S., You, S. & Zhang, L. 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*. 237(July 2018):111322. DOI: 10.1016 /j.rse.2019.111322

Trishchenko, A.P., Cihlar, J. & Li, Z. 2002. Effects of spectral response function on surface reflectance and NDVI measured with moderate resolution satellite sensors. *Remote Sensing of Environment*. 81(1). DOI: 10.1016/S0034-4257(01)00328-5

Tucker, C.J. & Maxwell, E.L. 1976. Sensor design for monitoring vegetation canopies. *Photogrammetric Engineering and Remote Sensing*. 42(11).

Tucker, C.J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*. 8(2). DOI: 10.1016/0034-4257(79)90013-0

Uddling, J., Gelang-Alfredsson, J., Piikki, K. & Pleijel, H. 2007. Evaluating the relationship between leaf chlorophyll concentration and SPAD-502 chlorophyll meter readings. *Photosynthesis Research*. 91(1):37–46. DOI: 10.1007/s11120-006-9077-5

UN 2015. Transforming our world: The 2030 agenda for sustainable development. General Assembley 70 session. United Nations. http://fra.europa.eu/en/law-reference/un-general-assembly-resolution-701-2015-transforming-our-world-2030-agenda

Upreti, D., Huang, W., Kong, W., Pascucci, S., Pignatti, S., Zhou, X., Ye, H. & Casa, R. 2019. A comparison of hybrid machine learning algorithms for the retrieval of wheat biophysical variables from sentinel-2. *Remote Sensing*. 11(5). DOI: 10.3390 /rs11050481

Ustin, S.L., Gitelson, A.A., Jacquemoud, S., Schaepman, M., Asner, G.P., Gamon, J.A. & Zarco-Tejada, P. 2009. Retrieval of foliar information about plant pigment systems from high resolution spectroscopy. *Remote Sensing of Environment*. 113(SUPPL. 1):67–77. DOI: 10.1016/j.rse.2008.10.019

Vanonckelen, S., Lhermitte, S. & Van Rompaey, A. 2013. The effect of atmospheric and topographic correction methods on land cover classification accuracy. *International Journal of Applied Earth Observation and Geoinformation*. 24(1):9–21. DOI: 10.1016/j.jag.2013.02.003

Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Networks* 1999, 10, 988-999.

Verger, A., Baret, F. & Camacho, F. 2011. Optimal modalities for radiative transfer-neural network estimation of canopy biophysical characteristics: Evaluation over an agricultural area with CHRIS/PROBA observations. *Remote Sensing of Environment*. 115(2):415–426. DOI: 10.1016/j.rse.2010.09.012

Vermote, E., Doxani, G., Gascon, F. & Roger, J.C. 2019. ACIX - Atmospheric atmospheric correction inter-comparison exercise. In *International Geoscience and Remote Sensing Symposium (IGARSS)*. V. 2019-January. DOI: 10.1109/IGARSS.2019.8900457.

Verrelst, J., Alonso, L., Camps-Valls, G., Delegido, J. & Moreno, J. 2012. Retrieval of vegetation biophysical parameters using Gaussian process techniques. *IEEE*

Transactions on Geoscience and Remote Sensing. 50(5 PART 2):1832–1843. DOI: 10.1109/TGRS.2011.2168962

Verrelst, J., M. E. Schaepman, B. Koetz, and M. Kneubühler. 2008. "Angular Sensitivity Analysis of Vegetation Indices Derived from CHRIS/PROBA Data." Remote Sensing of Environment 112 (5): 2341–2353. doi:10.1016/j.rse.2007.11. 001

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J.P., Camps-Valls, G. & Moreno, J. 2012. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sensing of Environment*. 118:127–139. DOI: 10.1016/j.rse.2011.11.002

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J.P., Camps-valls, G. & Moreno, J. 2012. Remote Sensing of Environment Machine learning regression algorithms for biophysical parameter retrieval : Opportunities for Sentinel-2 and -3. 118:127–139. DOI: 10.1016/j.rse.2011.11.002

Verrelst, J., Rivera, J.P., Gitelson, A., Delegido, J., Moreno, J. & Camps-Valls, G. 2016. Spectral band selection for vegetation properties retrieval using Gaussian processes regression. *International Journal of Applied Earth Observation and Geoinformation*. 52:554–567. DOI: 10.1016/j.jag.2016.07.016

Verrelst, J., Rivera, J.P., Moreno, J. & Camps-valls, G. 2013. ISPRS Journal of Photogrammetry and Remote Sensing Gaussian processes uncertainty estimates in experimental Sentinel-2 LAI and leaf chlorophyll content retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*. 86:157–167. DOI: 10.1016/j.isprsjprs.2013.09.012

Verrelst, J., Rivera, J.P., Veroustraete, F., Muñoz-Marí, J., Clevers, J.G.P.W., Camps-Valls, G. & Moreno, J. 2015. Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods - A comparison. *ISPRS Journal of Photogrammetry and Remote Sensing*. 108:260–272. DOI: 10.1016/j.isprsjprs .2015.04.013

Verrelst, J.; Alonso, L.; Camps-Valls, G.; Delegido, J.; Moreno, J. Retrieval of vegetation biophysical parameters using Gaussian process techniques. *IEEE Transactions on Geoscience and Remote Sensing* 2011, 50, 1832-1843.

Vierling, L. A., D. W. Deering, and T. F. Eck. 1997. "Differences in Arctic Tundra Vegetation Type and Phenology as Seen Using Bidirectional Radiometry in the Early Growing Season." Remote sensing of environment 60 (1): 71–82. doi:10.1016/ S0034-4257(96)00139-3

Viña, A., Gitelson, A.A., Nguy-Robertson, A.L. & Peng, Y. 2011. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. *Remote Sensing of Environment*. 115(12):3468–3478. DOI: 10.1016/j.rse.2011.08.010

Vincini, M., Calegari, F. & Casa, R. 2016. Sensitivity of leaf chlorophyll empirical estimators obtained at Sentinel-2 spectral resolution for different canopy structures. *Precision Agriculture*. 17(3):313–331. DOI: 10.1007/s11119-015-9424-7

Vuolo, F., Neugebauer, N., Bolognesi, S.F., Atzberger, C. & D'Urso, G. 2013. Estimation of leaf area index using DEIMOS-1 data: Application and transferability of a semi-empirical relationship between two agricultural areas. *Remote Sensing*. 5(3):1274–1291. DOI: 10.3390/rs5031274

Vuolo, F., Zóltak, M., Pipitone, C., Zappa, L., Wenng, H., Immitzer, M., Weiss, M., Baret, F., et al. 2016. Data service platform for Sentinel-2 surface reflectance and value-added products: System use and examples. *Remote Sensing*. 8(11). DOI: 10.3390/rs8110938

Walker, R.J. 2016. Population Growth and its Implications for Global Security. *American Journal of Economics and Sociology*. 75(4). DOI: 10.1111/ajes.12161

Wang, L., Chang, Q., Yang, J., Zhang, X. & Li, F. 2018. Estimation of paddy rice leaf area index using machine learning methods based on hyperspectral data from multi-year experiments. *PLoS ONE*. 13(12):1–16. DOI: 10.1371/journal.pone.0207624

Wang, L., Wang, P., Li, L., Xun, L., Kong, Q. & Liang, S. 2018. Developing an integrated indicator for monitoring maize growth condition using remotely sensed vegetation temperature condition index and leaf area index. *Computers and Electronics in Agriculture*. 152(17):340–349. DOI: 10.1016/j.compag.2018.07.026

Weiss, M. & Baret, F. 2016. S2ToolBox Level 2 products: LAI, FAPAR, FCOVER. Institut National de la Recherche Agronomique (INRA), Avignon.

Weiss, M. & Baret, F. 2016. S2ToolBox Level 2 products: LAI, FAPAR, FCOVER - Version 1.1. *Sentinel2 ToolBox Level2 Products*. 53. Available: http://step.esa.int/docs/extra/ATBD_S2ToolBox_L2B_V1.1.pdf

Weiss, M., and F. Baret 2020. S2ToolBox Level 2 products: LAI, FAPAR, FCOVER. V2.0 ed.

Weiss, M., Baret, F., Smith, G.J., Jonckheere, I. & Coppin, P. 2004. Review of methods for in situ leaf area index (LAI) determination Part II. Estimation of LAI, errors and sampling. *Agricultural and Forest Meteorology*. 121(1–2):37–53. DOI: 10.1016/j.agrformet.2003.08.001

Wen, D., Tongyu, X., Fenghua, Y. & Chunling, C. 2018. Measurement of nitrogen content in rice by inversion of hyperspectral reflectance data from an unmanned aerial vehicle. *Ciencia Rural*. 48(6):1–8. DOI: 10.1590/0103-8478cr20180008

Wen, G. Cahalan, R. F. Tsay, S. C. et al., "Impact of cumulus cloud spacing on Landsat atmospheric correction and aerosol retrieval," Journal of Geophysical Research: Atmospheres, 106(D11), 12129-12138 (2001).

Wilhelm, W.W., Ruwe, K. & Schlemmer, M.R. 2000. Comparison of three leaf area index meters in a corn canopy. *Crop Science*. 40(4). DOI: 10.2135/cropsci2000.4041179x

Wilson, R.T., Milton, E.J. & Nield, J.M. 2014. Spatial variability of the atmosphere over southern England, and its effect on scene-based atmospheric corrections. *International Journal of Remote Sensing*. 35(13). DOI: 10.1080/01431161.2014.939781

Wu, C., Liu, M., Liu, X., Wang, T. & Wang, L. 2019. Developing a new spectral index for detecting cadmium-induced stress in rice on a regional scale. *International Journal of Environmental Research and Public Health*. 16(23). DOI: 10.3390/ijerph16234811

Wu, C.; Niu, Z.; Tang, Q.; Huang, W. Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agricultural and forest meteorology* 2008, 148, 1230-1241.

Wu, J., Zhang, J., Lü, A. & Zhou, L. 2012. An exploratory analysis of spectral indices to estimate vegetation water content using sensitivity function. *Remote Sensing Letters*. 3(2):161–169. DOI: 10.1080/01431161.2011.551845

Wu, M., Wu, C., Huang, W., Niu, Z. & Wang, C. 2015. High-resolution Leaf Area Index estimation from synthetic Landsat data generated by a spatial and temporal data fusion model. *Computers and Electronics in Agriculture*. 115:1–11. DOI: 10.1016/j.compag .2015.05.003

Wulder, M.A., Loveland, T.R., Roy, D.P., Crawford, C.J., Masek, J.G., Woodcock, C.E., Allen, R.G., Anderson, M.C., et al. 2019. Current status of Landsat program, science, and applications. *Remote Sensing of Environment*. 225. DOI: 10.1016/j.rse.2019.02.015

Wulder, M.A., White, J.C., Goward, S.N., Masek, J.G., Irons, J.R., Herold, M., Cohen, W.B., Loveland, T.R., et al. 2008. Landsat continuity: Issues and opportunities for land cover monitoring. *Remote Sensing of Environment*. 112(3):955–969. DOI: 10.1016/j.rse .2007.07.004

Xiaoyan, W., Zhiwei, L., Wenjun, W. & Jiawei, W. 2020. Chlorophyll content for millet leaf using hyperspectral imaging and an attention-convolutional neural network. *Ciencia Rural*. 50(3). DOI: 10.1590/0103-8478cr20190731

Xie, Q., Dash, J., Huete, A., Jiang, A., Yin, G., Ding, Y., Peng, D., Hall, C.C., et al. 2019. Retrieval of crop biophysical parameters from Sentinel-2 remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation*. 80. DOI: 10.1016/j.jag.2019.04.019

Xiong, J., Thenkabail, P.S., Gumma, M.K., Teluguntla, P., Poehnelt, J., Congalton, R.G., Yadav, K. & Thau, D. 2017. Automated cropland mapping of continental Africa using Google Earth Engine cloud computing. *ISPRS Journal of Photogrammetry and Remote Sensing*. 126:225–244. DOI: 10.1016/j.isprsjprs.2017.01.019

Xu, M., Liu, R., Chen, J.M., Liu, Y., Shang, R., Ju, W., Wu, C. & Huang, W. 2019. Retrieving leaf chlorophyll content using a matrix-based vegetation index combination approach. *Remote Sensing of Environment*. 224(February):60–73. DOI: 10.1016/j.rse.2019.01.039

Xu, X.Q., Lu, J.S., Zhang, N., Yang, T.C., He, J.Y., Yao, X., Cheng, T., Zhu, Y., et al. 2019. Inversion of rice canopy chlorophyll content and leaf area index based on coupling of radiative transfer and Bayesian network models. *ISPRS Journal of Photogrammetry and Remote Sensing*. 150(February):185–196. DOI: 10.1016/j.isprsjprs.2019.02.013

Xu, Y., Smith, S.E., Grunwald, S., Abd-Elrahman, A. & Wani, S.P. 2017. Incorporation of satellite remote sensing pan-sharpened imagery into digital soil prediction and mapping models to characterize soil property variability in small agricultural fields. *ISPRS Journal of Photogrammetry and Remote Sensing*. 123. DOI: 10.1016/j.isprsjprs .2016.11.001

Yan, G., Hu, R., Luo, J., Weiss, M., Jiang, H., Mu, X., Xie, D., Zhang, W., et al. 2019. Review of indirect optical measurements of leaf area index: Recent advances, challenges, and perspectives. *Agricultural and Forest Meteorology*. 265(December 2018):390–411. DOI: 10.1016/j.agrformet.2018.11.033

Ye, F. 2018. Evolving the SVM model based on a hybrid method using swarm optimization techniques in combination with a genetic algorithm for medical diagnosis. *Multimedia Tools and Applications*. 77(3):3889–3918.

Yi, Q., Wang, F., Bao, A. & Jiapaer, G. 2014. Leaf and canopy water content estimation in cotton using hyperspectral indices and radiative transfer models. *International Journal of Applied Earth Observation and Geoinformation*. 33(1):67–75. DOI: 10.1016/j.jag.2014.04.019

Yin, G., J. Li, B. Xu, Y. Zeng, S. W, K. Yan, A. Verger, and G. Liu. 2020. "PLC-C: An Integrated Method for Sentinel-2 Topographic and Angular Normalization." IEEE Geoscience and Remote Sensing Letters 18 (8): 1446–1450. doi:10.1109/ LGRS.2020.3001905

Yu, F.H., Xu, T.Y., Du, W., Ma, H., Zhang, G.S. & Chen, C.L. 2017. Radiative transfer models (RTMs) for field phenotyping inversion of rice based on UAV hyperspectral remote sensing. *International Journal of Agricultural and Biological Engineering*. 10(4):150–157. DOI: 10.25165/j.ijabe.20171004.3076

Yu, L. & Liu, H. 2004. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*. 5:1205–1224.

Zaroug, M.A.H., Sylla, M.B., Giorgi, F., Eltahir, E.A.B. & Aggarwal, P.K. 2013. A sensitivity study on the role of the swamps of southern Sudan in the summer climate of North Africa using a regional climate model. *Theoretical and Applied Climatology*. 113(1–2). DOI: 10.1007/s00704-012-0751-6

Zhang, C. & Kovacs, J.M. 2012. The application of small unmanned aerial systems for precision agriculture: A review. *Precision Agriculture*. 13(6):693–712. DOI: 10.1007/s11119-012-9274-5.

Zhang, M., Su, W., Fu, Y., Zhu, D., Xue, J.H., Huang, J., Wang, W., Wu, J., et al. 2019. Super-resolution enhancement of Sentinel-2 image for retrieving LAI and chlorophyll content of summer corn. *European Journal of Agronomy*. 111(17):125938. DOI: 10.1016/j.eja.2019.125938

Zhang, M., Yuan, X. & Otkin, J.A. 2020. Remote sensing of the impact of flash drought events on terrestrial carbon dynamics over China. *Carbon Balance and Management*. 15(1):1–11. DOI: 10.1186/s13021-020-00156-1

Zhang, Y., Xia, C., Zhang, X., Cheng, X., Feng, G., Wang, Y. & Gao, Q. 2021. Estimating the maize biomass by crop height and narrowband vegetation indices derived from UAV-based hyperspectral images. *Ecological Indicators*. 129. DOI: 10.1016/j.ecolind.2021.107985

Zhao, B., Huang, B. & Zhong, Y. 2017. Transfer Learning with Fully Pretrained Deep Convolution Networks for Land-Use Classification. *IEEE Geoscience and Remote Sensing Letters*. 14(9). DOI: 10.1109/LGRS.2017.2691013

Zhao, C., Wang, Z., Wang, J., Huang, W. & Guo, T. 2011. Early detection of canopy nitrogen deficiency in winter wheat (Triticum aestivum L.) based on hyperspectral measurement of canopy chlorophyll status. *New Zealand Journal of Crop and Horticultural Science*. 39(4):251–262. DOI: 10.1080/01140671.2011.588713

Zheng, H., Cheng, T., Zhou, M., Li, D., Yao, X., Tian, Y., Cao, W. & Zhu, Y. 2019. Improved estimation of rice aboveground biomass combining textural and spectral analysis of UAV imagery. *Precision Agriculture*. 20(3):611–629. DOI: 10.1007/s11119-018-9600-7

Zhou, J., Khot, L.R., Boydston, R.A., Miklas, P.N. & Porter, L. 2018. Low altitude remote sensing technologies for crop stress monitoring: a case study on spatial and temporal monitoring of irrigated pinto bean. *Precision Agriculture*. 19(3):555–569. DOI: 10.1007/s11119-017-9539-0

Zhou, Z., Plauborg, F., Kristensen, K. & Andersen, M.N. 2017. Dry matter production, radiation interception and radiation use efficiency of potato in response to temperature and nitrogen application regimes. *Agricultural and Forest Meteorology*. 232:595–605. DOI: 10.1016/j.agrformet.2016.10.017

Zhu, Q., Chen, L., Hu, H., Pirasteh, S., Li, H. & Xie, X. 2020. Unsupervised Feature Learning to Improve Transferability of Landslide Susceptibility Representations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 13:3917–3930. DOI: 10.1109/JSTARS.2020.3006192

Zhu, Z., Bi, J., Pan, Y., Ganguly, S., Anav, A., Xu, L., Samanta, A., Piao, S., et al. 2013. Global data sets of vegetation leaf area index (LAI)3g and fraction of photosynthetically active radiation (FPAR)3g derived from global inventory modeling and mapping studies (GIMMS) normalized difference vegetation index (NDVI3G) for the period 1981 to 2011. *Remote Sensing*. 5(2). DOI: 10.3390/rs5020927

Zhu, Z., Wang, S. & Woodcock, C.E. 2015. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and Sentinel 2 images. *Remote Sensing of Environment*. 159. DOI: 10.1016/j.rse.2014.12.014

Zhu, Z., Wulder, M.A., Roy, D.P., Woodcock, C.E., Hansen, M.C., Radeloff, V.C., Healey, S.P., Schaaf, C., et al. 2019. Benefits of the free and open Landsat data policy. *Remote Sensing of Environment*. 224. DOI: 10.1016/j.rse.2019.02.016

Zhu, Z.M., Yang, C., Cao, M.M. and Liu, K.,. 2010. Changes of plant leaf area and its relationships with soil factors in the process of grassland desertification. *Chinese Journal of Ecology*. 29(12).

Zisi, T., Alexandridis, T.K., Kaplanis, S., Navrozidis, I., Tamouridou, A.A., Lagopodi, A., Moshou, D. & Polychronos, V. 2018. Incorporating surface elevation information in UAV multispectral images for mapping weed patches. *Journal of Imaging*. 4(11) :2384-2389. DOI: 10.3390/jimaging4110132

Zou, H.; Hastie, T.; Tibshirani, R. Sparse principal component analysis. *Journal of computational and graphical statistics* 2006, 15, 265-286.