Rice Yield Prediction in Ibagué, Colombia using a Multivariate Machine Learning Approach

Rice inspection in Tolima Department, Colombia, retrieved from Flickr - CIAT (2010).

James Kibble

MSc Remote Sensing and GIS

Aberystwyth University

2020

Acknowledgements

Firstly, I would like to express my gratitude to my dissertation supervisor Dr. Andy Hardy, for his continued support and valuable advice from beginning to the completion of this research. I would also like to thank Liliana Castillo Villamor for introducing me to the project and answering all my questions.

I express thanks to Meteoblue for granting access to their historic climate data repository, alongside Hacienda El Escobal, for sharing their extensive *in situ* field measurements, without which this investigation would not exist.

Lastly, I thank my partner Hannah and my family for supporting me along the way.

Abstract

Rice (*Oryza sativa* L.) is the second largest cultivated cereal crop globally, and the most affordable source of protein and carbohydrates in tropical South America. An increasingly erratic climate risks damaging agricultural production in the region, with rice vulnerable to environmental extremes and shifting weather patterns. In Colombia, temperatures will increase 5 to 7^{o}C, alongside a reduction of 10% precipitation rates by 2100. This poses a significant threat to regional food security, whereby 60% of current rice cultivated land may be untenable by 2050. Resultingly, a multivariate approach is required to mitigate the damages these projections pose. Through precision agriculture (PA), a selection of Earth Observation (EO) metrics, climate variables, and *in situ* field measurements were collated to establish a robust method for rice yield prediction in an area of Tolima department, Colombia.

A persistent issue with optical satellite EO in tropical regions is the high presence of cloud cover, limiting data availability. Filling these optical information gaps is crucial in the Tropics due to pertinent threats to food security. However, this can be mitigated using cloud penetrating Synthetic Aperture Radar (SAR) data. Resultingly, the relationship between Sentinel-2 generated vegetation indices (VIs), and Sentinel-1 SAR metrics was assessed at different rice phenological stages using machine learning techniques. Here, a maximum performance of R^2 0.583 was generated with the Normalised Difference Vegetation Index (NDVI) while rice was at the vegetative development stage. Future improvements were explored, notably through further phenological division to minimise dielectric influence.

Moreover, this investigation explored yield prediction capabilities of various rice cultivars, whereby reproductive EVI values, drought information, and seedling period proved most impactful to model prediction, generating robust predictions for Escobal 518 (R^2 0.949), Triunfo (R^2 0.697), and Fedearroz 68 (R^2 0.551). Additionally, contrasting variable feature importance suggest proposed methods can be harnessed for managerial decision-making, with some cultivars more suitable to projected regional climate fluctuations. Resultingly, this investigation presents a robust strategy for rice yield prediction in a location crucial to Colombian food security, through a multivariate machine learning approach, while exploring cloud cover mitigation to optimise remotely sensed data coverage.

Key words: *precision agriculture, rice, yield, prediction, food security, machine learning, Sentinel-1, Sentinel-2, cloud cover*

Table of Contents

Figures

Tables

List of Appendices

Chapter 1

1.0. Introduction

Demand for agricultural produce is predicted to increase over 50% by 2050 as the global population surpasses 9 billion (FAO, 2017; United Nations, 2019). The ability to predict agricultural yields accurately and methodically is increasingly crucial in meeting this challenge (Weiss et al., 2020). Effective yield prediction holds a multitude of advantages, including optimization of sustainable farming practices to meet projected climate changes (Wheeler and von Braun, 2013; Areal et al., 2018), stabilising food security (Di Falco et al., 2012; Biswas and Bhattacharyya, 2019; Weiss et al., 2020), and extrapolation to wider regions (Weiss et al., 2020).

Rice (*Oryza sativa* L.) is the second most cultivated cereal crop globally, with an approximate yield of 800 billion tonnes in 2018 (FAO, 2020). Owing to the inexpensive source of protein and carbohydrates, rice provides a vital component to South American diets (Zorilla et al., 2012). Colombia heavily relies on rice, accounting for an average intake of 37.7 kg per person annually (Delerce et al., 2016), representing both the greatest production value and the second most produced crop by area nationally (DANE, 2016; Arango-Londoño et al., 2020). However, risks to production are increasingly apparent, threatening both local and national socio-economic consequences (Meinke and Stone, 2005; Delerce et al., 2016).

Colombia generates lower rice yields compared to neighbouring countries, while consumption per capita continues to rise (Castro-Llanos et al., 2019). This has introduced expanding reliance on imports, potentially threatening national food security (Castro-Llanos et al., 2019). These pressures are compounded by increasingly erratic weather conditions, which will negatively influence agricultural output (Delerce et al., 2016: Jiménez et al., 2019). Colombia is projected to experience an increase in both magnitude and frequency of these events in the coming decades; between 2005 and 2100, temperatures will surge by 5 to 7 ^{o}C, alongside a 10% precipitation reduction (Pachauri et al., 2014). Such influences could reduce Colombian rice production by between 5 and 29% (Iizumi et al., 2014; Quevedo Amaya et al., 2019), yet even more concerning is Castro-Llanos et al.'s (2019) notion that 60% of Colombia's land currently cultivated for rice will be unmanageable by 2050, a reduction from 4.4 to 1.8 million hectares, owing to predicted rising temperatures and decreasing water availability. This is largely expected to effect land at lower elevations, while more elevated agricultural settings will be less impacted by climatic fluctuations (Castro-Llanos et al., 2019). Thus, the need for a robust

strategy to predict rice yields in areas with future scope for cultivation is crucial in ensuring national food security (Castro-Llanos et al., 2019; Jiménez et al., 2019; Weiss et al., 2020). A successful prediction strategy several months prior to harvest would prove invaluable to both regional farmers and national stakeholders (Noureldin et al., 2013).

The implementation of Precision Agriculture (PA) through remotely sensed data in conjunction with machine learning technology allows for such a strategy in a systematic and methodical manner (Liakos et al., 2018). PA is an information and technology-based approach to identify, analyse, and combat spatial and temporal crop fluctuations, aiming to optimise profits and sustainability, while minimising environmental damage (Lillesand et al., 2015; Finger et al., 2019). By harnessing a combination of EO metrics, environmental variables, and *in situ* field measurements, machine learning algorithms can be utilised to tackle weakened food security and climatic shifting on both a local and national level (Weiss et al., 2020). Machine learning is a valuable tool in PA practices, where modelling acts as a representation of the world based upon simplified assumptions (Spiegelhalter, 2019). PA provides a range of algorithms that harness large data volumes to automatically learn and improve upon existing correlations (Géron, 2019). This approach can ultimately uncover key factors determining yield production rates, allowing future improvements to agricultural practices (Chlingaryan et al., 2018; Weiss et al., 2020).

A valuable metric retrieved from remotely sensed data are vegetation indices (VIs), these being mathematical quantities derived from spectral band ratios to better discern vegetation properties (Wiegand et al., 1979; Lillesand et al., 2015). However, the tropical South American climate experiences significant cloud coverage, something problematic when using optical satellite data (Filgueiras et al., 2019). The implementation of cloud-penetrating Synthetic Aperture Radar (SAR), can combat this, allowing data collection unimpeded by weather conditions sunlight (Torres et al., 2012). Thus, opportunity exists for combining optical and SAR-based satellite data for a hybrid approach, maximising coverage and available information (Filgueiras et al., 2019). Consequently, development of an extensive model to accurately predict rice yield is required, specifically focused on Colombia, to strengthen national food security during projected climate shifts (Zabel et al., 2014; Castro-Llanos et al., 2019).

The research detailed in this investigation will be structured accordingly: a precise overview of research aims and objectives; an extensive review of academic literature surrounding key

research themes; a detailed and justifiable strategy for rice yield prediction appropriate to the study area, alongside a robust approach to mitigate cloud coverage to maximise data availability; a thorough presentation and interpretation of investigative results, succeeded by a rigorous discussion of findings in relation to wider literature; concluding remarks of the research presented, alongside potential future exploratory research avenues.

1.1. Aims and Objectives

This investigation aims to explore an effective and methodical approach to rice yield prediction within an area of central-western Colombia deemed suitable for future rice production (Castro-Llanos et al., 2019). A multivariate machine learning approach using remotely sensed data metrics, climate variables, and *in situ* field measurements will be explored to achieve this. To maximise optical data availability, cloud mitigation techniques will also be investigated. The completion of the investigative aims will be accomplished through the following objectives:

- Obtain and perform quality assurance on field data and climate variables from the study area, which have proved influential during previous yield prediction investigations.
- Identify appropriate EO data, undertaking necessary pre-processing and analysis to extract metrics.
- Investigate relationships between optically retrieved metrics and SAR data using machine learning techniques to explore cloud mitigation prospects.
- Using *in situ* GPS harvester measurements and EO metrics, determine vegetation index relationships to rice yield rates.
- Ascertain efficient and successful machine learning algorithms to predict yield rates in the study area, employing a range of appropriate collated environmental variables, remotely sensed data metrics, and *in situ* field measurements.
- Assess the prediction accuracy of cloud mitigation and yield prediction algorithms, while exploring the possibility for extrapolation to other locales.
- Provide a detailed description of recommendations from investigative findings to benefit stakeholders, strengthen food security, and establish future avenues of research.

2.0. Literature Review

A thorough review of PA practices will be delivered to explore the use of remote sensing techniques and machine learning for rice production. Specific focus will be given to combining multiple data sources for a hybrid approach, alongside the application of VIs. A detailed examination of past yield forecasting investigations will be presented, with significant emphasis on Colombia and neighbouring regions. This will provide a comprehensive understanding of surrounding literature and its influence upon the current body of work.

2.1. Remote sensing in agriculture

The main objective of PA is to harness remotely sensed spectral information to optimise managerial decision-making across both space and time (Whelan and Taylor, 2013). For effective implementation, remotely sensed data should fulfil several requirements: high spatial resolution for optimum data retrieval; high temporal resolution for sufficient crop phenological time-series coverage; high spectral resolution to enable detailed examination; and freely available open-source accessibility (Campos-Taberner et al., 2017; Weiss et al., 2020). Moreover, the platform used for remotely sensed data retrieval is an important consideration (Verger et al., 2014a; Zhou et al., 2017).

Zhou et al. (2017) demonstrated success with aerially-retrieved data for rice detection and yield forecasting in China, benefiting from the diminished effects of cloud coverage compared to optical satellite data. Yet aerial data collection is often time consuming and costly compared to satellite data, and simply unrealistic in inaccessible regions (Weiss et al., 2020). Contrastingly, satellite-based remote sensing often benefits from extensive open-source repositories with sufficient temporal resolution, proving more suitable in certain circumstances focussing on developing countries (Weiss et al., 2020). Moreover, PA technology has been adopted for agricultural vehicles, one example being specialist harvesters which collate crop weight samples among other variables throughout harvesting (Leroux et al., 2018). This information is spatially collected through a Geographic Positioning System (GPS), which allows a ‘within-plot’ yield map at regular intervals, rather than the more typical ‘overall-plot’ approach encompassing yield information per field (Pringle et al., 2003; Leroux et al., 2018).

PA has benefitted from the continued development of remotely sensed VIs; spectral transformations derived from multiple specific regions of the electromagnetic spectrum (Lillesand et al., 2015). These often provide a deeper understanding of plant properties than typical band analysis, forming an integral part of yield forecasting (Xue and Su, 2017; Chlingaryan et al., 2018).

2.2. Phenological monitoring with vegetation indices

VIs utilise specific spectral band combinations to highlight relationships between remotely sensed data and vegetation properties (Wiegand et al., 1979; Lillesand et al., 2015). VI implementation for PA is abundant in prior research, including such applications as yield prediction (Johnson et al., 2016; Shiu and Chuang, 2019), canopy radiation use efficiency (Garbulsky et al., 2011), and establishing nitrogen content (Clevers and Gitelson, 2013; Delloye et al., 2018), yet their prediction accuracy for vegetative parameters has been questioned (Marshall et al., 2016; Alvino and Marion, 2017). Owing to the extensive choice of VIs, robust justification is necessary to determine the most effective for the research purposes (Panda et al., 2010; Xue and Su, 2017).

Previous research has recognised the advantages of VIs in identifying rice phenology stages, generally divided into vegetative, reproductive, and ripening (Moldenhauer and Slaton, 2001; Kuenzer and Knauer, 2013; Ariza, 2019). Here, prophyll soil emergence initiates the vegetative stage, which involves significant plant height growth and leaf area expansion, alongside active tillering (Li et al., 2003; Yzarra Tito and Lopez Rios, 2011). Upon maximum tillering and stem elongation, the plant transitions to the reproductive stage, whereby energy is converted to panicle production and booting, while flowering follows the heading phase (Moldenhauer and Slaton, 2001). This marks the beginning of plant ripening, which is characterised by flowering, leaf senescence, and grain filling (Lin et al., 2014). Harvesting follows approximately 120 days after initial prophyll emergence, though this is dependent on rice cultivar and environmental conditions (Yoshida, 1981; Li et al., 2003; Kuenzer and Knauer, 2013). The array of growth stages presents altered spectral signatures, meaning rice phenology offers potential to be identified via VIs (Mosleh et al., 2015). Figure 2.1 highlights typical temporal variations in rice canopy spectra throughout development.

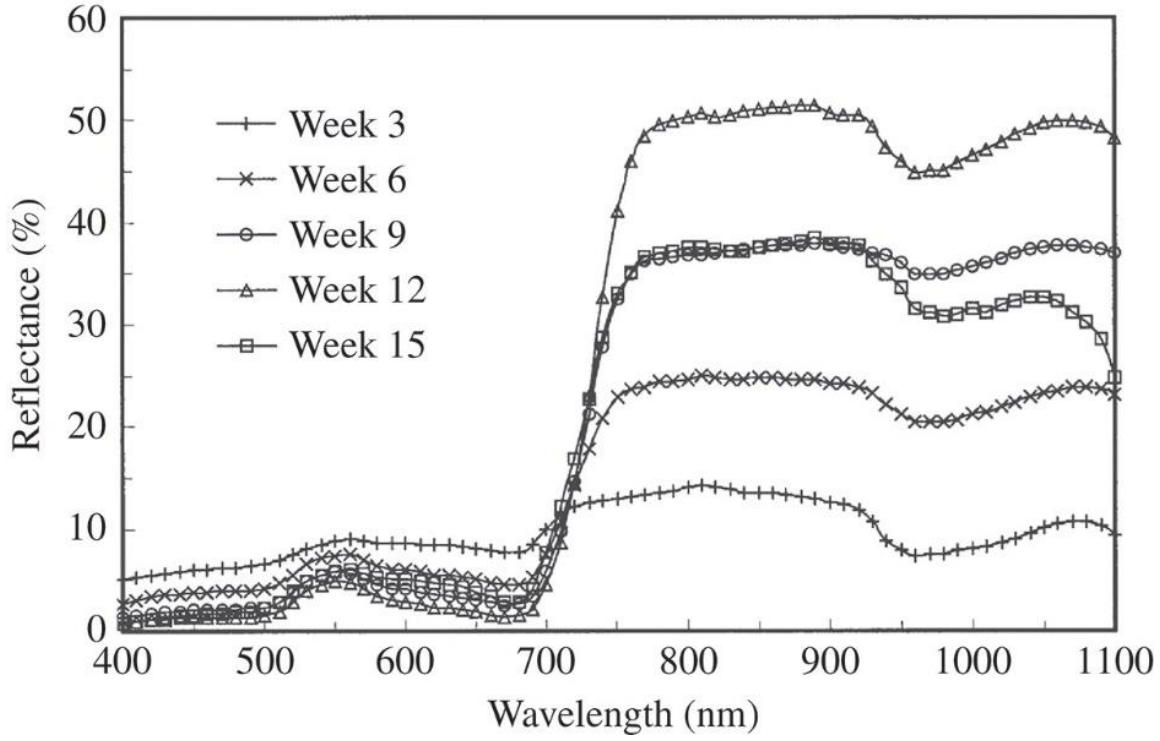

Figure 2. 1. A visual representation of the spectral variation of a typical rice canopy throughout development, outlining how such variations can be monitored to determine plant phenology (retrieved from Chang et al. (2005).

The Normalised Difference Vegetation Index (NDVI) has been used extensively during prior research to accurately forecast yield rates prior to harvesting (Rasmussen, 1997), and crop phenological stages (Ariza, 2019). The NDVI harnesses high red band spectral absorption associated with chlorophyll content, alongside strong reflectance upon the near infrared band, relating to mesophyll within the leaf structure, generating a notable 'red edge' response noticeable in Figure 2.1. (Sellers et al., 1992). Figure 2.2. provides a summarisation of rice phenological stages alongside corresponding NDVI values. Initially, lower NDVI values during germination relate to minimal vegetation. A positive correlation with NDVI and rice phenology exists following increases in vegetative density and leaf area. Here, rising chlorophyll content prompts wavelength absorption of the red band, while increased mesophyll levels following tillering and foliage development lead to heightened near infrared reflectance rates (Li et al., 2003; Yzarra Tito and Lopez Rios, 2011). As rice matures, a decreasing greenness, increasing yellowness, and reduced biomass following leaf decay result in a reduction in the NDVI (Kuenzer and Knauer, 2013; Mosleh et al., 2015). Therefore, a typical NDVI profile of healthy rice demonstrates a gradual rise, peaking at the reproductive stage approximately one to two months prior to harvest, followed by a decline as plant ripening

progresses (Kuenzer and Knauer, 2013; Mosleh et al., 2015; Ariza, 2019). This highlights how VIs can benefit PA practices, with values corresponding to plant development (Mosleh et al., 2015; Weiss et al., 2020).

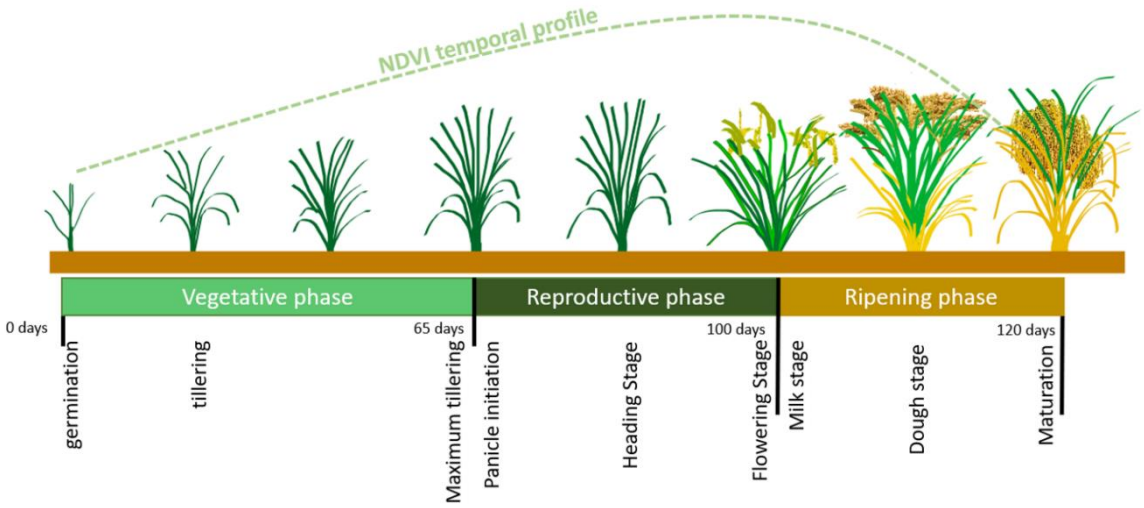

Figure 2.2. A general overview of the rice phenological stages throughout growth, namely vegetative, reproductive, and ripening, alongside corresponding NDVI value. (Modified from Kuenzer and Knauer (2013); Mosleh et al. (2015); Ariza (2019)).

Ariza (2019) utilised this relationship to investigate rice phenology identification in Colombia, generating Vis from Sentinel-2 and Landsat 7 and 8 imagery. By attaining NDVI values at 16-day intervals, Ariza (2019) predicted rice phenology with up to 72% accuracy using a Random Forest machine learning model. Additionally, Shihua et al. (2014) utilised the Enhanced Vegetation Index (EVI) to successfully establish rice growth cycles between vegetative, reproductive, and ripening stages, achieving a root mean square error (RMSE) of 10 days. Moreover, Shiu and Chuang (2019) used VIs alongside other variables for rice yield forecasting in Taiwan, applying the Support Vector Regression (SVR) machine learning method to achieve low error rates between 0.06% and 13.22%. The abundance of Vis has promoted their implementation in PA investigations, though this can be aided by other remotely sensed data, notably Synthetic Aperture Radar (SAR) (Weiss et al., 2020).

2.3. A hybrid data approach

Cloud coverage is a persistent obstacle in optical-based satellite investigations, particularly within tropical climates, causing difficulties in establishing robust data time-series', among other challenges (Filgueiras et al., 2019; Weiss et al., 2020). Data derived from airborne vehicles avoids this; Zhou et al. (2017) outlined the effectiveness of aerial-derived data for rice

detection and yield prediction, due to the diminished impact of cloud coverage. However, satellite-based remote sensing benefits from extensive open-source repositories and wider, continual coverage, often deemed more suitable in academic research, particularly within remote and developing locations (Weiss et al., 2020). Additionally, SARs cloud penetrating capabilities has opened research avenues whereby such data can be used in conjunction with optical satellite data to allow uninterrupted sensing, termed a hybrid approach (Filgueiras et al., 2019; Wu et al., 2019; Weiss et al., 2020). Table 2.1. provides a summary of prominent satellite-derived data successfully utilised during PA investigations.

Table 2.1. Details of satellite data successfully applied during previous PA investigations. Modified from Onojeghuo et al. (2018).

Wavelength	**Instrument (Agency)**	**Spatial resolution *(metres)***	**Spectral bands *(#)***	**Swath width *(km)***	**Spectral resolution *(µm)*/Waveband**	**Return period *(days)***
Multispectral	Worldview-3 (Digital globe)	1.24	28	13.1	8 MS: 0.4 - 1.04 8 SWIR: 1.195 - 2.365	<1
	RapidEye	5-6.5	5	78	VIS: 0.4 – 0.75 NIR: 0.75 – 1.3	1
	Sentinel-2A, B (ESA)	10, 20, 30 (VNIR)	13	290	VIS: 0.4 – 0.75 SWIR: 1.3 – 3.0	5
	Landsat 8 (USGS, NASA)	30	9	185	VIS: 0.4 – 0.75 NIR: 0.75 – 1.3 SWIR: 1.3 – 3.0	16
	PlanetScope	3	4	77	VIS: 0.42 – 0.70 NIR: 0.74 – 0.90	<1
Synthetic Aperture Radar (SAR)	Sentinel-1A C-Band SAR (ESA)	9, 20, 50	n/a	80, 250, 400	C-Band: 5.405 GHz; HH, VV, HH + HV, VV + VH; MW (1-100 cm)	6
	Sentinel-1B C-Band SAR (ESA)	9, 20, 50	n/a	80, 250, 400	C-Band (8-4 GHz)	6
	RADARSAT-2 (SAR) (Canadian Space Agency)	3-100	n/a	100-500	MW: 1-100 cm C-Band: 8-4 GHz	24

A hybrid approach has proven effective in prior PA investigations (Joshi et al., 2016; Filgueiras et al., 2019; Weiss et al., 2020). While optical data identifies vegetative spectral reflectance,

Synthetic Aperture Radar (SAR) data focuses on physical characteristics including leaf shape, size, structure, and water content (Woodhouse, 2005). Hybrid data can be advantageous in tropical regions such as Colombia, whereby persistent cloud coverage limits optical satellite usability (Yonezawa et al., 2012; Onojeghuo et al., 2018; Mansaray et al., 2020). Issues with SAR exist; random noise generation creates interpretative difficulties, while sensitivity to soil and vegetative water content owing to fluctuations in dielectric properties results in backscatter variability (Woodhouse, 2005; Vreugdenhil et al., 2018). However, the advantages offered by a hybrid approach are evident, whereby the combination of both optical and SAR information allow for greater exploration. Results have been encouraging in previous investigations; Filgueiras et al. (2019) demonstrated potential cloud cover mitigation during maize and soybean monitoring in Brazil, establishing a relationship between Sentinel-2 NDVI values and Sentinel-1 backscatter metrics. Filgueiras et al. (2019) trained the Random Forest (RF) algorithm with NDVI values alongside corresponding VV, VH, and normalised ratio procedure between bands (NRPB) values, allowing prediction of cloud masked NDVI values to an accuracy of R^2 0.975. Moreover, Wu et al. (2019) utilised a hybrid approach to identify rice plots damaged by typhoon-derived flooding in Zhejiang, China, where an accuracy of up to 93% was derived, again using RF. However, NDVI saturates once a vegetation density threshold is met and is an indicator of canopy greens, whereas radar backscatter responds to canopy architecture and moisture content (Woodhouse, 2005). This can result in discrepancies, whereby certain crops may in reality have low NDVI values, yet elevated backscatter values persist and cause discrepancies (Lillesand et al., 2015).

2.4. Rice yield forecasting and machine learning

Yield forecast delivery through the synthesis of multiple data sources is of consequence to numerous stakeholders, including local farmers, national governments, trade bodies, and international institutions (Kogan, 2019; Weiss et al., 2020). Further, precise yield forecasting is only increasing in importance as climate shifts become more impactful (Delerce et al., 2016). This is especially significant in less developed regions where food security can prove highly problematic (Castro-Llanos et al., 2019; Filippi et al., 2019; Kogan et al., 2019). Alongside this, agriculture in less developed countries can prove difficult to investigate, owing to increased heterogeneity, varied crop management, and smaller plot sizes (Castro-Llanos et al., 2019; Weiss et al., 2020). Complications are compounded in tropical regions, where limited seasonal variation leads to diverse sowing dates (Esquivel et al., 2018; Quevedo Amaya et al., 2019), while increased cloud coverage limits optical satellite data usability (Weiss et al., 2020).

As technology develops, the coalescence of remotely sensed data and machine learning may lead to assimilation where such issues are mitigated, perhaps allowing yield prediction much like weather forecasting (Weiss et al., 2020). This follows the ability of machine learning techniques to uncover detailed, non-linear relationships using many interconnected datasets, enabling efficient, unbiased decision-making (Chlingaryan et al., 2018). Provided is a review of investigations specifically harnessing machine learning technology for yield prediction.

Numerous machine learning algorithms have been applied to optimise rice yield prediction (Mishra et al., 2016), namely Artificial Neural Networks (ANN) (Safa et al., 2004; Ji et al., 2007; Gandhi et al., 2016), Support Vector Machines (SVM) (Jaikla et al., 2008; Ruß, 2009; Dey et al., 2017), Support Vector Regression (SVR) (Jaikla et al., 2008; Shiu and Chuang, 2019), Regression Trees (RT) (Kim and Lee, 2016; Chlingaryan et al., 2018), Random Forest (RF) (Jeong et al., 2016), and K-Nearest Neighbour (KNN) (Chlingaryan et al., 2018). However, model accuracy is strongly reliant on algorithm choice, alongside the quality of data; noise, erroneous inputs, bias, and overall relevance to the objective must be considered (Chlingaryan et al., 2018; Liakos et al., 2018). Indeed, Halevy et al. (2009) has argued that data volume is more relevant to achieving higher performance than model selection. Thus, it is important to ensure data is of sufficient quality, while an array of models are utilised to determine the most suitable for the intended research (Kim and Lee, 2016).

Chlingaryan et al.'s (2018) extensive review of machine learning for yield prediction concludes that the future of agriculture rests on synthesising remote sensing metrics, machine learning technology, and environmental variables. This has been widely demonstrated during rice yield prediction; Yaghouti et al. (2019) used VIs derived from Landsat 7 data for yield prediction in northern Iran. Using linear regression, they identified relationships using the NDVI at the end of the reproductive stage, producing an R^2 0.71 between predicted and actual rice yields. Further, their most accurately predicted variety produced a root mean squared error (RMSE) 272 kg ha-1 and normalised root mean squared error (NRMSE) 6%. Sarker et al. (2012) established that temperature and precipitation variables have a particularly significant contribution to rice yield prediction accuracy. Both variables are greatly influenced by a changing climate, demonstrating yield disparity is reliant on multiple factors (Verger et al., 2014b; Weiss et al., 2020). Moreover, this outlines the increased likelihood of model success when independent, environmental variables are synthesised with further metrics (Delerce et al., 2016; Liakos et al., 2018).

Demonstrating this, Jaikla et al. (2008) used SVR to predict rice yields with climate variables and *in situ* field measurements in Taiwan, resulting in a 2.9% error rate. Additionally, Yawata et al. (2019) used satellite data to assess approximately 3500 rice plots, with the intention of aiding national food security in Japan; using NDVI values from RapidEye and SPOT-6 satellite data, linear regression model prediction was improved by MAE 2.5% compared to conventional methods. Furthermore, Dammalage and Shanmugam (2018) used low resolution Landsat 8 and MODIS data for rice yield prediction in Polonnaruwa District, Sri Lanka for food security enhancement, primarily with the NDVI and EVI. EVI established a greater accuracy of 83.7% one month prior to harvest during heading and flowering stages, allowing accurate yield predictions approximately one month in advance.

However, simple regression techniques can be limiting, with more complex modelling valuable during PA investigations for multiple reasons. Firstly, linear regression analysis relies on an assumption of normality, whereby the sample data requires a bell-shaped distribution to avoid overt bias (Géron, 2019). Further, some relationships may be non-linear, and would resultingly not be detected via simple linear methods (Deisenroth et al., 2020). Model performance inflation can also occur from multicollinearity, whereby variables share similar relationship trends, causing model overfitting and reduced prediction accuracy with unseen data samples (Géron, 2019). Finally, by utilising more complex models, data in a categorical format can be harnessed to greater effect alongside continuous information. However, increased model complexity often comes at the cost of computational efficiency, thus a balance is necessary depending on available resources and data volume (Géron, 2019; Deisenroth et al., 2020).

Gandhi et al. (2016) utilised a more complex approach, specifically an ANN 10-fold cross-validation method for rice yield prediction in Maharashtra state, India. Using crop information, satellite data, and climate variables, a maximum prediction accuracy of R^2 0.975 was produced, alongside mean absolute error (MAE) of 0.0526 ton/ha, and RMSE of 0.1527 ton/ha (Gandhi et al., 2016). Gilardelli et al. (2019) also demonstrated the importance of data assimilation in PA, emphasising the ability of remotely sensed data to accentuate crop canopy variations at a sub-plot scale, something not possible with general climate variables. Here, rice yield prediction accuracy in northern Italy increased by over 2%, achieving an R^2 value of 0.79, through inclusion of VIs, *in situ* field information, and climatic variables.

Also demonstrating more advanced modelling, Zhang et al. (2019a) investigated rice yield forecasting in the Sahel, West Africa, with regards to climate influences. It was determined

that the biggest drivers influencing rice yield were precipitation rates, followed by maximum and minimum temperature. Here, ANN proved most successful, when compared alongside Gradient Boosting Regression (GBR) and multiple linear regression, yielding a performance of R^2 0.952, and MAE 0.115 ton/ha (Zhang et al., 2019a). The projected decreasing rainfall alongside increasing temperatures is predicted to cause a dramatic decline in yield rates in the coming years, supported by Zhang et al.'s (2019a) findings. Following this conclusion, Zhang et al. (2019a) were able to establish management adaption recommendations for continued rice cultivation, including catchment basins to fully harness precipitation, alongside adaptable irrigation techniques to maximise water availability. Such recommendations for direct farmer-level adoption are often overlooked but can be critical in establishing successful investigative outcomes (Weiss et al., 2020). Zhao et al. (2013) also examined rice management recommendations through machine learning, developing a PA system to improve yield and nitrogen use efficiency in northeast China, resulting in increased grain yield rates of 10% for local regional farmers.

Analysis of regions closely aligned to the study area is useful to establish appropriate variables. Ji et al. (2007) examined rice yield prediction in mountainous areas within Fujian province, China, comparing the accuracy of ANN and linear regression analysis. Here, ANN achieved an R^2 value of 0.67, while linear regression analysis achieved a lower R^2 value of 0.52. This investigation is of intrigue because it shares a similar climate and elevation to the present thesis' study area, albeit on a different continent. By studying research closely aligned the current body of work, knowledge gaps and influential components can be identified (Weiss et al., 2020). Thus, prior research of rice yield prediction specifically in Colombia will be explored.

2.5. Prior research in Colombia

Research surrounding Colombian rice production is significant due to its direct relation to the present thesis, generally sharing climate projections and data coverage (Pachauri et al., 2014). Quevedo Amaya et al.'s (2019) investigation researched the most effective sowing date for optimal rice yield in Colombia. Concerning 10 sowing periods between 2015 and 2016, dry matter and plant height was studied, while environmental variables pertaining to temperature, precipitation and radiation were collated from a local weather station in Tolima department (Quevedo Amaya et al., 2019). Decision Tree (DT) algorithms were harnessed, while feature significance was determined through least squares regression. Here, solar radiation demonstrated the strongest relationship to yield, concluding that sowing in May and December would maximise corresponding yield rates (Quevedo Amaya et al., 2019). Alongside sharing a

similar study area, this investigation is relevant to the present research due to its strong focus on identifying prominent variables relating to rice yield in preparation for the changing climate.

Ariza (2019) also studied within Tolima department, whereby VIs were utilised to identify rice phenology stages. This is valuable as the relatively stable climate allows rice production throughout the year, leading to varied coinciding development stages. By attaining NDVI values at 16-day intervals with Landsat 7 and 8 data, together with climatic variables and *in situ* field measurements, Ariza (2019) successfully established three machine learning approaches to predict rice growth periods: RF, SVR, and GBR, achieving overall accuracies of 71.8%, 71.2%, and 60.9% respectively. This demonstrates how phenological identification can be undertaken using a variety of data to a high correlation, though results are reliant on cloud cover relief.

Likewise, Delerce et al. (2016) assessed the relationship between both irrigated and rainfed rice yields to climate variables in Colombia, surveying production variations between cultivars and phenological stages for improved growing practices. In choosing a machine learning approach, Delerce et al. (2016) considered the presence of substantial data noise, non-linear relationships, and the need to identify the most relevant independent metrics while limiting data dimensionality. Such factors are equivalent to those faced during the present thesis. Delerce et al.'s (2016) investigation uncovered rice cultivar as the greatest influence upon yield prediction, alongside weather variables divided by phenology, with climate data contributing between 6% and 46% to yield variation. Cultivars tended to react differently to climate scenarios and phenological stages; increased temperatures at the reproductive stage negatively influenced yield in one cultivar, yet the opposite occurred during ripening, while another cultivar was negatively impacted by elevated temperatures at both the vegetative and reproductive stages (Delerce et al., 2016). Exploring this connection between cultivar, phenology, and climate is appropriate to ensure maximum future rice yields, whereby managerial recommendations can be given based upon projected weather conditions (Delerce et al., 2016).

Delerce et al.'s (2016) investigation differed from the present thesis, whereby it relied solely on climatic variables and *in situ* field measurements. This investigation will therefore build upon prior research with the addition of EO metrics, allowing further understanding of the influence cultivar, phenological stages, and climate variables have on rice yield. By highlighting previous relevant work, key findings and alternative methods can be considered,

while existing knowledge gaps can be identified and explored. Table 2.2 presents a summary of previous research deemed most pertinent to this investigation, focussing on both rice yield prediction and cloud mitigation techniques.

Table 2.2. A summary of investigations most pertinent to the present thesis regarding rice yield prediction and cloud mitigation.

Author(s)	Study Area	Sensor/Data	Methodology	Accuracy	Synopsis
Quevedo Amaya *et al*. (2019)	Armero-Guayabal, Tolima, Colombia	LAI, max diurnal temp, max night temp, min night temp, accum. solar radiation, accum. rainfall, avg. relative humidity.	DT	Overall accuracy: 92%	• LAI and solar radiation bestow greatest influence on yield. • Determined that sowing in May and December optimised yields.
Shiu and Chuang (2019)	Central Taiwan	47 variables consisting of vegetation and texture indices derived from SPOT-7, including NDVI, SAVI, RVI.	SVR, MLR, GWR	Error rate was between 0.06% and 13.22%; GWR highest performing.	• GWR performed best following feature selection (Pearson's correlation) • Required more complete satellite timeseries • Field measurements retrieved from hand sampling introduced limitations
Yaghouti et al. (2019)	Northern Iran	Landsat 7 derived VIs: NDVI, SAVI, LAI, DVI.	MLR	R^2 0.71; RMSE 272 kg ha-1; NRMSE 6%	• Identified strong relationship between yield and NDVI values at rice flowering stage.
Delerce et al. (2016)	Tolima, Colombia	*In situ* data, daily max and min temp, precipitation, relative humidity, solar radiation.	RF	Overall R^2: 0.267 and 0.502 at Saldaña and Villavicencio, respectively.	• Different cultivars and phenological stages demonstrate varied reactions to climate variables, particularly temperature.
Wu et al. (2019)	Zhejiang, China	Sentinel-1 (VV, VH); Sentinel-2 (NDVI, EVI)	RF	93%; kappa 0.9 (VH + NDVI) and 85%; kappa 0.8 (VV + NDVI)	• Rice paddy identification by merging Sentinel-1 and Sentinel-2 metrics • Combination of VIs and backscatter yielded overall accuracy of up to 93%.
Filgueiras et al. (2019)	Bahia, Brazil	Sentinel-1 (VV, VH, NRPD); Sentinel-2 (NDVI)	RF; SVM; GBM	RF: R^2 0.975, RMSE 0.036, MAE 0.020; SVM: R^2 0.920, RMSE 0.036, MAE 0.042; GBM: R^2 0.932, RMSE 0.059, MAE 0.040.	• Attempted to mitigate cloud cover of Sentienl-2 imagery using Sentinel-1 derived metrics

Chapter 2

3.0. Methods

This investigation seeks to establish a methodical approach to rice yield prediction at Hacienda El Escobal, a farm located on the Ibagué plateau in Tolima Department, central-western Colombia. This is among the 40% of agricultural land deemed suitable for rice production by 2050 following climatic changes (Castro-Llanos et al., 2019), making it crucial in maintaining national food security. The investigation will therefore provide both local and national insight into rice yield prediction, while attempting cloud cover mitigation techniques to maximise optical satellite coverage. The work will follow four fundamental steps: data pre-processing and quality assurance, derivation of relevant metrics, model selection and application, and a robust assessment of model accuracy. Using the Python programming language throughout, data pre-processing was performed in the Linux interface, while all machine learning modelling was completed using Google Collaboratory in the Jupyter notebook cloud environment. A workflow summary detailing the methodology is presented in Figure 3.1.

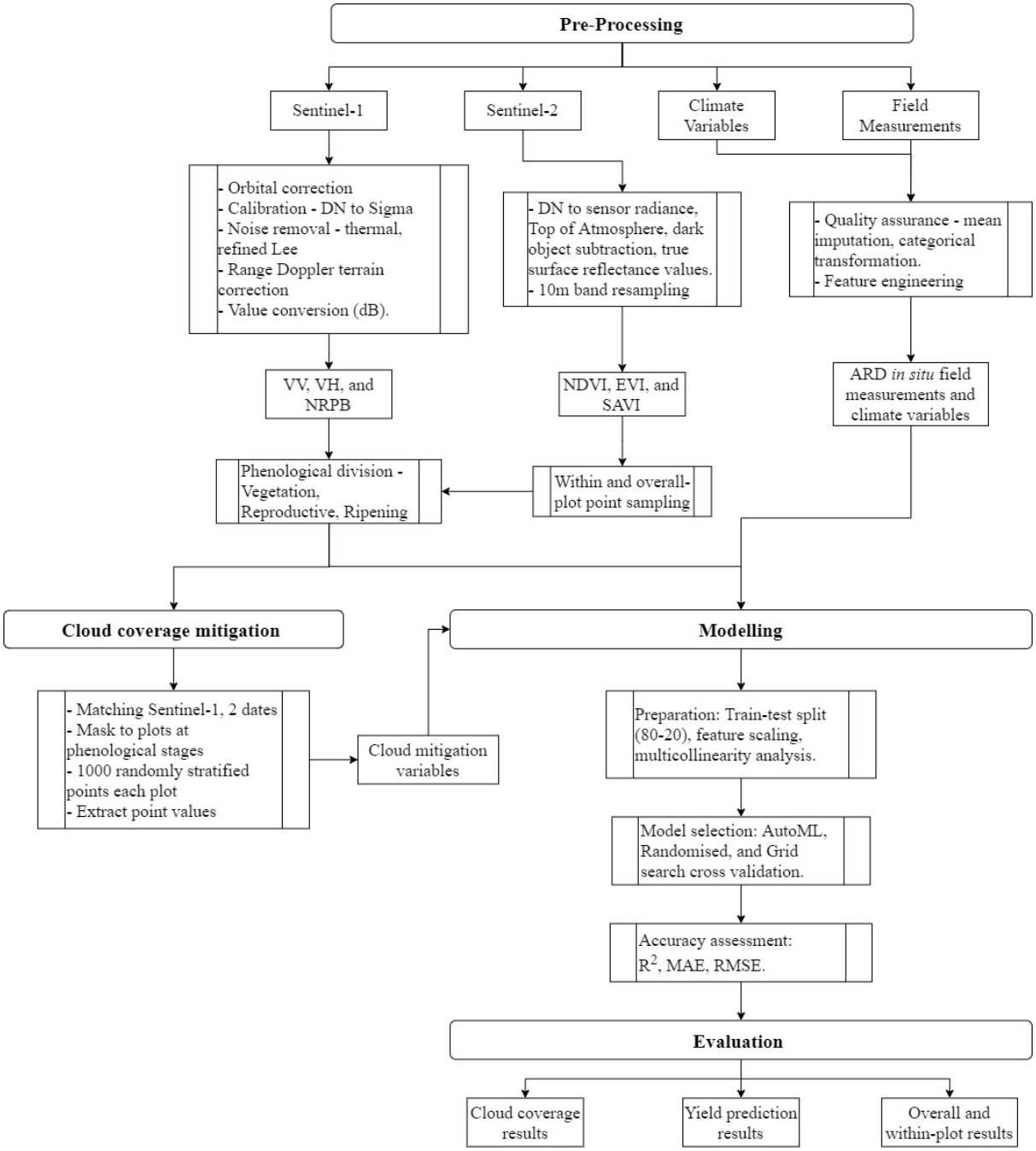

Figure 3. 1. A summarised display of the methods presented as a workflow, highlighting key stages, processes, and outputs.

3.1. Study area

Selecting an appropriate study area was dictated by several factors, namely sufficient satellite coverage, alongside accessibility to *in situ* field measurements and robust climate variables. Hacienda El Escobal, a farm located on the eastern outskirts of Ibagué, the regional capital city of Tolima Department, Colombia, facilitated these requirements. Featuring significant topographic variation with fine, inceptisol soils, Tolima is the greatest rice producing department in Colombia (Castilla-Lozano et al., 2011; Delerce et al., 2016). The tropical environment generates similar temperatures annually, while precipitation is bimodally distributed, with wet seasons across March, April, and May (MAM), and September, October, and November (SON) (Delerce et al., 2016; Esquivel et al., 2018).

Hacienda El Escobal is located within the Colombian Andes mountain range between 4.3^{o} N to 4.4^{o} N latitude and 75.2^{o} W to 75.0^{o} W longitude, situated approximately 1000 metres above sea level. Previous research indicates this will prove a suitable elevation for rice cultivation by 2050 following climatic shifts, meaning increased understanding of future yields in this area is desirable for regional and national food security (Castro-Llanos et al., 2019). The study area is located on the Ibagué plateau extending eastwards and surrounded by mountains, providing a source of rich and fertile soil. Experiencing bi-modal wet seasons means rice is cultivated throughout the year. Thus, crop phenological stages occur at varying stages dependent on plot. The study area remains irrigated via the contour-levee technique due to surrounding topography, enabling a consistent water depth across topography when available. Figure 3.2. provides a detailed overview of the study area and the specific rice plots studied during this investigation.

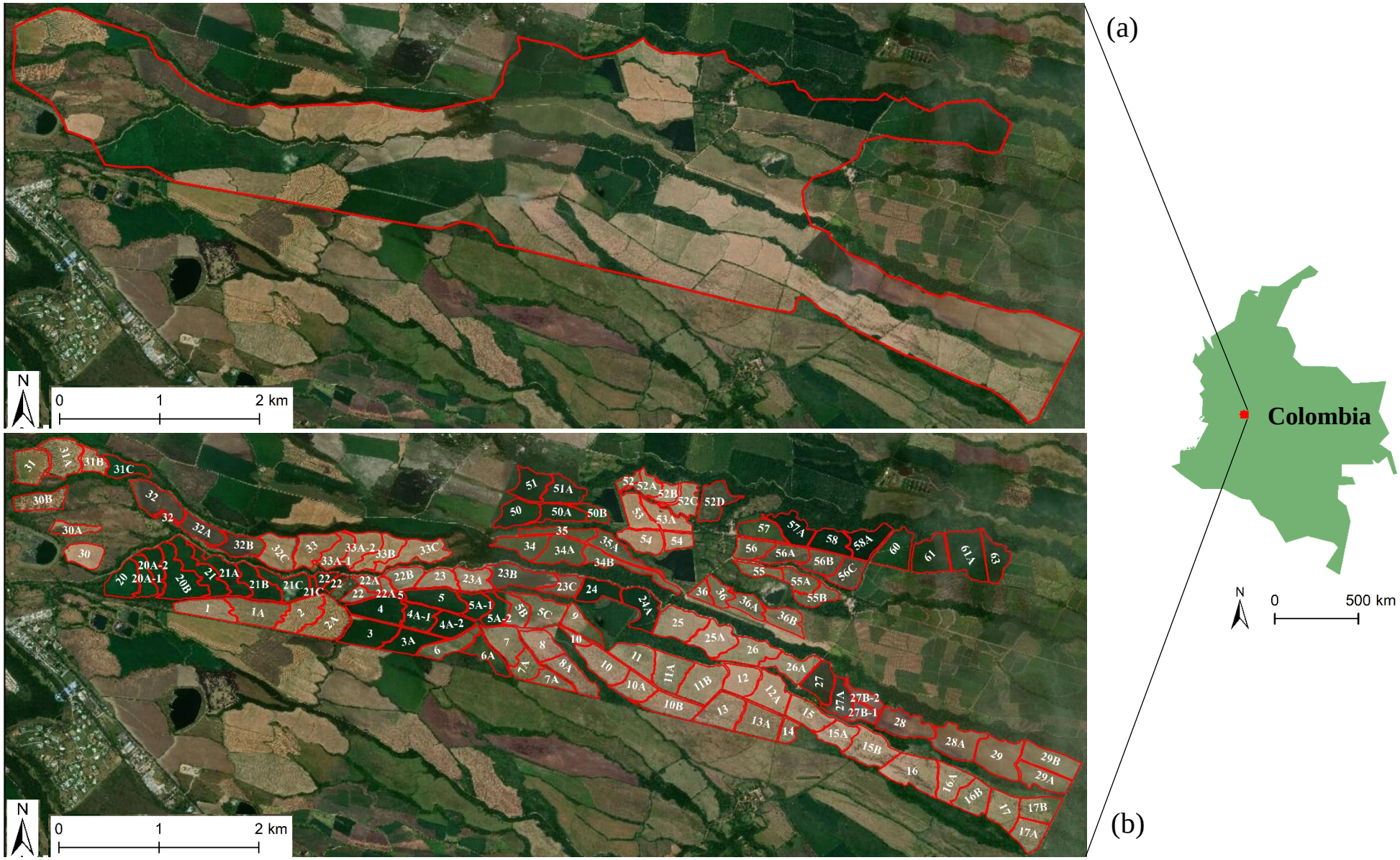

Figure 3.2. A context map of the study area presenting: (a) an outline of the area under investigation; (b) individual rice plots.

3.2. Research data

Obtaining appropriate data from robust sources was important in ensuring reliable investigative results. Some farms maintain private agricultural records useful for research purposes (Delerce et al., 2016); Hacienda El Escobal holds various records from roughly 5 harvests. Encompassing approximately 1,250 hectares of rice cultivation, the farm provides information on past yields for entire plots, alongside variables collated from GPS-mounted harvesters including sample yield data within plots (Elescobal, n.d.). Moreover, information on sowing, initial emergence, and harvest dates is available, alongside cultivar specification.

The study area receives coverage from the optical Sentinel-2 satellite, and the Sentinel-1 Synthetic Aperture Radar (SAR) satellite constellations, with revisit times of 5 and 6 days respectively (Clerici et al., 2016). This ensures sufficient spatial, temporal, and spectral coverage at suitable resolutions (Campos-Taberner et al., 2017; Weiss et al., 2020), while maintaining open-source accessibility from the European Space Agency's (ESA) Scientific Data Hub (Scihub.copernicus.eu, n.d.). Additionally, the proximity of several weather stations, alongside permissible access to historic climate simulation data (Meteoblue, n.d.), provided a thorough historical catalogue of pertinent variables. Therefore, appropriate climate data could be gathered from multiple sources for robust coverage. Some variables with multiple data sources were combined via averaging for a greater representation of the study area, while other variables were established via feature engineering. Using multiple sources was not possible in all circumstances, as some weather station data was stored monthly, while Meteoblue's simulation data related to daily periods. Such details, alongside a thorough justification of all variables used throughout the investigation, are presented in Table 3.1.

Table 3.1. *An overview of the data retrieved, pre-processed, and transformed for the purposes of rice yield prediction modelling in the study area.*

Data Type	Metric (Unit)	Description and Justification	Source
Field Measurements	Yield (kg/ha) - Within and overall-plot	346 harvests across 132 individual plots were harnessed between 2016 and 2019 for the purposes of this investigation, allowing collection of within-plot data from harvester samples, alongside overall plot yield data. Yield was generated by dividing plot production (kg) by plot area (hectares) data, retrieved from *in situ* field measurements supplied by Hacienda El Escobal.	Hacienda El Escobal
	Cultivar	Cultivar variation is evident in the dataset, notably Escobal 312, Escobal 417, Escobal 518, Fedearroz 67, Fedearroz 68, Maja 6, Orizica, Panorama 394, and Triunfo. Cultivar has a significant impact on yield rates, owing to crop height, vegetative density, and growth period (Kuenzer and Knauer, 2013; Delerce et al., 2016; Zhou et al., 2017).	
	Season	Owing to the impact of Tolima's bimodal seasonal variation on crop yield (Delerce et al., 2016; Esquivel et al., 2018), data was divided based upon sowing date into December-January-February (DJF), March-April-May (MAM), June-July-August (JJA), and September-October-November (SON).	
	Seedling stage	Using field measurements, seedling stage is time from sowing to emergence. Seedling stage is an important determinant of rice yield rates in Colombia (Delerce et al., 2016; Quevedo-Amaya et al., 2020), and elsewhere amongst other cereal crops (Shivrain et al., 2009; Wang et al., 2009).	
	Growth time	Refers to the period between crop emergence and harvest. (Thippani et al., 2017)	
Earth Observation	NDVI Vegetative NDVI Reproduction NDVI Ripe	The Normalised Difference Vegetation Index (NDVI) is a widely used metric during agricultural investigations owing to its strong correlation to vegetative presence and rice yield prediction capabilities (González-Betancourt and Mayorga-Ruíz, 2018; Shiu and Chuang, 2019; Yaghouti et al., 2019; Wu et al., 2019). Additionally, the NDVI, in combination with the EVI and SAVI, were harnessed by Munibah et al (2019) while investigating rice growth phases and yield relationships using Sentinel-2 data.	Sentinel-2 data (Appendix A).
	EVI Vegetative EVI Reproduction EVI Ripe	The Enhanced Vegetation Index (EVI) can prove beneficial in agricultural monitoring, owing to its lessened impact from saturation experienced by the NDVI from high density vegetative coverage (Zhang et al., 2019b). Recent use of the EVI during rice cultivation further highlights its effectiveness (Munibah et al., 2019; Wu et al., 2019).	
	SAVI Vegetative SAVI Reproduction SAVI Ripe	The Soil Adjusted Vegetation Index (SAVI) corrects any influence of soil brightness from areas of low vegetation, a useful characteristic while investigating crops at varied growth stages (Huete, 1988). SAVI has also been successfully used for similar purposes to the current investigation (Shiu and Chuang, 2019; Yaghouti et al., 2019).	
Earth Observation	Sentinel-1 VH Polarisation	Metrics derived from Sentinel-1 data for application to cloud mitigation research. The implementation of VV, VH, and NRPB index follows the methodology of Filgueiras's (2019) investigation into cloud mitigation in Brazil by establishing a	Sentinel-1 data retrieved from

	Sentinel-1 VV Polarisation Sentinel-1 NRPB	relationship to the NDVI, and prior links between backscatter rations and VI values in agricultural settings (Veloso et al., 2017).	Scihub.copernicus.eu (n.d.).
Climate variables	Precipitation (mm)	The cumulative amount of precipitation recorded at the study area throughout the growing period. Precipitation is a key determinant in other machine learning-driven rice yield research, owing to its projected decline as the climate changes (Gandhi et al., 2016; Quevedo Amaya *et al.,* 2019; Zhang et al., 2019a).	Meteoblue (n.d.)., Perales Airport Weather Station
	Average temperature (ºC)	The average temperature recorded throughout the growing period. An increase in temperature is a key projection of the changing regional climate, causing significant impact upon yield rates (Quevedo Amaya *et al.,* 2019; Zhang et al., 2019a).	
	Wet frequency (days	Calculated as the number of days whereby daily precipitation surpasses the precipitation average for each given month through the growing period. Wet day frequency has been found to impact rice yield variability more significantly than cumulative precipitation amounts in prior research (Revadekar and Preethi 2012; Fishman 2016), while Fernandes et al. (2020) also found that wet day frequency is a better yield indicator than total precipitation amounts, albeit only between June and August.	Meteoblue (n.d.).
	Maximum and Minimum Temp. (ºC)	The maximum and minimum recorded temperatures throughout the growing period. The variables proved a beneficial input in previous work regarding rice yield prediction (Sumith et al., 2002; Maruyama, 2013), including Tolima Department, Colombia (Quevedo Amaya *et al.,* 2019).	
	Drought frequency (days)	Number of days per month where no precipitation was recorded, measured throughout the growing period. Provides an alternative avenue to wet frequency, whereby the impact of decreasing rainfall can be evaluated, a key element of the changing climate in the region (Heinemann and Sentelhas, 2011; Heinemann et al., 2015).	
	Average relative humidity (%)	The average recorded humidity recorded at the study area throughout the growing period. Beneficial input in previous work regarding rice yield prediction (Sumith et al., 2002; Maruyama, 2013), including Tolima Department, Colombia (Quevedo Amaya *et al.,* 2019).	
	Sunlight (mins)	The cumulative amount of sunlight in minutes received by the study area over the course of the growing period. Beneficial in previous work regarding rice yield prediction in Tolima Department, Colombia (Quevedo Amaya *et al.,* 2019).	

3.3.0. Data Pre-processing

Prior to modelling, data required pre-processing and preparation. This section provides information on these steps.

3.3.1. Sentinel-2 Pre-processing

Sentinel-2 data was retrieved from the European Space Agency's (ESA) Scientific Data Hub (Scihub.copernicus.eu, n.d.), covering the period where *in situ* field measurements were available between 2016 and 2019. Sentinel-2 data required conversion to Analysis Ready Data (ARD), achieved using Atmospheric and Radiometric Correction of Satellite Imagery (ARCSI) software (Bunting, 2014). This process involved the conversion of Digital Number (DN) pixel values to sensor radiance, Top of Atmosphere (ToA) reflectance generation, and application of the Dark Object Subtraction (DOS) method (Chavez, 1996) to transform data to surface reflectance.

Raw satellite data is supplied as DNs, these being pixel values prior to conversion to a quantitative measure such as radiance (Mather and Koch, 2011). Radiance is understood as a measure of wavelength energy in watts obtained by a sensor, radiated by a unit area, per solid angle of measurement, per nanometre; this equation is abbreviated to $W \cdot sr^{-1} \cdot nm^{-1}$ (Mather and Koch, 2011), calculated as follows:

Eq. (1)

$$L\lambda = \left(\frac{LMAX\lambda - LMIN\lambda}{Qcal\ min - Qcal\ max}\right)(Qcal - Qcal\ min) + LMIN\lambda$$

Conversion to ToA reflectance is necessary as radiance energy and at-sensor measurements diverge, owing to different solar zenith angles following time variations between data acquisitions, alongside disparities in solar irradiance from spectral band sensor differences (Lillesand et al., 2015). The equation follows:

Eq. (2)

$$\rho\lambda = \frac{\pi \cdot L\lambda \cdot d^2}{ESUN\lambda \cdot \cos(\theta s)}$$

Establishing surface reflectance mitigates pixel value distortion from radiation scattering and absorption, a result of atmospheric aerosols and water vapour (Bunting, 2014). DOS adopts the

premise that the darkest pixel values from each image band are the product of atmospheric scattering, requiring removal (Chavez, 1996). Following this, Sentinel-2 data was considered an accurate representation of surface reflectance.

Owing to the study area's tropical location, cloud masking was necessary to remove distorted imagery (Filgueiras et al., 2019). Following Wang et al.'s (2016) data fusion methodology, band resolution was resampled from 20 m to 10 m via the nearest neighbour method using ARCSI, with bands subsequently sharpened using a 7x7 pixel filter for band specific linear regression (Bunting, 2014). Additionally, bands 1, 9, and 10 were excluded from ARD as they are typically used for atmospheric corrections (Wang et al., 2016). Table 3.2. details the resulting Sentinel-2 band information used during this investigation.

Table 3.2. Sentinel-2 band information in ARD format, following all necessary pre-processing measures.

Band Number	Band Details	Central Wavelength (nm)
1	Blue	496.6
2	Green	560
3	Red	664.5
4	Red 5	703.9
5	Red 6	740.2
6	Red 7	782.5
7	NIR 8	835.1
8	NIR 8a	864.8
9	SWIR 1	1613.7
10	SWIR 2	2202.4

3.3.2. Vegetation Indices preparation

Determining the most suitable VIs required careful consideration, with the NDVI, EVI, and SAVI deemed the most appropriate for the investigation, owing to their success in past research in comparable locales (Panda et al., 2010; Cheng and Wu, 2011; Noureldin et al., 2013; Son et al., 2014; Zhou et al., 2017; González-Betancourt and Mayorga-Ruíz, 2018; Shui and Chuang, 2019; Yaghouti et al., 2019; Wu et al., 2019). The selection of the NDVI, EVI, and SAVI was reinforced by Munibah et al (2019), who opted for these indices during a similar investigation into rice yield prediction in climatically comparable West Java, Indonesia, using Sentinel-2

data. Furthermore, Duan et al. (2019) successfully used these indices during rice yield prediction with airborne data, demonstrating their value for possible extrapolation.

VIs were batch-generated within the Linux interface using Python RSGISLIB (Bunting et al., 2014). Sentinel-2's band 8a was designated the suitable NIR band during VI calculations, owing to its compatibility with other sensors (Zhang et al., 2018), alongside increased precision in previous PA investigations compared to band 8 (Li et al., 2017a; Zhang et al., 2017). Further, its application in VIs has shown less inclination to saturation, providing more valuable data (Tesfaye and Awoke, 2020). The NDVI, EVI, and SAVI were calculated using the equations presented in Table 3.3., where *Red* (664.5 nm), *NIR* (864.8), and *Blue* (496.6 nm) are Sentinel-2 bands 3, 8, and 10, respectively.

Table 3.3. The VIs harnessed during the investigation, alongside specific formulation, and band information.

Vegetation Index	Equation	Band wavelengths (μm)	Justification
NDVI (eq.3)	$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$	$NDVI_{B8a}$ (864.8, 664.5)	Rouse et al. (1973)
EVI (eq.4)	$EVI = 2.5\,\frac{(NIR - Red)}{\left((NIR + C1\,Red - C2\,Blue) + L\right)}$	EVI_{B8a} (864.8, 664.5, 496.6, C1 = 6, C2 = 7.5, L = 1, G = 2.5)	Liu and Huete (1995)
SAVI (eq.5)	$SAVI = \frac{(NIR - Red)}{(NIR + Red + L)(1 + L)}$	$SAVI_{B8a}$ (864.8, 664.5, L = 0.5)	Huete (1988)

An example of generated indices is presented in Figure 3.3., displaying NDVI, EVI, and SAVI values from the study area on 31st October 2018.

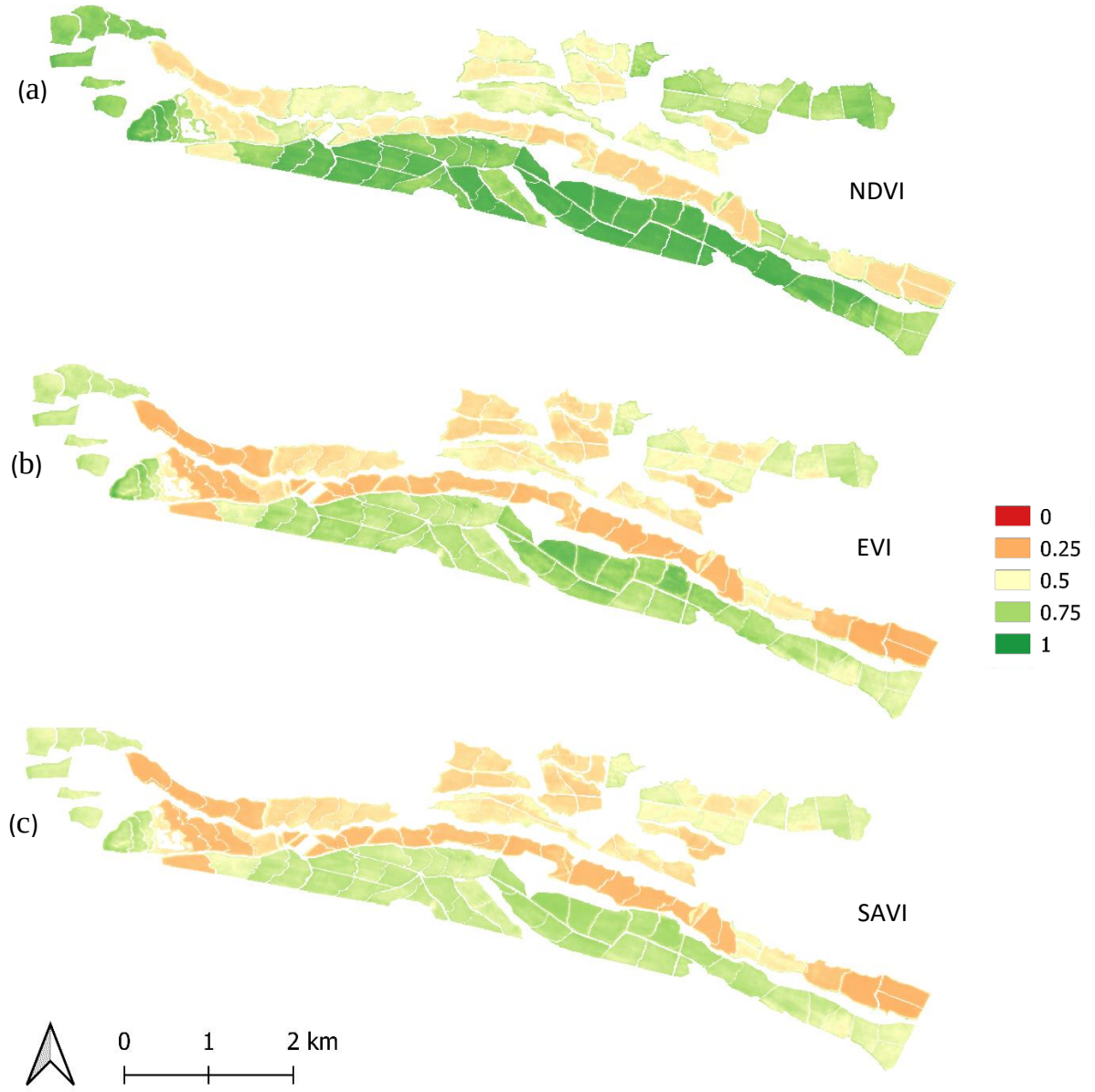

Figure 3.3. An overview of generated VIs covering all plots within the study area, specifically (a) NDVI, (b) EVI, and (c) SAVI. Variation in phenological stages is evident. Generated from Sentinel-2 imagery captured on 31st October 2018

Correlations between VI values and yield rates were assessed via two methods, namely an overall and within-plot approach. An overall-plot approach collated average NDVI, EVI and SAVI values at the vegetative, reproductive, and ripening phenological stages, which were then compared to *in situ* overall plot yield data for each harvest. Phenology division has proven valuable during past research assessing rice yield prediction (Delerce et al., 2016; Ariza, 2019).

The alternative within-plot method retrieved the same VI values from phenological stages, though these were gathered from individual *in situ* GPS sample points, corresponding to locations of harvester sample yield collections. Resultingly, each sample had associated index values. Necessary data preparation was achieved as follows; after generation and phenological division, values were collated from each index using QGIS point sampling tool from GPS sample points supplied by the Hacienda El Escobal, sufficiently covering each pixel. Using Panek and Gozdowski's (2020) method, overall-plot data was then generated by averaging all VI point values per plot. Contrastingly, within-plot data was gathered from all harvester sampling points again using QGIS point sampling tool, meaning each sample weight had corresponding VI values.

Phenological division was crucial during analysis, achieved by stacking all Sentinel-2 scenes from a specific plot throughout a growing season according to *in situ* field measurements, enabling a time-series analysis of key growth stages (Ariza, 2019). Figure 3.4. provides a representation of this, whereby stacked VI values are generated to allow analysis of crop phenological development. Information on emergence and harvest dates for individual plots allowed a precise overview of crop development stages. Scenes were assigned as either vegetative, reproductive, or ripening, encompassing roughly 65, 35, and 20 days respectively, though this varies by cultivar (Kuenzer and Knauer, 2013; Mosleh et al., 2015). Stages captured with multiple times by satellite were combined via averaging, while plots with insufficient coverage were removed. Where only one phenology was unobtainable, an average value was imputed using corresponding plots of the same cultivar and development stage, with Scikit-Learn's 'SimpleImputer' function (Pedregosa et al., 2011).

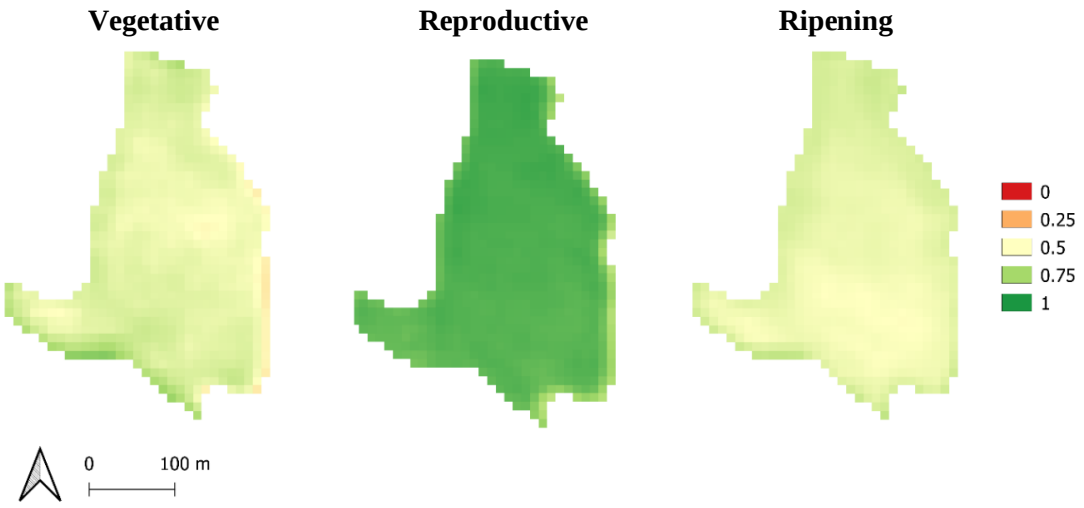

Figure 3.4. A representation of NDVI data masked to plot 27a and divided to three phenological stages, namely vegetative, Reproductive, and Ripening.

Using this combination of *in situ* data with typically observed rice phenological sequencing (Kuenzer and Knauer, 2013; Mosleh et al., 2015; Ariza, 2019), an accurate representation of crop development was established in the study area. This is demonstrated in Figure 3.5., which details plotted NDVI values to demonstrate the trend of rice growth from the vegetative stage, peaking at the reproductive stage, and beginning to fall upon ripening (Kuenzer and Knauer, 2013; Mosleh et al., 2015; Ariza, 2019). This trend was considered to accurately divide data by phenology.

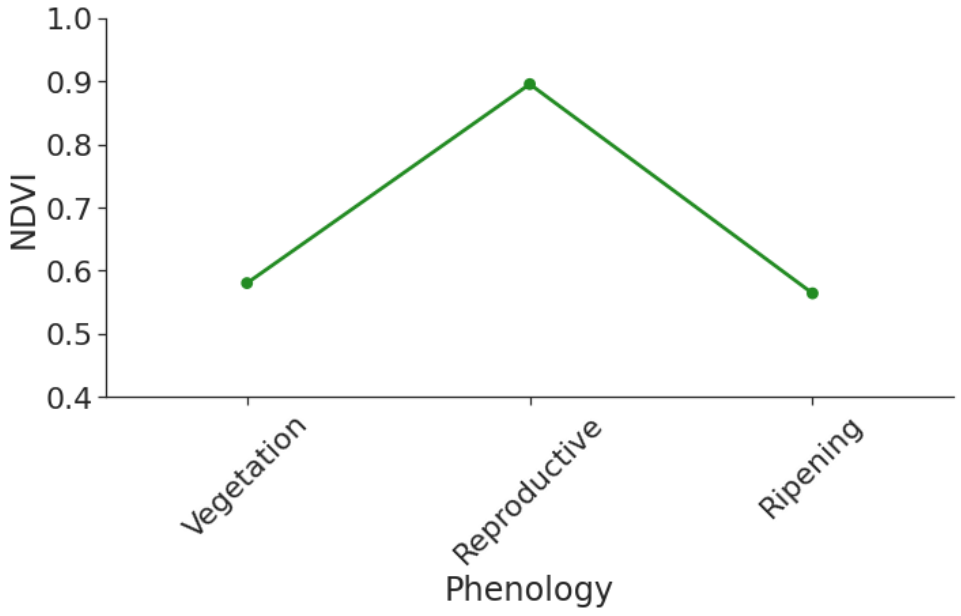

Figure 3.5. A plotted line graph detailing the corresponding NDVI values to phenological stage plotted from stages from plot 27a in Figure 3.4. This rising, peaking, and falling trend is typical rice response to NDVI and other VIs as detailed is prior investigations (Kuenzer and Knauer, 2013; Mosleh et al., 2015; Ariza, 2019).

3.3.3. Sentinel-1 Pre-processing

Sentinel-1 data required pre-processing to avoid error propagation during cloud mitigation analysis (Filgueiras et al., 2019). Following Filgueiras et al.'s (2019) methodology, Sentinel-1 data captured on the same day as corresponding Sentinel-2 imagery was identified and retrieved in level 1 Ground-Range Detected (GRD) format, producing 8 matches (Appendix A). Pre-processing was achieved using ESA's Sentinel Application Platform (SNAP) Sentinel-1 Toolbox version 7.0.3., with steps automated using the Graph Builder (SNAP, n.d.). Both Vertical Vertical (VV) and Vertical Horizontal (VH) SAR polarisations were harnessed to determine their effectiveness during modelling, alongside a generated normalised ratio procedure between bands (NRPB) index, for additional correlation testing (Filgueiras et al., 2019).

Firstly, orbital data correction, automatically retrieved from ESA's product metadata, was assigned to each image, allowing positioning between the orbital track and sensor. DNs were radiometrically calibrated and normalised, while backscatter image intensity values were converted to sigma nought, which refers to reflective strength in terms of the geometric cross section of a conducting sphere that would give rise to the same level of reflectivity (Filgueiras et al., 2019; Filliponi, 2019; Truckenbrodt et al., 2019). Noise removal followed, whereby thermal-induced disparities to pixel data were normalised throughout each scene (Filliponi, 2019). Additional granular noise resulting from scattering was mitigated with the Refined Lee filter using KNN to determine average values across neighbouring pixels, and removing anomalous values (Yommy et al., 2015; Filliponi, 2019).

Range Doppler Terrain Correction mitigated geometric distortion from Sentinel-1's side-looking method of data capture; by applying the SRTM 1 Second HGT Digital Elevation Model (DEM), pixel distortions were rectified (Fillioponi, 2019). This was performed using bilinear interpolation resampling, owing to use in prior investigations (Truckenbrodt et al., 2019), and aligned to the target Coordinate Reference System (CRS), UTM Zone 18 / World Geodetic System 1984. Data was converted from a unitless coefficient to decibels (dB) through a logarithmic transformation procedure, allowing comparison to VI values (Fillioponi, 2019).

3.4. Cloud cover mitigation

Cloud coverage persists in the study area's tropical climate, causing optical data voids. To mitigate information gaps and enable improved rice monitoring, attempts to determine correlations between VI data and backscatter metrics were explored, following prior success (Filgueiras et al., 2019). Filgueiras et al. (2019) identified a strong relationship between NDVI and backscatter metrics in Brazil, with an R^2 value of 0.98 (Filgueiras et al., 2019). Therefore, this was replicated with subsequent pre-processing steps detailed in Section 3.3.3. VV and VH polarisations were harnessed to generate the NRPB index, due to its relationship to VIs in agricultural settings (Veloso et al., 2017; Vreugdenhil et al., 2018; Filgueiras et al., 2019). NRPB is calculated via:

Eq. (6)

$$NRPB = \frac{(\sigma VH - \sigma VV)}{(\sigma VH + \sigma VV)}$$

Following this, VV, VH, and NRPB variables were evaluated to determine their correlation to VI values of the same capture date. Variables were stacked by capture data, ensuring pixels

from each metric covered the same respective area using the 'Set output file resolution' QGIS tool (de Leeuw et al., 2014). An example of all six variables from 31st October 2018 is displayed in Figure 3.6.

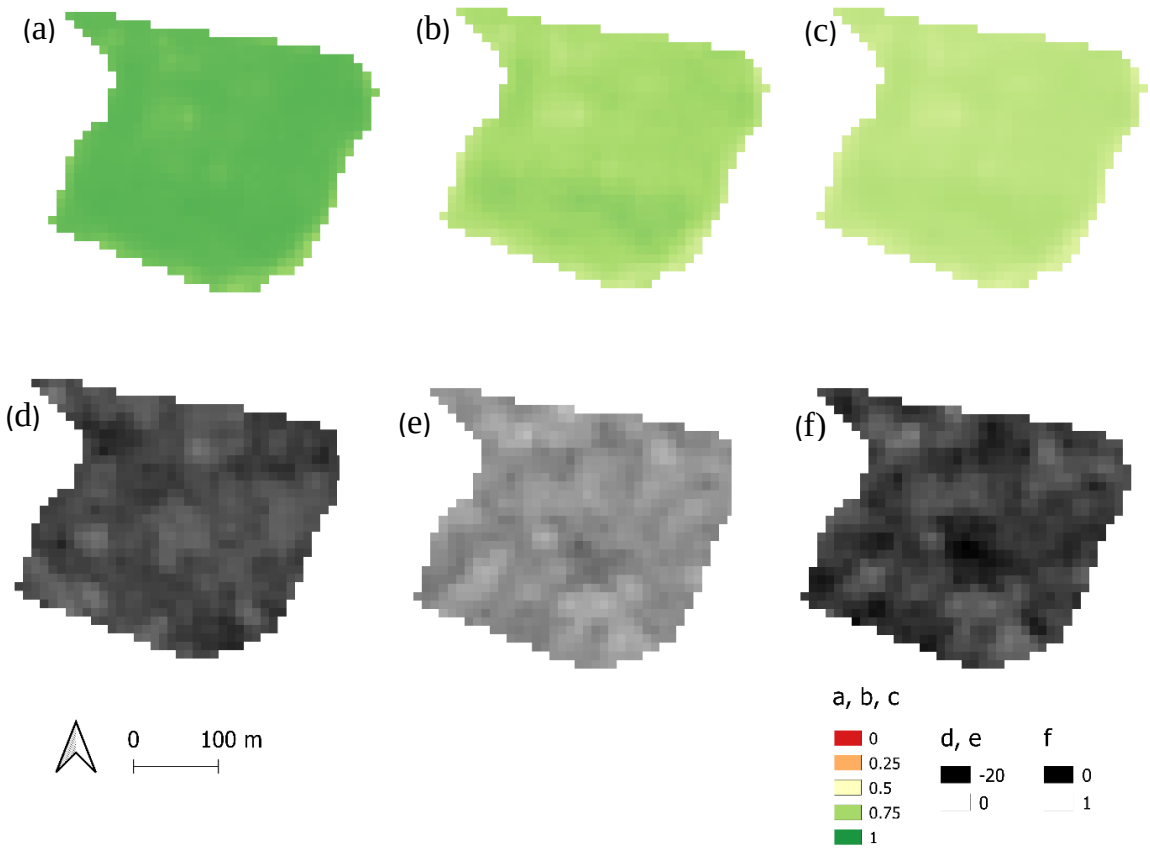

Figure 3.6. An example of each variable used for cloud mitigation research: (a) NDVI, (b) EVI, (c) SAVI, (d) VH, (e) VV, and (f) NRPB. All variables have been clipped to Plot 12, whereby values can be retrieved from generated points for correlation modelling. Data capture date was 31st October 2018.

1000 randomly stratified points per plot were generated, with a negative 10 m buffer to eliminate boundary and GPS error. Point values from variables were extracted via QGIS's sample raster values tool. Following this, collated values were arranged by VI and phenological stage, allowing further analysis through machine learning techniques. Previous PA research with SAR data reinforces this, where the best performances were attained using a machine learning approach (Sivasankar et al., 2018; Filgueiras et al., 2019).

3.5.0. Modelling process

Several factors were considered when determining the appropriate modelling approach for this investigation, including: the presence of erroneous data, likely non-linear relationships between variables, high multicollinearity, and identifying the most impactful data (Delerce et al., 2016).

The following section details the resulting modelling process, concerning both cloud mitigation and rice yield prediction.

3.5.1. Quality assurance and pre-processing

Quality assurance is critical when preparing raw data for machine learning to avoid inaccuracies associated with poor data structure and erroneous features (Géron, 2019). Using Scikit-Learn's 'SimpleImputer' function (Pedregosa et al., 2011), columns with missing data were replaced with mean values from available information. In addition, categorical variables required transformation into a numerical format for model compatibility (Géron, 2019; Deisenroth et al., 2020). Scikit-Learn's 'OneHotEncoding' function (Pedregosa et al., 2011) was harnessed, which converts categorical strings to numerical values in accordance with category (Deisenroth et al., 2020).

Prior to model implementation, data was separated into training and testing sets for dependent and independent variables. While the training data is used to fit the model, test data allows model assessment with an unseen data selection. This enables assessment of model generalization, this being its ability to adapt to unseen datasets (Deisenroth et al., 2020). Using Scikit-Learn's 'train_test_split' function (Pedregosa et al., 2011), values were split 80% training and 20% testing, allowing robust performance analysis (Deisenroth et al., 2020). Further, feature scaling was required to merge all variables to the same planar scale, crucial to avoid algorithms automatically assuming higher values have elevated influence, while also allowing faster convergence for increased computational efficiency (Deisenroth et al., 2020). Using Scikit-Learn's 'StandardScaler' (Pedregosa et al., 2011), variables were transformed to scale, computing mean and standard deviation upon the training set to implement upon the test set.

3.5.2. Variable multicollinearity and feature selection

Multicollinearity relates to substantial dependency between variables, skewing model performance (Chlingaryan et al., 2018; Géron, 2019). This causes potential inflated error if multiple variables are similar, increased model variance, and substantial dependence on limited data, termed overfitting (Dormann et al., 2013; Chlingaryan et al., 2018). By mitigating multicollinearity, feature selection generally allows faster and more efficient computation, resulting in a more robust model (Chlingaryan et al., 2018; Géron, 2019).

Variable multicollinearity analysis utilised Pearson's correlation coefficient of determination. Positive 1 signifies total linearly positive correlation, negative 1 displays total linearly negative

correlation, while 0 signals no linear collinearity. Values greater than positive and negative 0.7 were considered eligible for removal owing to significant multicollinearity (Dormann et al., 2013). The equation is as follows:

Eq. (7)

$$r = \frac{n(\sum xy)(\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

As Pearson's coefficient only considers linear relationships, an alternative method was necessary; the Variance Inflation Factor (VIF) examines multicollinearity through ordinary least squares regression analysis (Deisenroth et al., 2020). This measures variance increase based upon multicollinearity between variables (Deisenroth et al., 2020). A score threshold of 10 was used, with greater values indicating significant multicollinearity requiring removal (Dormann et al., 2013). The VIF equation follows:

Eq. (8)

$$VIF = \frac{1}{1 - R_i^2}$$

3.5.3. Model selection

Initial model selection employed successful algorithms from prior investigations for preliminary performance analysis. Moreover, an Automated Machine Learning (AutoML) approach, specifically the TPOT library (Olson and Moore, 2016), was used to gain concealed insight into model hyperparameters and feature selection criteria. TPOT's stochastic nature allows an initial overview of models to explore further (Olson and Moore, 2016).

Alongside TPOT, the following models were implemented using Scikit-Learn library (Pedregosa et al., 2011) with default settings to develop a general performance overview: Simple Linear Regression (SLR), Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), Gradient Boosted Regression (GBR), and Extreme Gradient Boosting (XGBoost). SLR is applied where only one independent variable was present, where ordinary least squares with two-dimensional data was harnessed to determine the linear function best predictive of the dependent variable (Géron, 2019). Where multiple independent variables were available, MLR was used, following the same approach but with dimensions in accordance with the number of variables. The RF ensemble regression approach used decision

trees to establish the most accurate performance metrics (Brieman, 2001; Chlingaryan et al., 2018). Additionally, SVR predicts planar data distributions through conversion of non-linear data to higher dimension feature space (Smola and Schölkopf, 2004). GBR is an additive machine learning approach, whereby model performance is optimised through error rates of the gradient loss function for a given number of weak prediction models (Géron, 2019). XGBoost is a modified version of GBR, which follows the same scalable process, but with increased speed (Chen and Guestrin, 2016).

The most successful preliminary models were tuned using Scikit-Learn's 'RandomizedSearchCV' function (Pedregosa et al., 2011), a grid of hyperparameters defined and randomly sampled via K-fold cross validation by a given number of iterations. Values were further explored via Scikit-Learn's 'GridSearchCV'. This differs from RandomizedSearchCV by evaluating all defined hyperparameter combinations, allowing a more focussed approach. Figure 3.7. visualises of both cross-validation methods, outlining how both in conjunction with one another can maximise model performance.

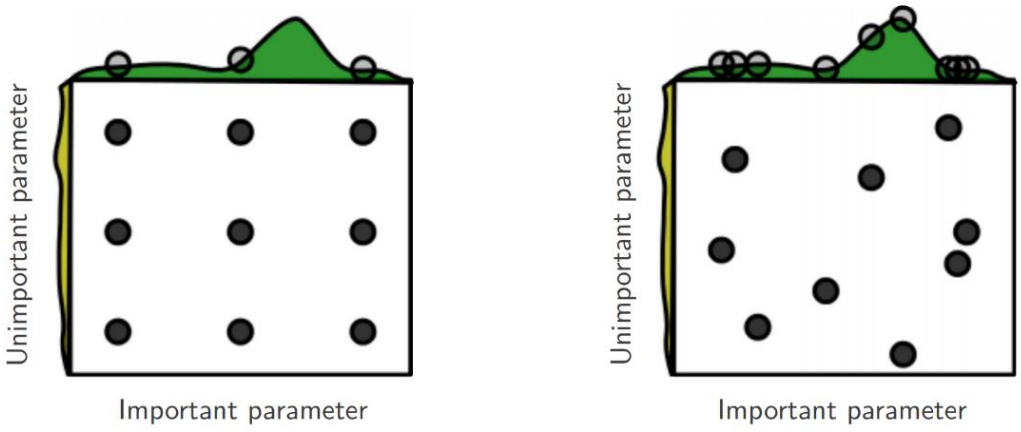

Figure 3. 7. A diagram displaying (a) the grid search cross validation and (b) randomised search cross validation. This outlines how the two methods can produce varied results. Figure retrieved and modified from Bergstra and Bengio (2012).

3.6. Model accuracy assessment

Modelling is essentially a representation of reality through observed data, meaning residual error is an inevitable consequence and can be used to determine accuracy (Spiegelhalter, 2019). This is achieved by effectively measuring the distance between the predictor variable vector and the target value vector (Géron, 2019). Following division between training and testing,

robust accurate assessment of training data could be achieved using unseen values (Deisenroth et al., 2020). Consequently, various accuracy metrics were generated for rigorous model performance analysis.

Ranging between 0 and 1, the R^2 coefficient of determination measures variability of predicted values in relation to actual data (Géron, 2019; Deisenroth et al., 2020). An R^2 greater than 0.6 indicates high correlation, though error values should be considered (Alexander et al., 2015; Iizuka et al., 2020). R^2 is calculated via the following equation, y being the independent variable, $\hat{y}$ referring to the predicted y value, and $\bar{y}$ the mean value of y:

Eq. (9)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2}$$

The Mean Absolute Error (MAE) is alternative performance metric that gives little credence to large outlier errors (Deisenroth et al., 2020), calculated via:

Eq. (10)

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

The Root Mean Squared Error (RMSE) provides information on model error rates, with increased sensitivity to large outlier errors (Deisenroth et al., 2020). The RMSE is calculated through the following equation:

Where n refers to the number of instances, y_i refers to the vector of values of the i^{th} instance in the dataset, and $\hat{y}$ is the predicted value of y.

Eq. (11)

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

Chapter 3

4.0. Results

4.1. Cloud cover mitigation

A summary of cloud mitigation results is displayed in Table 4.1. Following multicollinearity analysis, VV alongside the NRPB index proved the most appropriate independent variables. Consequently, the VH polarisation was removed from further analysis due to a weaker correlation, likely due to differing crop canopy reactions compared to VV (Woodhouse, 2005). Additionally, NDVI values demonstrated the strongest relationship to Sentinel-1 metrics from the three VI's among all phenological stages. XGBoost was generally the best performing algorithm, during both vegetative and reproductive stages.

Research revealed notable correlations within vegetative and reproductive phenological stages. Although the vegetative stage displayed the highest R^2 output, analysis of the reproductive stage proved more consistent, generating an R^2 value of over 0.530 for each model, while maintaining minimal RMSE and MAE values. NDVI was the most effective VI predictor, whereby results of 0.583 (R^2), 0.114 (RMSE), and 0.155 (MAE) for vegetative stage and 0.578 (R^2), 0.016 (RMSE), and 0.020 (MAE) for reproductive stage were generated using XGBoost. These results present a relatively consistent correlation, with specified SAR metrics accounting for almost 60% of NDVI variation approximately one month prior to harvest. This demonstrates the possibility for optical cloud mitigation during this crop development period. However, further error reductions and correlation analysis is necessary before wider implementation (Alexander et al., 2015).

Elevated MAE and RMSE values during the vegetative phase perhaps stem from the phenology's timeframe, covering a period of development longer than the reproductive and ripening stages combined (Kuenzer and Knauer, 2013; Mosleh et al., 2015). The vegetative phenology also sees large structural change, from initial prophyll emergence to canopy closure during maximum tillering and stem elongation, meaning Sentinel-1's high backscatter response to plant architecture likely caused substantial error variations. Further, fluctuating dielectric properties responding to soil moisture adjustments likely contributed to error, whereby early-stage vegetative would be dominated by soil, while later-stage vegetative following canopy closure would diminish its impact on backscatter (Bousbih et al., 2017; Filgueiras et al., 2019). Consequently, a higher variation in backscatter values during the more precise stages of the

vegetative phenology, namely germination, tillering, and stem elongation (Zheng et al., 2016), may have triggered elevated error.

Contrastingly, soil and ripening stages performed poorly; soil garnered a maximum R^2 value of 0.078, with 0.033 (RMSE) and 0.042 (MAE) using RF. As previously mentioned, the influence of fluctuating soil dielectric properties likely influenced this, whereby changing soil moisture levels dramatically alter backscattering. Moreover, ripening scored a maximum R^2 0.373, 0.068 (RMSE), and 0.058 (MAE) using RF. During this stage, spectral response changes dramatically following plant yellowing and panicle decay as chlorophyll content decreases, limiting correlations (Kuenzer and Knauer, 2013).

Table 4.1. A display of results obtained through cloud mitigation by combining VI and backscatter values across multiple dates. Following analysis, VV and NRPB backscatter values were used due to correlation to all VIs, while maintaining little multicollinearity. Highlighted values indicate significant results.

Algorithm	Vegetation Index (VI)	Soil			Vegetative			Reproductive			Ripening		
		R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)
Multiple linear	NDVI	-0.017	0.032	0.038	0.323	0.160	0.197	0.568	0.017	0.020	0.180	0.066	0.081
	EVI	0.001	0.024	0.031	0.289	0.119	0.150	0.471	0.035	0.043	0.029	0.068	0.084
	SAVI	-0.035	0.020	0.027	0.299	0.106	0.133	0.552	0.023	0.029	0.0428	0.053	0.064
RF	NDVI	0.078	0.033	0.042	0.551	0.125	0.165	0.536	0.017	0.021	0.373	0.048	0.058
	EVI	0.115	0.022	0.029	0.532	0.095	0.128	0.280	0.045	0.055	0.068	0.060	0.075
	SAVI	0.153	0.019	0.025	0.563	0.082	0.110	0.355	0.030	0.037	0.105	0.043	0.054
SVR	NDVI	0.067	0.033	0.042	0.391	0.159	0.196	0.530	0.015	0.020	0.086	0.06	0.076
	EVI	0.048	0.021	0.030	0.373	0.121	0.152	0.320	0.051	0.059	-0.060	0.071	0.085
	SAVI	0.016	0.021	0.028	0.374	0.094	0.135	0.369	0.034	0.040	-0.019	0.051	0.062
GBR	NDVI	-0.053	0.030	0.038	0.358	0.145	0.192	0.555	0.017	0.020	0.132	0.067	0.084
	EVI	-0.066	0.025	0.032	0.315	0.107	0.147	0.456	0.035	0.044	0.039	0.068	0.084
	SAVI	-0.059	0.021	0.028	0.336	0.095	0.130	0.541	0.024	0.029	0.043	0.053	0.064
XGBoost	NDVI	-0.082	0.077	0.039	0.583	0.114	0.155	0.578	0.016	0.020	0.142	0.068	0.083
	EVI	0.021	0.023	0.031	0.533	0.087	0.122	0.414	0.036	0.045	0.030	0.077	0.106
	SAVI	-0.086	0.021	0.028	0.552	0.077	0.106	0.504	0.024	0.030	0.029	0.053	0.065

4.2. Overall and within-plot approaches

VI values generated via overall (Table 4.2.) and within-plot (Table 4.3.) methods enabled a robust assessment of the most appropriate research direction for the investigation. The overall-plot approach produced significantly greater correlation for all VIs and phenological stages compared to within-plot. However, even the greatest correlation (R^2 0.316) proved insufficient for direct rice yield prediction, with almost 70% of variation unaccounted for. The reproductive stage was the only phenology to demonstrate any notable correlation during analysis, while vegetative and ripening stages proved insignificant. Minimal correlation at the vegetative stage could be attributed to its longer timeframe, allowing increased opportunity for value discrepancies (Kuenzer and Knauer, 2013). Meanwhile, the ripening stage experiences varied panicle growth and decaying biomass through plant maturity, likely resulting in spectral variations that diminish correlation performance (Sakamoto et al., 2011; Zhou et al., 2017).

EVI at the reproductive stage displayed the greatest relationship to yield with the Random Forest algorithm, with R^2, MAE, and RMSE of 0.316, 701.193, and 885.332, respectively. Overall, three algorithms generated R^2 values above 0.250, two of which using EVI, the other with SAVI. NDVI performed worst, achieving a maximum R^2 0.172. No correlative relationships were uncovered during within-plot analysis. Here, a substantial volume of harvester samples corresponded to individual VI values owing to the spatial resolution of Sentinel-2 data. This is illustrated in Figure 4.1., with several GPS yield samples partnered with one VI pixel value. Clearly this creates ambiguity, whereby VI pixel size corresponds poorly to within-plot yield samples, likely producing minimal correlation. Contrastingly, the overall-plot approach offers a more general overview of yield and VI values, avoiding the inconsistency posed by varied sample values within each VI pixel. If remotely sensed data of a higher spatial resolution were acquired such as PlanetScope, harvester sample values would have been designated more spatially precise VI values, potentially offering increased precision via the within-plot technique (Houborg and McCabe, 2016; Houborg and McCabe, 2018).

Resultingly, an overall-plot approach demonstrates elevated performance when assessing the relationship between *in situ* yield data and VI values. Contrastingly, within-plot yield samples would benefit from greater imagery resolution for correlative analysis, possibly via PlanetScope or a more precise airborne approach, though this is beyond the scope of the investigation (Houborg and McCabe, 2016; Houborg and McCabe, 2018). The generally low correlation for both approaches may be dependent on exclusion of other variables understood

to impact rice yield rates, meaning further analysis of the overall-plot approach alongside additional data is crucial to maximise performance.

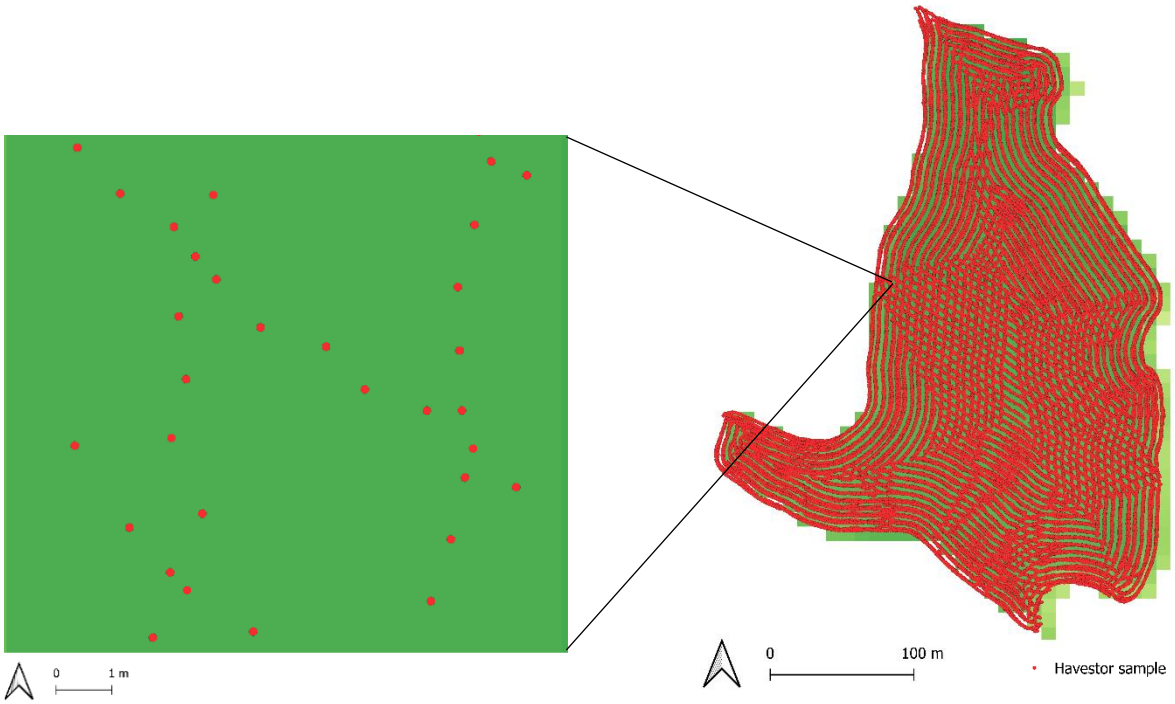

Figure 4. 1. An illustration of GPS yield samples in relation to VI pixel size, generated from Sentinel-2 data, demonstrating resolution discrepancy during within-plot yield prediction.

Table 4.2. A table displaying the performance of each individual VI and corresponding phenological stage in predicting rice yield (kg/ha) in the study area at an overall-plot level. Highlighted values indicate significant results.

Algorithm	Vegetation Index (VI)	Vegetative			Reproductive			Ripening		
		R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)
Simple linear	NDVI	0.006	860.723	1089.484	0.124	765.046	1022.849	0.004	872.077	1090.752
	EVI	0.016	859.927	1084.293	0.196	773.655	980.214	0.023	853.712	1080.164
	SAVI	0.004	856.972	1090.485	0.230	748.813	958.889	-0.006	867.309	1095.853
RF	NDVI	-0.017	873.9307	1102.22	0.162	780.795	982.706	0.084	853.659	1024.847
	EVI	-0.070	877.157	1107.578	0.316	701.193	885.332	0.031	861.387	1054.326
	SAVI	-0.087	898.904	1116.216	0.250	757.335	946.605	0.087	835.318	1022.949
SVR	NDVI	-0.051	927.501	1187.76	0.061	852.422	1122.835	0.028	889.455	1142.395
	EVI	-0.021	925.781	1170.805	0.255	777.117	1000.037	0.047	900.203	1131.099
	SAVI	-0.067	963.47	1197.453	0.122	816.097	1085.969	0.109	861.286	1093.875
GBR	NDVI	0.00	877.439	1092.492	0.174	749.274	993.373	0.074	844.210	1051.403
	EVI	0.029	844.518	1076.617	0.177	794.604	991.316	0.060	828.225	1059.334
	SAVI	-0.019	871.212	1103.445	0.256	753.241	942.477	0.029	857.728	1077.139
XGBoost	NDVI	-0.034	899.497	1111.312	0.172	750.473	994.549	0.102	838.105	1035.687
	EVI	-0.009	862.058	1097.73	0.179	798.166	990.49	0.061	829.971	1058.726
	SAVI	-0.064	883.526	1127.345	0.242	759.345	951.219	0.027	853.358	1078.222

Table 4. 3. A table displaying the performance of each individual VI and corresponding phenological stage in predicting rice yield (kg/ha) in the study area at a within-plot level. Highlighted values indicate significant results.

Algorithm	Vegetation Index (VI)	Vegetative			Reproductive			Ripening		
		R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)
Simple linear	NDVI	0.000	750.345	1010.640	0.001	729.927	976.350	0.015	738.888	1003.466
	EVI	0.001	749.471	1009.000	0.002	729.004	975.911	0.010	740.393	1006.212
	SAVI	0.001	749.498	1009.981	0.000	729.904	976.601	0.011	739.931	1005.323
RF	NDVI	0.027	726.077	985.608	0.011	722.986	971.364	0.013	738.195	1004.465
	EVI	0.024	733.994	998.059	0.001	729.169	976.183	0.010	738.024	1006.244
	SAVI	0.004	745.492	1008.621	0.002	729.196	975.685	0.022	735.884	999.754
SVR	NDVI	0.021	733.442	999.055	0.020	738.210	997.730	0.023	734.846	1003.862
	EVI	0.020	734.421	999.116	0.006	744.748	1005.166	0.031	732.472	1000.076
	SAVI	0.002	742.216;	1008.285	0.002	746.252	1006.897	0.025	737.111	1002.864
GBR	NDVI	0.027	731.291	996.934	0.024	716.973	964.905	0.029	732.576	996.281
	EVI	0.032	730.642	994.099	0.004	728.116	974.931	0.023	732.626	999.381
	SAVI	0.010	743.098	1005.264	-0.011	730.724	982.022	0.025	734.039	998.305
XGBoost	NDVI	0.027	731.063	996.984	0.022	718.517	965.780	0.030	732.140	995.941
	EVI	0.032	731.296	994.007	-0.004	727.669	978.733	0.023	732.026	999.428
	SAVI	0.011	743.297	1004.912	-0.012	731.673	982.491	0.025	734.504	998.235

4.3. Cultivar model inclusion

Additional *in situ* field measurements and climate variables were harnessed to improve correlations attained during overall-plot analysis. Rice cultivar information was utilised, owing to its high feature importance during prior investigations (Kuenzer and Knauer, 2013; Delerce et al., 2016; Zhou et al., 2017). To maximise available data and maintain robustness, an arbitrary count threshold of 10 yield values per cultivar was established, as per prior investigations (Delerce et al., 2016; Géron, 2019). Resulting cultivars were as follows: Fedearroz 67, Fedearroz 68, Escobal 417, Escobal 518, TRIUNFO, Orizica, and Maja 6. Figure 4.2 visualises recorded yield counts for each cultivar.

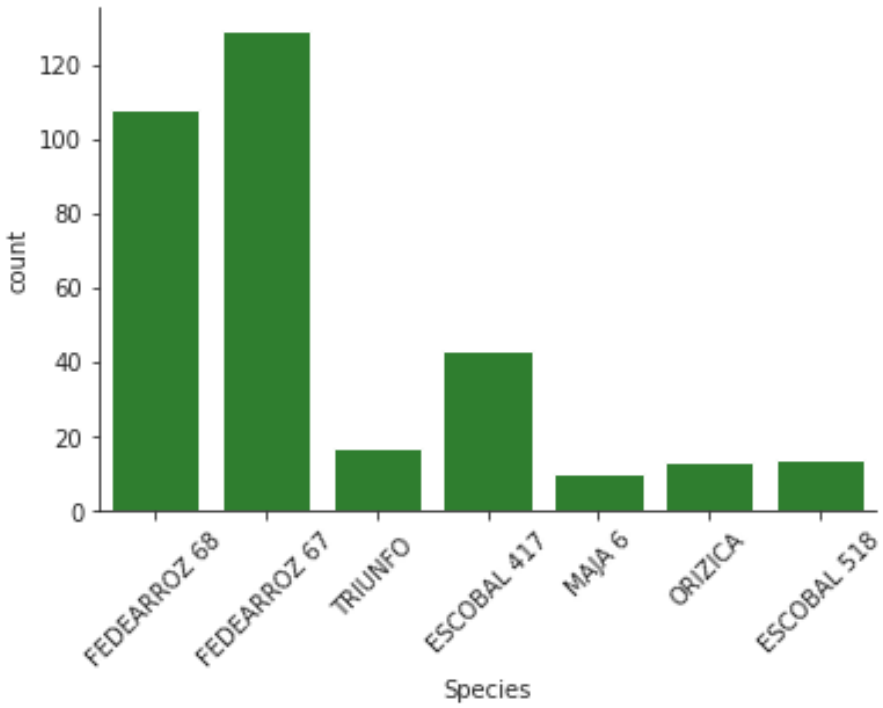

Figure 4.2. A bar plot presenting cultivar count for data used during yield prediction modelling.

Inclusion of cultivar information, alongside corresponding VI values and phenological stages, produced some encouraging results (Table 4.4.), with multiple model performance metrics indicating high correlation (Alexander et al., 2015). As with overall-plot analysis, the reproductive phenological stage provided the strongest relationship to yield, whereby plant booting has previously indicated future yield capacity (Zhou et al., 2017; Wang et al., 2019a). Escobal 518 demonstrated the strongest relationship, accounting for the two greatest correlations, whereby GBR generated R^2 0.901, MAE 203.853, and RMSE 347.827, while XGBoost produced R^2 0.890, MAE 300.931, and RMSE 366.701. This establishes the capacity

for yield prediction with Escobal 518, whereby EVI values captured one to two months prior to harvest account for over 90% of data variation. Maja 6 and Fedearroz 68 also showed potential, garnering R^2 0.853 and 0.595, respectively. Contrastingly, Escobal 417 displayed minimal yield relationships, while Orizica delivered reasonable correlations at vegetative (R^2 0.744) and ripening (0.577) stages, while the reproductive stage proved insignificant.

Clearly yield prediction capacity fluctuates depending on cultivar and phenology, but additional data would be beneficial for providing a more detailed impression (Géron, 2019). However, Fedearroz 67 produced comparatively poor correlation, achieving a maximum R^2 value of 0.385 (MAE: 884.897, RMSE: 1096.000), while holding the largest data quantity of all cultivars examined. This emphasises that although data volume often allows greater accuracy, obvious discrepancy between cultivars exist.

Table 4.4. The correlative performance of specific rice cultivars in predicting yield in the study area, with reference to the most successful algorithms during initial performance analysis. Highlighted values indicate significant results.

Cultivar	Algorithm	VI	Vegetative			Reproductive			Ripening		
			R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)
Escobal 417	XGBoost	NDVI	-0.024	829.884	1049.017	-0.729	1280.455	1363.441	-0.369	1080.488	1212.988
		EVI	-0.271	1061.451	1168.723	-2.495	1756.431	1938.369	-0.326	1099.681	1193.974
		SAVI	-1.143	1124.470	1517.682	-0.743	1290.904	1368.984	-0.302	1113.275	1182.857
	RF	NDVI	-1.020	1164.928	1501.468	-0.461	653.163	882.460	-0.360	676.384	851.360
		EVI	-0.410	658.279	867.031	-0.367	672.303	853.526	-0.280	670.218	826.044
		SAVI	-0.441	1107.406	1268.234	-0.995	1255.914	1492.131	-0.543	1179.711	1312.209
	GBR	NDVI	0.072	781.285	998.681	-0.352	1128.681	1205.579	-0.270	1045.471	1168.384
		EVI	-0.333	1110.649	1197.023	-0.236	1063.556	1152.691	-0.352	1128.681	1205.579
		SAVI	-0.995	1127.036	1464.348	-0.352	1128.681	1205.579	-0.352	1128.681	1205.579
Escobal 518	XGBoost	NDVI	-0.135	946.938	1178.869	0.730	402.907	575.211	-0.908	1276.365	1528.209
		EVI	-0.906	1356.973	1527.501	0.890	300.931	366.701	-0.292	1126.041	1257.535
		SAVI	-0.016	1049.004	1115.077	0.642	641.050	662.307	0.193	792.126	994.117
	RF	NDVI	-0.139	1324.101	1324.303	0.417	944.893	947.651	-0.139	1324.101	1324.303
		EVI	-0.139	1324.101	1324.303	0.383	902.297	974.883	-0.139	1324.101	1324.303
		SAVI	-1.151	1816.113	1819.817	0.656	723.630	727.640	-0.139	1324.101	1324.303
	GBR	NDVI	-0.123	948.111	1172.638	0.606	644.954	694.197	-1.062	1346.219	1588.662
		EVI	-1.445	1525.879	1729.984	0.901	203.853	347.827	-0.078	1040.761	1148.532
		SAVI	0.302	876.014	924.555	0.673	531.570	632.773	0.412	717.835	848.435
Fedearroz 67	XGBoost	NDVI	0.023	1055.734	1381.237	0.164	1004.736	1277.178	0.019	1058.266;	1383.545
		EVI	-0.038	1108.979	1423.799	0.385	884.897	1096.000	-0.006	1122.860	1401.186
		SAVI	0.017	1069.536	1385.306	0.271	936.899	1192.705	-0.024	1101.430	1413.851
	RF	NDVI	-0.023	950.098	1147.407	-0.245	926.594	1098.663	0.059	940.616	1100.264
		EVI	-0.168	966.825	1225.645	0.059	904.520	1100.481	-4.10	1080.465	1346.780
		SAVI	-0.201	998.848	1243.084	-0.134	864.944	1048.419	0.034	927.991	1114.776
	GBR	NDVI	0.013	1065.970	1388.059	0.189	1018.002	1258.249	0.096	1013.307	1328.207
		EVI	-0.006	1080.427	1401.102	0.271	945.823	1193.097	-0.003	1083.173	1399.221
		SAVI	0.008	1071.069	1391.500	0.211	968.793	1240.973	0.028	1059.072	1377.124
Fedearroz 68	XGBoost	NDVI	0.004	943.843	1213.257	0.340	783.820	987.640	-2.086	1089.285	1336.346
		EVI	0.181	880.535	1100.354	0.595	630.846	771.386	-0.046	967.813	1243.329

		SAVI	0.121	902.307	1139.670	0.398	773.409	943.522	0.021	930.303	1202.800
	RF	NDVI	0.059	710.125	899.813	-0.473	793.213	1125.961	-0.059	761.013;	954.541
		EVI	0.182	681.233	839.186	0.124	847.096	963.155	-0.215	954.812	1134.176
		SAVI	-0.251	811.095	1037.450	-0.016	616.180	935.218	-0.284	999.990	1165. 934
	GBR	NDVI	0.331	813.682	994.038	0.358	764.253	973.966	-0.047	991.520	1243.883
		EVI	0.227	897.286	1068.617	0.559	658.907	807.406	-0.030	972.875	1233.796
		SAVI	0.203	880.956	1085.460	0.532	702.213	831.861	0.019	942.788	1204.019
	XGBoost	NDVI	-0.316	636.977	720.737	-1.039	883.987	897.248	-0.119	628.202	664.657
		EVI	-0.031	632.555	637.740	-0.956	631.519	878.484	0.539	391.719	426.566
		SAVI	-0.06	883.987	897.248	0.101	594.696	595.569	0.535	428.056	428.595
	RF	NDVI	0.049	936.665	945.572	0.853	152.327	182.698	0.363	769.164	773.870
Maja 6		EVI	0.086	923.217	926.965	0.609	298.072	298.096	-1.590	750.266	766.902
		SAVI	0.234	848.510	848.554	0.736	194.268	244.731	0.342	601.143	786.510
	GBR	NDVI	-0.535	951.936	1044.044	-2.634	1368.670	1606.396	-4.455	1787.822	1968.254
		EVI	-2.243	1262.087	1517.552	-1.590	1183.043	1356.165	-2.065	1095.621	1475.368
		SAVI	-3.204	1291.854	1727.851	-0.675	932.308	1090.711	-4.038	1650.570	1891.523
	XGBoost	NDVI	-0.605	978.312	1067.435	-3.367	1505.253	1761.005	-5.300	1938.856	2115.145
		EVI	-3.326	1380.719	1752.603	-1.482	1180.266	1327.671	-2.420	1267.176	1558.406
		SAVI	-3.810	1423.697	1848.123	-1.269	1098.424	1269.443	-3.959	1509.949	1876.611
	RF	NDVI	0.744	504.623	515.445	-3.642	2149.687	2192.742	-1.503	1426.896	1610.221
Orizica		EVI	-0.286	1078.269	1154.250	-1.404	1575.820	1578.080	-2.492	1245.300	1901.948
		SAVI	-0.082	853.423	1058.750	-1.200	1441.812	1509.482	0.577	587.900	662.267
	GBR	NDVI	-0.535	951.936	1044.044	-2.243	1262.087	1517.552	-4.455	1787.822	1968.254
		EVI	-2.243	1262.087	1517.552	-1.590	1183.043	1356.165	-2.065	1095.621	1475.368
		SAVI	-3.204	1291.854	1727.851	-0.675	932.308	1090.711	-4.038	1650.570	1891.523
	XGBoost	NDVI	0.266	1084.876	1226.319	-2.113	2052.034	2527.073	0.149	1187.172	1320.978
		EVI	0.557	820.749	953.667	0.374	1015.046	1133.411	-0.363	1631.698	1672.163
		SAVI	0.010	1260.584	1425.164	-0.034	1217.347	1456.676	-0.048	1241.695	1466.291
	RF	NDVI	-0.238	810.079	950.486	-3.813	1487.699	1874.303	-1.833	1156.564	1437.906
TRIUNFO		EVI	-0.704	999.240	1216.438	0.687	397.276	521.644	-3.614	1759.769	1835.271
		SAVI	-0.233	709.679	1034.907	-1.977	989.704	1607.909	-1.394	1270.330	1441.755
	GBR	NDVI	0.237	1104.279	1250.770	-1.174	1718.844	2112.193	-0.021	1314.160	1447.119
		EVI	0.672	749.043	820.202	0.393	897.639	1115.481	-0.208	1542.953	1574.181
		SAVI	0.051	1239.992	1395.358	-0.045	1218.553	1463.910	-0.106	1358.680	1506.243

4.4. *In situ* field measurements and climate variables

Building upon performance improvements offered by cultivar specification, further *in situ* field measurements and climate variables were explored. Following the same multicollinearity and modelling methods previously detailed, additional data was utilised for improved prediction performance and feature importance analysis. Figure 4.3. presents a Pearson's coefficient heatmap, demonstrating collinearity between variables. Here, stronger collinearity between remotely sensed metrics is present, most prominently between shared phenological stages of NDVI, EVI, and SAVI values. Climate variables and *in situ* field measurements demonstrate lower collinearity, though still notable between solar radiation, drought days, and relative humidity, all linking to temperature variations.

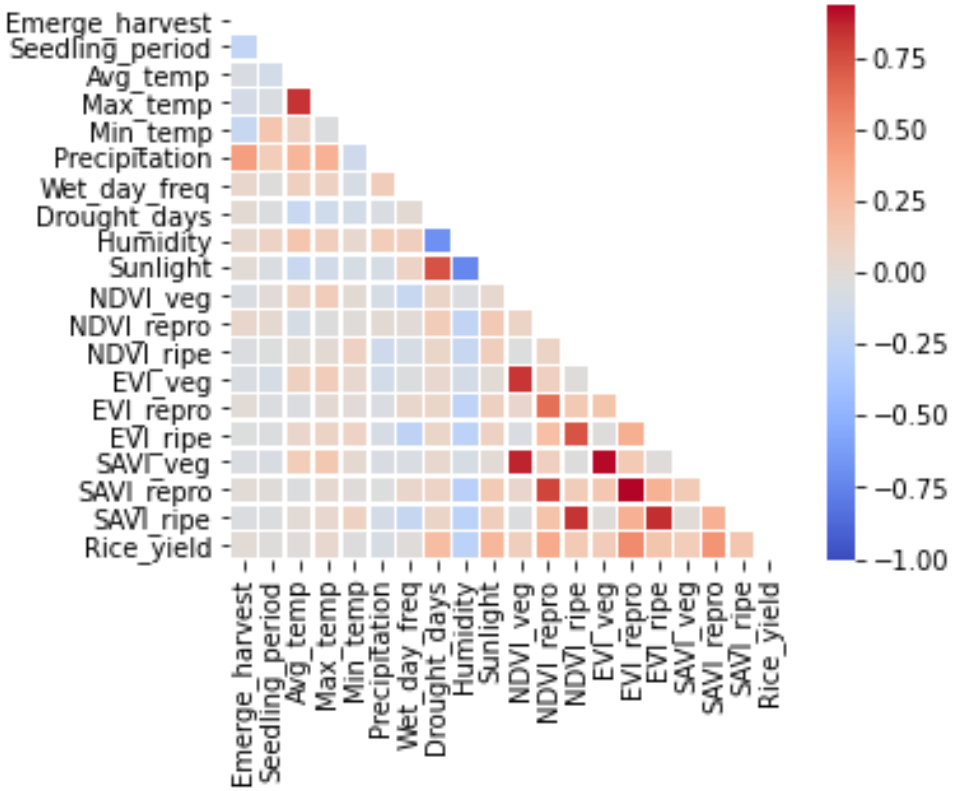

Figure 4.3. A heatmap presenting Pearson's correlation analysis of all variables used during rice yield prediction modelling.

The VIF was also employed to identify non-linear collinearity (Géron, 2019; Deisenroth et al., 2020). Following this, resulting variables were modelled to establish the prediction performance for each cultivar with the additional variables. Table 4.5. presents the most encouraging results for each rice cultivar.

Two variables, namely the seedling period and number of drought days per month, both strengthened yield prediction performance, though not in all circumstances. Additionally, the inclusion of VI values at specific phenological stages appeared a crucial component for every model tested, demonstrating the importance of EO during rice yield prediction. The EVI and reproductive phenology proved dominant during analysis, included in most cultivar models, demonstrating reproductive booting as a strong determinant of yield (Zhou et al., 2017). The EVI vegetative phenology displayed some success, specifically for the Orizca cultivar, achieving an R^2 of 0.744. Conversely, ripening failed to attain the highest performance in any model, likely due to varied panicle growth and decaying biomass following plant maturity, causing unpredictable spectral variations (Sakamoto et al., 2011; Zhou et al., 2017). Escobal 518 proved the most encouraging cultivar for yield prediction, attaining a performance of 0.949 (R^2), 191.059 (MAE), and 250.934 (RMSE), with XGBoost. This was achieved with the inclusion of both seedling and drought variables, alongside EVI values at the reproductive stage. Therefore, the modelling process accounted for 95% of yield variation for Escobal 518 in the study area.

The majority of *in situ* data and climate variables demonstrated substantial collinearity, deeming them unsuitable for model inclusion to avoid overfitting (Géron, 2019; Deisenroth et al., 2020). This can be interpreted in different ways; data could have been of poor quality, though retrieval from multiple sources weakens this argument. Alternatively, unlike EO data, climate variables were not divided phonologically, instead encompassing entire growing periods. Judging by the improvements that VI division brought to performance, the same pre-processing to climate variables may have had a similarly positive impact (Delerce et al., 2016). Even so, the addition of these variables was clearly beneficial for some cultivars, providing interesting connotations to the investigation.

Where the inclusion of additional data presented no performance improvements, values previously derived from VIs and phenological division were used. Maja 6's R^2 value of 0.853 was generated using only NDVI values at the reproductive stage, with no benefit from further variables. This scenario could be due to the lower value count for some cultivars, providing less information for robust modelling, or was simply less impacted by other variables (Delerce et al., 2016). Furthermore, the research covered a relatively short period, approximately 5 harvests, meaning climate variables could have had limited time to demonstrate noticeable influence. This is consequential considering the influence that the warmer El Niño and cooler La Niña events exert on Colombian rice production, possibly causing disparities within the

independent variables (Esquivel et al., 2018; Cai et al., 2020). For example, El Niño flooding between 2016 and 2017 heavily impacted Colombian agriculture, which may have influenced modelling performance due to limited data availability (Cai et al., 2020).

Table 4. 5. Using algorithms harnessed during specific cultivar analysis, maximum performance metrics of each cultivar using additional climate data and in situ field measurements is displayed to explore yield influences. Results are highlighted where improvements have been established with the addition of these variables.

Cultivar	Variables	Algorithm	R^2	MAE (kg/ha-1)	RMSE (kg/ha-1)
Escobal 417	EVI_veg, drought days, seedling stage	RF	-0.154	921.971	1113.821
	NDVI veg	XGBoost	-0.024	829.884	1049.017
	EVI_veg, drought days, seedling stage	GBR	-0.279	904.893	1172.329
Escobal 518	NDVI_repro, drought days, seedling stage	RF	0.914	251.880	324.922
	EVI_repro, drought days, seedling stage	XGBoost	0.949	191.059	250.934
	EVI_repro	gbr	0.901	203.853	347.827
Fedearroz 67	SAVI_repro, drought days, seedling stage	RF	0.458	857.075	1029.052
		XGBoost	0.427	875.655	1057.166
	EVI_repro, frought days, seedling	GBR	0.426	910.477	1098.223
Fedearroz 68	SAVI_repro, drought days, seedling stage	RF	0.551	690.969	814.171
	EVI_repro, drought days, seedling stage	XGBoost	0.556	668.298	810.161
	SAVI_repro, drought days, seedling stage	GBR	0.468	766.418	930.124
Maja 6	NDVI_repro	RF	0.853	152.327	182.698
	EVI veg	XGBoost	-0.031	632.555	637.740
	EVI_repro, drought, seedling	GBR	0.118	430.929	590.034
Orizica	EVI_veg	RF	0.744	504.623	515.445
	SAVI_repro, drought	XGBoost	0.124	560.908	587.967
	SAVI repro	GBR	-0.675	932.308	1090.711
TRIUNFO	EVI_repro	RF	0.687	397.276	521.644
	EVI_repro, drought, seedling	XGBoost	0.697	761.188	853.997
	EVI_veg	GBR	0.672	749.043	820.202

To better understand the impact of each variable on yield prediction modelling, Scikit-Learn's feature importance function was implemented for the best performing cultivars (Pedregosa et al., 2011). Figure 4.4. presents resulting analysis, visualising variable influence on the best performing models. While EO metrics appeared dominant throughout, cultivars evidently react differently to drought and seedling period variables. This may suggest such analysis can be used for managerial decision-making, with cultivars less impacted by climate variables more robust to environmental fluctuations (Delerce et al., 2016). Notably, more extreme temperatures and lower precipitation can prolong rice seedling stage (Delerce et al., 2016; Quevedo et al., 2020), in turn reducing yield volume (Gan et al., 1992). Therefore, Escobal 518's reduced seedling period influence compared to Triunfo indicates it is more appropraite as climate change continues to impact Colombia. Moreover, Fedearroz 67 and Fedearrox 68 appeared to react comparably to variables, though Fedearroz 67 displayed slightly more influence from reproductive EVI values.

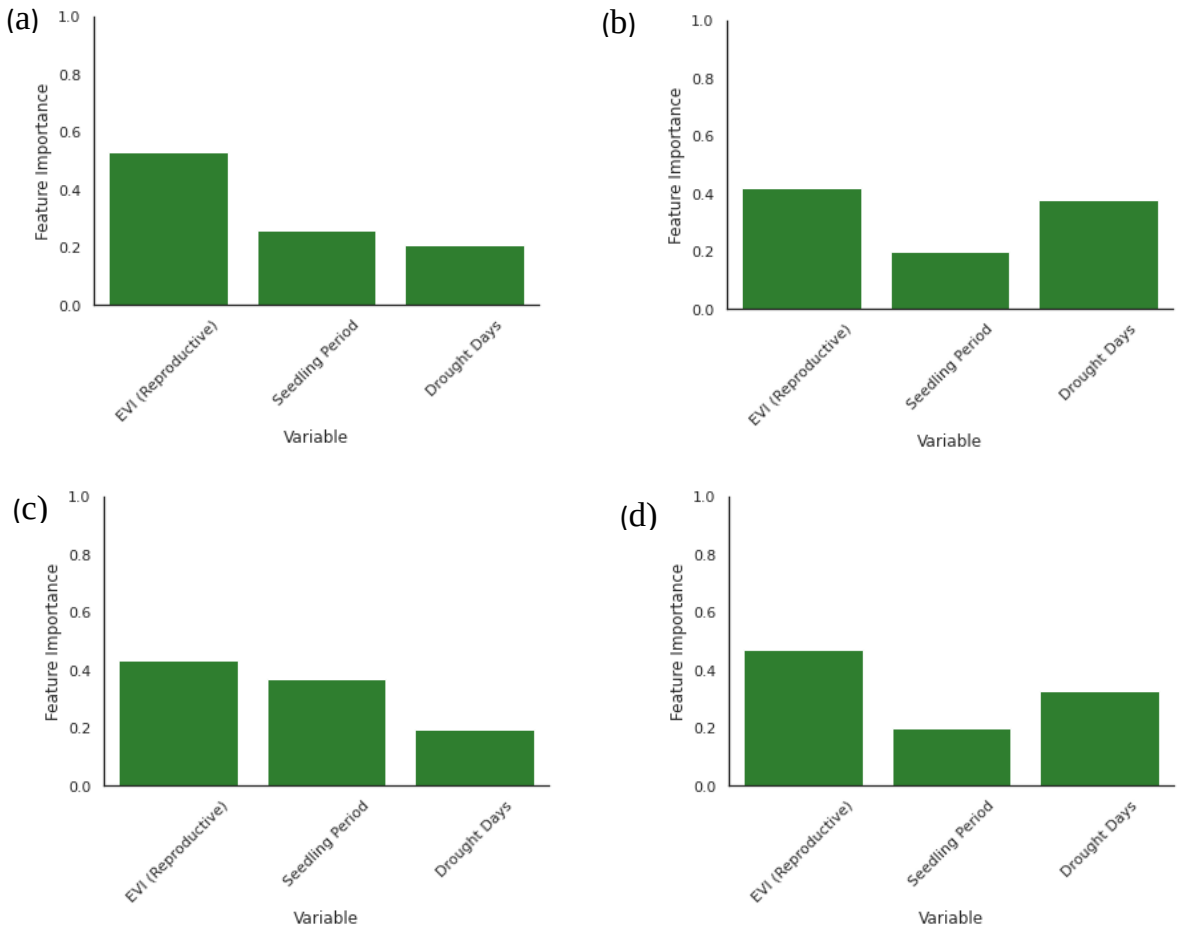

Figure 4.4. Feature importance of variables for best performing species: (a) Escobal 518 (XGBoost); (b) Fedearroz 68 (XGBoost); (c) Triunfo (XGBoost); (d) Fedearroz 67 (XGBoost).

Figure 4.5. visualises the most encouraging model performances influenced by EO metrics, drought information, and sowing periods. Here predicted yield values are plotted against actual yield data, presenting cultivar specific trends. Though Escobal 518 achieved the highest performance, the test set volume is minimal compared to Fedearroz 67 and Fedearroz 68. This perhaps influences performance, highlighting the need for further data to reinforce investigative findings.

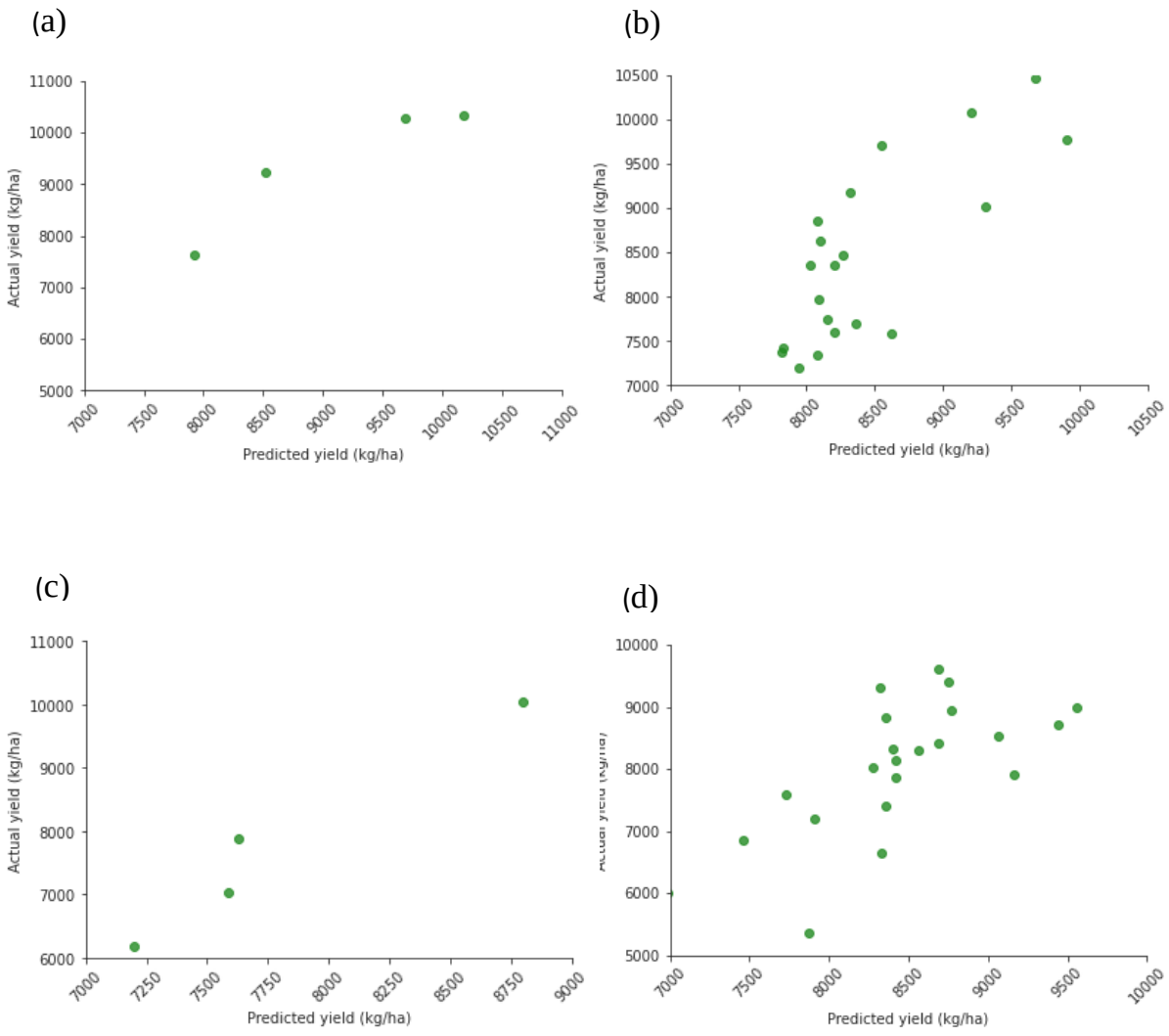

Figure 4.5. Scatter plots displaying the predicted yield vs actual yield values from the highest performing models using EO metrics, climate variables, and in situ field measurements, specifically: (a) Escobal 518 (XGBoost); (b) Fedearroz 68 (XGBoost); (c) Triunfo (XGBoost); (d) Fedearroz 67 (XGBoost).

Chapter 4

5.0. Discussion

5.1. Cloud coverage mitigation

Considering the encouraging correlation and minimal error detected at the reproductive phenology, opportunity for cloud cover mitigation of optical satellite coverage approximately one to two months prior to harvest exists. Additionally, the vegetative stage generated the highest R^2 value during analysis (R^2 0.583), albeit with larger error fluctuations between models. Similar to Filgueiras et al.'s (2019) investigation, the inclusion of the NRPB alongside VV improved correlations, whereby the ratio of VV and VH allows greater model generalization (Vreugdenhil et al., 2018), though equivalent levels of correlation were not found.

The poor relationship at the ripening stage appears significant; Figure 5.1. displays the typical appearance of rice at each phenological stage, highlighting a clear increase in yellowing during maturity. Past research has demonstrated VI value saturation following plant yellowing, witnessed during flowering and ripening, which is linked to declining chlorophyll content (Shen et al., 2010, Haagsma, 2015). Here, increasing yellowness in conjunction with reduced biomass as leaves decay cause lower VI values (Kuenzer and Knauer, 2013; Mosleh et al., 2015; Ariza, 2019). Yet as backscatter values are concerned with plant canopy structure and moisture content, this variation shown in the optical wavelengths would not be replicated, leading to a diminished correlation.

Figure 5.1. A representation of typical rice appearance at the (a) vegetative, (b) reproductive, and (c) ripening stages of phenological development. Specific focus is given to largely green canopies at both the vegetative and reproductive stages, followed by significant yellowing upon maturity. Modified from Yang et al. (2016) and He et al. (2018).

The average values of each phenological stage for the NDVI and VV polarisation reinforce this (Figure 5.2.). Here, a steep decline at the ripening stage associated with NDVI is not replicated to the same extent by VV, which instead displays a value only slightly less than the preceding reproductive stage. Prior investigations support this synopsis; Zhou et al. (2017) determined varied panicle growth and ripening stages can cause difficulties in identifying rice yield due to introduced spectral variation from decaying biomass, while Sakamoto et al. (2011) concluded that colour variation at later phonologies reduced rice identification and yield prediction accuracy. As backscatter values are dependent upon canopy structure and plant moisture content, its unfeasible to apply this cloud mitigation method as plant maturity progresses, and explains the low correlation at the ripening stage. However, Vreugdenhil et al. (2018) emphasised significant backscatter sensitivity to leaf water content, meaning leaf decay associated with ripening may explain the slight backscatter reduction observed (Figure 5.2.). Overall, a more precise assessment of phenology at the vegetative stage appears the most encouraging route to improving cloud mitigation.

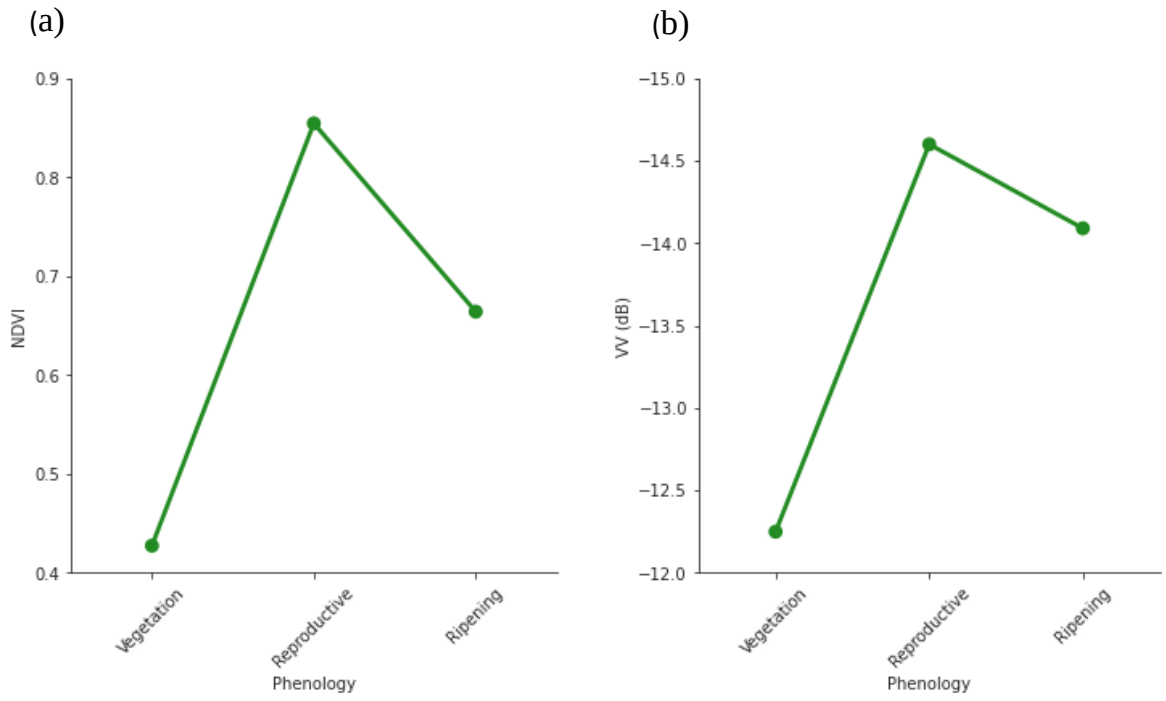

Figure 5.2. A plot displaying the average value of (a) NDVI and (b) VV backscatter at each phenological stage from cloud mitigation data. Peak values are reached for both at the reproductive stage (0.85 and -14.7 respectively). Both variables appear to mimic each other through the vegetative and reproduction stages, though both values recede from their peak during ripening, VV backscatter is less impacted compared to the change from vegetative to reproductive, while the NDVI declines to a greater extent.

Sentinel-1's C-band does not have the capacity to interact with rice in the same way as the more penetrative L-band, meaning there is significant interaction within the plant canopy. This further supports the variations in backscatter caused by canopy structural change during vegetative, while the spectral change associated with ripening was inadequately recorded with SAR C-band (Figure 5.2.). The minimal response to soil demonstrates the benefits of more in-depth phenology spitting, as dialectic properties dictated by moisture content dominate backscatter (Bousbih et al., 2017). In much the same way, a proportion of recorded vegetative values will be influenced by the dialectic constant during the early stages of rice emergence (Bousbih et al., 2017). By removing this period from analysis, correlations may strengthen. Duan et al. (2019) detailed a preference for more precise phenology division while investigating rice yield prediction in Hubei Province, China. Here, an uneven presence of panicles and leaves caused a diverse spectral response, leading to diminished crop parameter estimations (Duan et al., 2019). The vegetative stage generally covers a period double that of reproductive and ripening stages (Kuenzer and Knauer, 2013), yet provided the highest performance throughout cloud mitigation modelling (R^2 0.583). Following past research, significant fluctuations in both reflectance and backscatter are therefore likely at this stage due to the extended period of plant development, meaning more precise phenological splitting to include germination, tillering, and stem elongation may diminish the high levels of error, improving cloud mitigation capabilities (Zheng et al., 2016; Filgueiras et al., 2019; Zhang et al., 2019b). Tillering and stem elongation stages are less impacted by soil moisture and corresponding dielectric properties, as the closed canopy means minimal soil is visible to influence backscatter (Bousbih et al., 2017; Filgueiras et al., 2019).

Zhang et al. (2019b) noted the strong yield prediction capacity during stem elongation within the vegetative phenology, alongside booting at the reproductive stage. This coincides with encouraging correlations found during the investigation, meaning further phenological splitting has potential to develop cloud mitigation at growth stages crucial for yield prediction (Zhang et al., 2019b). By focusing on more precise phenology division, particularly at the vegetative stage following canopy closure, error may be minimised by ensuring all values are retrieved at a period of closed canopy with little influence from soil dielectric properties (Filgueiras et al., 2019; Zhang et al., 2019b). This is valuable for maximising the available optical satellite data at a key growth period for yield prediction (Filgueiras et al., 2019; Zhang et al., 2019b). Successful implementation and extrapolation of this approach elsewhere in Colombia and

beyond could therefore contribute to strengthening food security (Castro-Llanos et al., 2019; Jiménez et al., 2019; Weiss et al., 2020).

Due to the influence of these factors on past investigations, accounting for rice cultivar and season has potential to improve cloud mitigation correlations owing to associated spectral variations (Kuenzer and Knauer, 2013; Delerce et al., 2016; Zhou et al., 2017). However, the additional data preparation and complications involved may prove difficult and largely unrealistic for wider extrapolation.

5.2. Overall and Within-Plot

Determining the most appropriate plot approach was important in optimising yield prediction within the study area. The poor performance generated via the within-plot approach was likely related to insufficient Sentinel-2 spatial resolution. Further, GPS yield sampling used during within-plot analysis held notable limitations, whereby significant value discrepancy within VI pixels impacted modelling (Leroux et al., 2018). This inconsistency can result from various scenarios; sample values are highly dependent on harvester operator skill, whereby vehicle speed, cutting head angle, and presence of foreign materials can all influence recorded yield data (Blackmore, 1999; Arslan and Colvin, 2002; Leroux et al., 2018). An overall-plot approach proved a successful alternative, albeit with disadvantages. For example, lower data volume provided decreased performance certainty and modelling compared to the abundant GPS sample points (Géron, 2019). In response, cultivar values below a count of 10 were removed to mitigate fluctuations associated with limited data (Delerce et al., 2016; Géron, 2019).

Satellite data with increased spatial resolution has been harnessed during previous investigations to achieve improved yield prediction performance. Noureldin et al. (2013) acquired 10 m spatial resolution SPOT data, where VIs harnessing red and near-infrared spectral reflectance consistently achieved R^2 values above 0.80. Additionally, the application of 3 m spatial resolution Quickbird data has been used to generate VIs for grain yield prediction using GPS yield samples, whereby Yang et al. (2006) established a maximum R^2 value of 0.81 through SLR. However, these approaches were unsatisfactory for the present investigation; Quickbird is no longer operational, while SPOT data is commercially distributed, diminishing open-accessibility. Additionally, these approaches centred on an individual plot of one cultivar over a single harvest, while the present investigation examined multiple rice varieties and unstructured growing seasons across several plots. However, the development of satellite constellations such as those provided by Planet Labs may combat the limited spatial-temporal

frequency of Sentinel-2 data for more precise phenological division (Houborg and McCabe, 2016; Houborg and McCabe, 2018). Though sensor inconsistencies within the constellations relating to spectral bandwidth and radiometric quality, alongside their commercial nature, demonstrate this approach is constrained (Houborg and McCabe, 2018).

Alternatively, the opportunity to further develop within-plot research is presented by Wang et al. (2019b) in Jiangsu Province, China, which experiences a humid subtropical monsoon climate comparable to the study area. Encouraging results were produced with a SAR simple difference (SSD) index, which harnessed the variation between VH at the end of rice tillering and grain filling phenological stages, achieving an RMSE of 740 kg/ha-1 (Wang et al., 2019b), an improvement on the highest RMSE from overall-plot analysis (885 kg/ha-1). Thus, exploring within-plot yield prediction with Sentinel-1 data is possible, bypassing unfavourable weather conditions, while remaining freely accessible. The implementation of Sentinel-2 can still be considered an effective option given the circumstances, albeit through an overall-plot approach. Moreover, Sentinel-2 data allows for possible extrapolation to other suitable rice growing areas in Colombia and beyond given its open-source accessibility, as detailed by Castro-Llanos et al. (2019). However, including further climate variables and *in situ* field measurements proved critical to further optimise yield prediction performance (Delerce et al., 2016).

5.3. Enhancing model yield prediction capacity

Following past research (Delerce et al., 2016), division of overall-plot data by rice cultivar proved an interesting development, achieving higher prediction capacity. Indeed, the maximum overall-plot performance of R^2 0.316 was elevated following cultivar specification, with notably strong collinearity displayed by Escobal 518 (R^2 0.901), Maja 6 (R^2 0.853), Orizca (R^2 0.744), and Triunfo (R^2 0.687). Likewise, 29 of the 63 generated cultivar models achieved higher than the best overall-plot performance, further demonstrating the effectiveness of cultivar division during prediction modelling. Results suggest some cultivars are more responsive to modelling and specific climate variables, a factor to be considered by farmers for optimum yield production as the climate alters (Delerce et al., 2016). Similar scenarios have been observed in previous research; Delerce et al.'s (2016) investigation in Colombia uncovered that rice cultivar had the greatest influence on yield rates, alongside weather variables at the reproductive stage. Delerce et al. (2016) revealed significant yield variation between cultivars during specific climate scenarios, thereby allowing cultivar recommendations for farmers where such climates were anticipated. Moreover, alternating

yield rates based upon climate variables allowed farmers to adequately adapt to the projected climate change in the region, whereby cultivars more suitable to decreasing precipitation rates and increasing temperatures were identified (Shah et al., 2011; Delerce et al., 2016).

Following Delerce et al.'s (2016) trajectory, the present investigation went on to assess a range of previously detailed climate variables and *in situ* field measurements to optimise cultivar-specific modelling. Again, performance variations between cultivars was evident, with some showing no improvements with additional climate variables. However, seedling period and the number of drought days both demonstrated performance improvements for some cultivars, while other variables displayed excessive collinearity to the dependent variable. For example, the addition of these variables alongside reproductive EVI values for Escobal 517 produced an exceptional R^2 of 0.949, accounting for 95% of rice yield variation. Further, Escobal 518's reduced seedling period influence compared to Triunfo following feature importance analysis suggests it is a more appropraite cultivar as climate change impacts Colombia, providing useful advice for regional decision making. This demonstrates the value in incorporating a range of data sources, including EO metrics, climate variables, and *in situ* field measurements during rice yield investigations. Delerce et al. (2016) also demonstrated that cultivar, seedling, and drought influence benefited rice yield prediction, yet they achieved a maximum R^2 of 0.502 at Saldaña, Colombia, demonstrating the importance of EO data inclusion during such research.

The obvious benefits seedling and drought information bring is significant in the context of a changing climate and food security. The influence of drought in the study area has substantial literary support; Heinemann et al. (2015) studied drought impacts on upland rice production in neighbouring Brazil, whereby 44% of cultivated rice was most impacted by drought, particularly during the reproductive period. Moreover, Heinemann and Sentelhas (2011) concluded that increased drought levels were the dominant abiotic stress upon upland rice production during prior research. Similarly, Delerce et al. (2016) demonstrated that rainfall frequency appeared the most important feature for yield prediction for certain cultivars, notably in the vegetative stage. To combat the negative climatic influence on rice yield, Heienemann et al. (2015) recommended an adaptation approach, one option being selection of the most resilient cultivars for maximum production, a method also proposed by Delerce et al. (2016). Owing to the influence of drought, a similar strategy could be implemented at Hacienda El Escobal, whereby cultivars less affected by the drought variable may be more suitable for future cultivation to minimise food insecurity. Through feature importance analysis, Escobal 518 and Triunfo indicated less influence from drought compared to sowing period, while both

Fedearroz 67 and Fedearroz 68 exhibited the opposite relationship, with substantial drought-based impact. One could therefore recommend Escobal 518 and Triunfo as suitable cultivars for future production, owing to their reasonable prediction accuracy and reduced impact from lower precipitation rates.

The influence of seedling period during modelling is also of note following similar findings in Colombia (Delerce et al., 2016; Quevedo-Amaya et al., 2020). Indeed, extreme temperatures and low precipitation rates can prolong the seedling stage of rice (Delerce et al., 2016; Quevedo et al., 2020), which in turn is linked to reduced yield capacity (Gan et al., 1992). During Delerce et al.'s (2016) investigation regarding the most suitable sowing periods, they determined that rice sown throughout the year encountered varying weather events, with crops planted in April and May found to be beneficial for maximising yield (Delerce et al., 2016). Consequently, the influence of sowing period on the present investigation likely relates to the climatic variations noted by Delerce et al. (2016), whereby certain months are more suitable for rice sowing than others. Gan et al. (1992) reinforces this, arguing that shorter sowing periods typically correspond to improved yield rates, implying that minimising this growth stage is advantageous. Quevedo-Amaya et al. (2020) found a similar influence during research into rice yield optimisation in a region approximately 20km from Hacienda El Escobal. Concentrating on the Fedearroz 68 cultivar throughout 2017, Quevedo-Amaya et al. (2020) concluded that sowing date selection had no impact on production costs yet can increase profitability by up to 26% due to influence on yield. Correspondingly, the present investigation found the inclusion of sowing period also improved yield prediction accuracy for the Fedearroz 68 cultivar, though drought offered greater influence. This reinforces findings by Delerece et al (2016) and Quevedo-Amaya et al. (2020) regarding sowing impact on regional rice production, which appears to be dictated by annual climate variations. Owing to fluctuations and extreme events following climate change, one can expect the significance of sowing period to further increase, reinforcing its importance during prediction modelling (Delerce et al., 2016; Quevedo-Amaya et al., 2020). However, other influencing factors exist; mechanical characteristics of the seedbed, such as tillage practices, also contribute to the seedling stage (Blacklow, 1972).

Impacts of the warmer El Niño and cooler La Niña events on Colombian rice production should also be considered; the former tends to produce negative precipitation anomalies, while the latter sees positive precipitation anomalies (Poveda et al., 2001; Esquivel et al., 2018; Cai et al., 2020). The 2016-2017 El Niño flooding event is evidence of this; damage to Colombian agriculture output following crop destruction may have skewed modelling due to the lack of

years evaluated (Cai et al., 2020). As climate change progresses, so too will the intensity of extreme weather events, strengthening the need for maximising the data volume for robust analysis (Esquivel et al., 2018; Quevedo Amaya et al., 2019; Cai et al., 2020).

The phenological splitting of climate data, as applied to EO metrics, may have resulted in further correlations and more vigorous results. Climate data during the reproductive stage has shown significance to yield prediction in prior research, (Delerce et al., 2016; Iizumi et al., 2018; Das et al., 2020), much like that observed with EO metrics in the current investigation. Therefore, splitting climate variables at phenological stages could prove beneficial instead of analysing the entire growing period (Delerce et al., 2016). It can be theorised that the dominance of EO metrics in the best performing models could thus be related to their phenological splitting. In contrast, climate variables simply covered the whole growth period, perhaps weakening their correlation. However, van Oort et al. (2011) discussed the sensitivity of yield prediction models to accurately determine phenological stages, meaning further precision should be approached cautiously and is beyond the scope of the present investigation.

5.4. Extrapolation

Extrapolation of methods should be explored as climate change persists in tropical regions, risking food security. The presented cloud mitigation technique offers a rudimentary path for wider extrapolation to regions impacted by optical cloud masking due to the open-accessibility of Sentinel-1 and Sentinel-2 data. Furthermore, while rice cultivation is widespread in the tropics, potential exists for extrapolation to alternative crops. Filgueiras et al. (2019) demonstrated this, using a similar technique for maize and soybean monitoring in Brazil. Yet the maximum performance attained in the present investigation (R^2 0.583, MAE 0.114, RMSE 0.155) requires additional development prior to further application to reduce uncertainty. This investigation thoroughly explored options and recommends a more precise division of phenology at the vegetative stage to achieve this, reducing the large variation in plant structure and dielectric constant currently observed prior to canopy closure at the vegetative stage. By focussing analysis following canopy closure during maximum tillering and stem elongation stages, the high error rates will likely be mitigated (Filgueiras et al., 2019; Zhang et al., 2019b), allowing extrapolation of this technique to fill optical data gaps in the wider tropics.

Colombia is climatically diverse, offering difficulties in national-scale yield prediction extrapolation However, the robustness of the present investigation is supported by prior research within close proximity and elevation to the study area, whereby cultivar, seedling period, and drought have all been confirmed as significant factors in rice yield forecasting in

Tolima (Delerce et al., 2016; Quevedo et al., 2020). This strengthens the investigative findings and suggests potential for model extrapolation at least within Tolima department, this being the greatest rice producing department in Colombia (Castilla-Lozano et al., 2011; Delerce et al., 2016). Exploring this further, Castro-Llanos et al.'s (2019) research into the impact of climate change on Colombia's cultivated rice is significant. As the present study area is focused upon agricultural land situated at a suitable elevation for continued rice cultivation by 2050, this investigation's yield prediction methodology has potential to be extrapolated to the remaining 40% of suitable agricultural land in Colombia due to their shared characteristics (Castro-Llanos et al., 2019). Specifically, this land has sufficient elevation to avoid the impacts of rising temperatures and diminishing water availability (Castro-Llanos et al., 2019). Therefore, the method proposed in this investigation may prove a crucial part in maintaining future national food security. Feature importance analysis of specific cultivars holds additional value for extrapolation; one could recommend Escobal 518 and Fedearroz 68 as suitable cultivars for projected climate changes in the future, owing to their reasonable prediction accuracy and lessened impact from both drought and seedling period. Communicating such findings to the Colombian government and farmers through a set of guidelines could therefore be a valuable approach to extrapolating the proposed methods for the benefit of both regional and national food security (Jiménez et al., 2019).

5.5. Research limitations

Several limitations likely incurred influence on results. Data collection and modelling preparation is one example; during data quality assurance, imputation was harnessed to fill missing data values to maximise data volume during modelling. The imputation of average values introduced synthetic data not entirely representative of reality, bringing a degree of uncertainty (Géron, 2019; Deisenroth et al., 2020). This imputation, in conjunction with random error associated with instrument measurement limitations, contributes to error propagation, bringing uncertainty to statistical analysis (Deisenroth et al., 2020). Further inaccuracy likely arises from field measurements following human error; variables such as sowing and emergence dates relied upon human monitoring to determine their occurrence, which was likely not universal for all plants within a plot. Similarly, harvesting often happened over several days, yet was recorded within the dataset as the last day of harvesting for the whole plot. Such error seems unavoidable through overall-plot analysis, though alternative estimation methods not reliant on humans could replace this (Delerce et al., 2016).

Furthermore, the data availability covered approximately five harvests per plot. For a more robust analysis, it would be beneficial to cover a greater timeframe, especially considering the unpredictability brought by El Niño and La Niña weather events, and their resulting impact on Colombian rice production (Esquivel et al., 2018; Cai et al., 2020). This reinforces the importance of farms maintaining thorough and accurate agricultural datasets to maximise the PA modelling capabilities (Delerce et al., 2016). Specifically, a greater data volume across a wider period would have assisted in differentiating cultivar performance and overall prediction to a higher degree. As argued by Halevy et al. (2009), the volume of useful data is more valuable for maximising model performance than the complexity of the algorithm used.

The limitations of climate variables is also worth exploring; though the influence of seedling period and drought are of interest given their literary backing, evidence suggests that further performance improvements are likely if climate data were divided by phenology, instead of encompassing the entire growing period (Fageria, 2007; Heinemann et al., 2015; Delerce et al., 2016). Using just remotely sensed data, the present investigation found that the reproductive stage is most correlative to yield values, therefore rice yield can be predicted to the greatest accuracy approximately 1 to 2 months prior to harvest. If climate variables were to be divided in the same manner, model robustness and performance would likely increase, alongside the ability to identify the most suitable phenology with all variables (Fageira, 2007; Heinemann et al., 2015; Delerce et al., 2016). However, Meteoblue was the only available source collating daily climate data, while weather stations surrounding the study area supplied monthly information, lessening the opportunity for phenology division. Thus, supplementary daily climate data would benefit the investigation (Fageira, 2007; Heinemann et al., 2015). Such an approach is supported by past research; Heinemann et al. (2015) found drought at the reproductive phenology the leading factor impacting rice yield in neighbouring Brazil. Here, Heienemann et al. (2015) stated the need for phenology splitting when examining climate variable's impact on growth, with specific mention of drought, a significant influence on yield in this investigation. Indeed, Fageira (2007), also noted that abiotic factors such as radiation, temperature, and drought are ultimately most impactful to rice yield during the vegetative and reproductive stages, suggesting that results in the present research could be further optimised through phenological division.

Chapter 5

6.0. Conclusions

6.1. Cloud mitigation and rice yield prediction

Optical satellite data is clearly a valuable component in tropical PA, but availability can be limited due to persistent cloud coverage. Following extreme climate projections and concerns surrounding regional food security, it is essential that these information gaps are filled with accurate data. During this investigation, cloud-penetrating Sentinel-1 metrics, specifically VV, VH, and NRPB, were modelled alongside NDVI, EVI, and SAVI generated from optical Sentinel-2 data. Metrics were divided into phenology due to evidence of varied results from these periods (Filgueiras et al., 2019, Zhang et al., 2019b). Correlations for cloud mitigation proved encouraging at reproductive and vegetative stages (R^2 0.578 and 0.583, respectively) through machine learning techniques, using NDVI, VV, and NRPB metrics. Though these were not sufficient to mitigate regional cloud cover in a robust manner, findings allowed exploration for further development. Namely, elevated error rates at the vegetative stage resulted from substantial plant structural changes during this growth period, alongside varied influence of dielectric properties during open and closed canopy coverage (Filgueiras et al., 2019; Zhang et al., 2019b). Resultingly, this investigation proposes that more precise phenology division at the vegetative stage, specifically during maximum tillering and stem elongation, may reduce error fluctuations and improve cloud mitigation (Filgueiras et al., 2019, Zhang et al., 2019b). Here, the impact of soil moisture and corresponding dielectric properties has little influence on backscatter values, as the closed canopy reduces soil exposure (Bousbih et al., 2017; Filgueiras et al., 2019).

Additionally, rice yield prediction was explored through machine learning, harnessing EO metrics, climate variables, and *in situ* field measurements. Effective yield prediction within the tropics is an important research avenue for regional food security, following Castro Llanos et al.'s (2019) investigation, which established only 40% of current rice plots in Colombia will be appropriate for production by 2050. The current study area has sufficient elevation to be classified as suitable, thus establishing a robust prediction model in Hacienda El Escobal is crucial for food security and extrapolation to wider areas. Further, the ability to predict yield rates several months in advance can prove paramount to meet demands of both regional farmers and national food security (Noureldin et al., 2013).

Findings demonstrated an overall-plot approach had higher yield prediction capabilities compared to the use of GPS yield samples using VI values, likely due to Sentinel-2's spatial resolution. However, maximum performance of R^2 0.316 at the reproductive stage suggests further variables were required to improve prediction capabilities. Following prior research (Delerce et al. 2016), accounting for different cultivars improved performance metrics dramatically; notably strong collinearity was displayed by Escobal 518 (R^2 0.901), Maja 6 (R^2 0.853), Orizca (R^2 0.744), and Triunfo (R^2 0.687). Reproductive stage EVI values proved most advantageous to modelling, whereby plant booting has previously proven a strong indicator of rice yield (Zhou et al., 2017; Wang et al., 2019a). Additionally, the inclusion of climate variables and *in situ* field measurements reinforced prediction capabilities in some circumstances, including Escobal 518 (R^2 0.949), Triunfo (R^2 0.697), and Fedearroz 68 (R^2 0.551), whereby reproductive stage EVI, drought information, and seedling period proved most impactful.

Feature importance analysis allowed for exploration of variables impact, where the influence of drought and seedling data corresponded to previous investigations within proximity to the study area, strengthening obtained results (Delerce et al., 2016; Quevedo-Amaya et al., 2020). This allowed identification of cultivars more resilient to projected climate scenarios, which is important information for managerial decision-making in the region. For example, Escobal 518's reduced seedling period influence compared to Triunfo suggests it is a more appropraite cultivar as climate change impacts Colombia, whereby more extreme temperatures and lower precipitation can prolong the rice seedling stage (Delerce et al., 2016; Quevedo et al., 2020), in turn reducing yield volume (Gan et al., 1992). However, the dominance of EO metrics during all feature importance analysis suggests a more precise phenology division for climate variables could uncover supplementary correlations.

The combination of satellite-borne rice yield prediction and cloud mitigation determined during this investigation provides valuable results to ensure regional food security, which would benefit from further exploration. These strategies are an essential component to strengthen rice producing regions in Colombia and surrounding tropical areas, especially with the onset of increasing climate extremes (Castro-Llanos et al., 2019; Weiss et al., 2020). Future avenues of research are therefore proposed.

6.2. Future Research Avenues

More precise phenological splitting in all aspects of the investigation, including cloud cover mitigation and climate variables during yield prediction, would positively influence the results.

Regarding cloud mitigation, both the findings and wider literature suggest the division of the vegetative stage into more precise periods may improve performance from R^2 0.583, while reducing error rates (Bousbih et al., 2017; Filgueiras et al., 2019; Zhang et al., 2019b). Currently, the vegetative phenology experiences large-scale structural change from initial emergence to full stem elongation, during which the canopy is both open and closed. Resulting backscatter demonstrates strong fluctuations due to the changing plant architecture and dielectric properties from soil moisture, reflecting the high error rates. By focussing on maximum tillering and stem elongation, soil moisture will likely have a lessened impact on backscatter values due to canopy closure (Filgueiras et al., 2019; Zhang et al., 2019b). Therefore, more precise division of the vegetative stage could enhance cloud mitigation results above R^2 0.583, enabling extrapolation to other areas of Colombia and other tropical regions, where filling missing data values is crucial (Castro-Llanos et al., 2019; Jiménez et al., 2019; Weiss et al., 2020). Such precision was beyond the scope of this investigation, though the introduction and continued development of satellite constellations such as those provided by Planet Labs may combat the somewhat insufficient spatial temporal frequency in the future (Houborg and McCabe, 2016; Houborg and McCabe, 2018).

A similar direction should be explored during rice yield prediction, where division of climate variables into phenological stages to match EO data could improve results. This is demonstrated through feature importance analysis, which clearly presents EO metrics as the dominant predictor variable for all models. Moreover, drought and seedling variables alongside remotely sensed metrics improve yield prediction in most circumstances, achieving a maximum R^2 value of 0.949. EO metrics show yield forecasting is possible one to two months prior to harvest with EVI data during the reproductive period, garnering a maximum R^2 of 0.901. By including climate variables also divided by phenology, prediction accuracy will likely be strengthened, with added capacity to pinpoint the best stage of plant development for yield prediction (Fageira, 2007; Heinemann et al., 2015, Delerce et al., 2016). This direction is supported by prior investigations (Fageira, 2007; Heinemann et al., 2015, Delerce et al., 2016; Quevedo Amaya et al., 2019), whereby the combination of both remotely sensed and climate variables split by phenology would allow more robust identification of specific time periods best suited for yield prediction.

Extrapolation of methods for use across a wider area, alongside development of a user-focused design for farmers, could prove an efficient path for local-level adoption and impact. Sotelo et al. (2020) curated an extensive, user-focused climate application service for rice and maize

cultivation in Colombia, and the integration of remotely sensed metrics could provide additional advantages, judging by its high feature importance demonstrated in Hacienda El Escobal. An integration of remotely sensed data and climate variables into Sotelo et al.'s (2020) user-focused application could be a pivotal step in combating the food insecurity that the region faces. Ultimately, communication of the investigation's methods and findings are key to ensuring actual adoption to benefit managerial decision-making (Whelan and Taylor, 2013; Young and Verhulst, 2017; Sotelo et al., 2020). Thus, this investigation has potential to be adapted into a wider tool to provide a user-focused system for rice cultivation using multiple data sources.

References

Alexander, D.L., Tropsha, A. and Winkler, D.A. (2015). Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of chemical information and modeling*, *55*(7), pp.1316-1322.

Ali, S., Liu, Y., Ishaq, M., Shah, T., Ilyas, A. and Din, I.U. (2017). Climate change and its impact on the yield of major food crops: Evidence from Pakistan. *Foods*, *6*(6), p.39.

Alvino, A. and Marino, S. (2017). Remote Sensing for Irrigation of Horticultural Crops. *Horticulturae*, 3(2), p.40.

Arango-Londoño, D., Ramírez-Villegas, J., Barrios-Pérez, C., Bonilla-Findji, O., Jarvis, A. and Uribe, J.M. (2020). Cierre de brechas de rendimiento en los sistemas colombianos de siembra directa de arroz: un análisis de frontera estocástica/Closing yield gaps in Colombian direct seeding rice systems: a stochastic frontier analysis. *Agronomía Colombiana*, *38*(1).

Areal, F.J., Jones, P.J., Mortimer, S.R. and Wilson, P. (2018). Measuring sustainable intensification: Combining composite indicators and efficiency analysis to account for positive externalities in cereal production. *Land use policy*, *75*, pp.314-326.

Ariza, A.A. (2019). *Machine Learning and Big Data Techniques for Satellite-Based Rice Phenology Monitoring* (Doctoral dissertation, The University of Manchester (United Kingdom)).

Arslan, S. and Colvin, T.S. (2002). Grain yield mapping: Yield sensing, yield reconstruction, and errors. *Precision Agriculture*, *3*(2), pp.135-154.

Azzari, G., Jain, M. and Lobell, D.B. (2017). Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. *Remote Sensing of Environment*, *202*, pp.129-141.

Barnes, E.M. and Baker, M.G. (2000). Multispectral data for mapping soil texture: possibilities and limitations. *Applied Engineering in Agriculture*, *16*(6), p.731.

Bastiaanssen, W.G., Molden, D.J. and Makin, I.W. (2000). Remote sensing for irrigated agriculture: examples from research and possible applications. *Agricultural water management*, *46*(2), pp.137-155.

Bauer, M.E. and Cipra, J.E. (1973). Identification of agricultural crops by computer processing of ERTS MSS data.

Becker-Reshef, I., Vermote, E., Lindeman, M. and Justice, C. (2010). A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. *Remote sensing of environment*, *114*(6), pp.1312-1323.

Benedict, H.M. and Swidler, R. (1961). Nondestructive method for estimating chlorophyll content of leaves. *Science*, *133*(3469), pp.2015-2016.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, *13*(1), pp.281-305.

Birth, G.S. and McVey, G.R. (1968). Measuring the color of growing turf with a reflectance spectrophotometer 1. *Agronomy Journal*, *60*(6), pp.640-643.

Biswas, R. and Bhattacharyya, B. (2019). Rice yield prediction in lower Gangetic Plain of India through multivariate approach and multiple regression analysis. *Journal of Agrometeorology*, *21*(1), pp.101-103.

Blacklow, W.M. (1972). Influence of Temperature on Germination and Elongation of the Radicle and Shoot of Corn (Zea mays L.) 1. *Crop Science*, *12*(5), pp.647-650.

Blackmore, S. (1999). Remedial correction of yield map data. *Precision agriculture*, *1*(1), pp.53-66.

Bolton, D.K. and Friedl, M.A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, *173*, pp.74-84.

Bousbih, S., Zribi, M., Lili-Chabaane, Z., Baghdadi, N., El Hajj, M., Gao, Q. and Mougenot, B. (2017). Potential of Sentinel-1 radar data for the assessment of soil and cereal cover parameters. *Sensors*, *17*(11), p.2617.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), pp.5-32.

Bunting, P. (2014) Introduction to ARCSI for Generating Analysis Ready Data (ARD), Aberystywyth, available: https://www.arcsi.remotesensing.info/tutorials/ARCSI_Intro_Tutorial_compress.pdf [Accessed 29 June 2020].

Bunting, P., Clewley, D., Lucas, R M., and Gillingham, S. (2014). The Remote Sensing and GIS Software Library (RSGISLib), Computers & Geosciences. Volume 62, pp.216-226.

Campos-Taberner, M., García-Haro, F., Camps-Valls, G., Grau-Muedra, G., Nutini, F., Busetto, L., Katsantonis, D., Stavrakoudis, D., Minakou, C., Gatti, L., Barbieri, M., Holecz, F., Stroppiana, D. and Boschetti, M. (2017). Exploitation of SAR and Optical Sentinel Data to Detect Rice Crop and Estimate Seasonal Dynamics of Leaf Area Index. *Remote Sensing*, 9(3), p.248.

Campos-Taberner, M., García-Haro, F.J., Camps-Valls, G., Grau-Muedra, G., Nutini, F., Crema, A. and Boschetti, M. (2016). Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring. *Remote Sensing of Environment*, *187*, pp.102-118.

Carlson, T.N. and Ripley, D.A. (1997). On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote sensing of Environment*, *62*(3), pp.241-252.

Castilla-Lozano LA, Vanegas J, Rodriguez J, Uribe-Velez D. (2011). Evaluación de la fertilidad en suelos de las zonas arroceras de Tolima y Meta. In: Uribe-Velez D, Melgarejo LM, editors. Ecología de microorganismos rizosféricos asociados a cultivos de arroz de Tolima y Meta. Bogotá. p.33–52.

Castro-Llanos, F., Hyman, G., Rubiano, J., Ramirez-Villegas, J. and Achicanoy, H. (2019). Climate change favors rice production at higher elevations in Colombia. *Mitigation and Adaptation Strategies for Global Change*, 24(8), pp.1401-1430.

Chang, K.W., Shen, Y. and Lo, J.C. (2005). Predicting rice yield using canopy reflectance measured at booting stage. *Agronomy Journal*, *97*(3), pp.872-878.

Chavez, P.S. (1996). Image-based atmospheric corrections-revisited and improved. *Photogrammetric engineering and remote sensing*, *62*(9), pp.1025-1035.

Chen, J.M. and Black, T.A. (1992). Defining leaf area index for non-flat leaves. *Plant, Cell & Environment*, *15*(4), pp.421-429.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794.

Cheng, Q. and Wu, X. (2011). Mapping paddy rice yield in Zhejiang Province using MODIS spectral index. *Turkish Journal of Agriculture and Forestry*, *35*(6), pp.579-589.

Chlingaryan, A., Sukkarieh, S. and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, pp.61-69.

Clay, D.E., Kim, K.I., Chang, J., Clay, S.A. and Dalsted, K. (2006). Characterizing water and nitrogen stress in corn using remote sensing. *Agronomy Journal*, *98*(3), pp.579-587.

Clerici, N., Valbuena Calderón, C.A. and Posada, J.M. (2017). Fusion of Sentinel-1A and Sentinel-2A data for land cover mapping: a case study in the lower Magdalena region, Colombia. *Journal of Maps*, *13*(2), pp.718-726.

Clevers, J.G. and 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*, pp.344-351.

Confalonieri, R., Acutis, M., Bellocchi, G., Cerrani, I., Tarantola, S., Donatelli, M. and Genovese, G. (2006). Exploratory sensitivity analysis of CropSyst, WARM and WOFOST: a case-study with rice biomass simulations.

Confalonieri, R., Rosenmund, A.S. and Baruth, B. (2009). An improved model to simulate rice yield. *Agronomy for Sustainable Development*, *29*(3), pp.463-474.

Dammalage, T.L. and Shanmugam, T. (2018). Use of Satellite Remote Sensing for Rice Yield Estimation: A Case Study of Polonnaruwa District, Sri Lanka. *Asian Journal of Advances in Agricultural Research*, pp.1-9.

Das, S., Kumar, A., Barman, M., Pal, S. and Bandopadhyay, P. (2020). Impact of Climate Variability on Phenology of Rice. In *Agronomic Crops* (pp. 13-28). Springer, Singapore.

De Leeuw, J., Vrieling, A., Shee, A., Atzberger, C., Hadgu, K.M., Biradar, C.M., Keah, H. and Turvey, C. (2014). The potential and uptake of remote sensing in insurance: A review. *Remote Sensing*, *6*(11), pp.10888-10912.

Deisenroth, M., Faisal, A. and Ong, C. (2020). *Mathematics For Machine Learning*. Cambridge University Press, pp.1-417.

Delerce, S., Dorado, H., Grillon, A., Rebolledo, M., Prager, S., Patiño, V., Garcés Varón, G. and Jiménez, D. (2016). Assessing Weather-Yield Relationships in Rice at Local Scale Using Data Mining Approaches. *PLOS ONE*, 11(8), pp.1-25.

Delloye, C., Weiss, M. and 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*, pp.245-261.

Dey, U.K., Masud, A.H. and Uddin, M.N. (2017). Rice yield prediction model using data mining. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE,* pp.321-326.

Di Falco, S., Yesuf, M., Kohlin, G. and Ringler, C. (2012). Estimating the Impact of Climate Change on Agriculture in Low-Income Countries: Household Level Evidence from the Nile Basin, Ethiopia. *Environmental and Resource Economics*, 52(4), pp.457-478.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J. and Münkemüller, T. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), pp.27-46.

Duan, B., Fang, S., Zhu, R., Wu, X., Wang, S., Gong, Y. and Peng, Y. (2019). Remote estimation of rice yield with unmanned aerial vehicle (UAV) data and spectral mixture analysis. *Frontiers in plant science*, *10*, p.204.

Elescobal. (n.d.). *Hacienda – El Escobal*. [online] Available at: https://elescobal.com/ [Accessed 29 June 2020].

Esquivel, A., Llanos-Herrera, L., Agudelo, D., Prager, S.D., Fernandes, K., Rojas, A., Valencia, J.J. and Ramirez-Villegas, J. (2018). Predictability of seasonal precipitation across major crop growing areas in Colombia. *Climate Services*, *12*, pp.36-47.

Fageria, N.K. (2007). Yield physiology of rice. *Journal of plant nutrition*, *30*(6), pp.843-879.

FAO. (2017). The future of food and agriculture–Trends and challenges. *Annual Report*.

FAO. (2020). *FAOSTAT*. [online] Fao.org. Available at: <http://www.fao.org/faostat/en/#data/QC/visualize> [Accessed 28 June 2020].

Fernandes, K., Muñoz, A.G., Ramirez-Villegas, J., Agudelo, D., Llanos-Herrera, L., Esquivel, A., Rodriguez-Espinoza, J. and Prager, S.D. (2020). Improving Seasonal precipitation forecasts for agriculture in the Orinoquía Region of Colombia. *Weather and Forecasting*, *35*(2), pp.437-449.

Filgueiras, R., Mantovani, E., Althoff, D., Fernandes Filho, E. and Cunha, F. (2019). Crop NDVI Monitoring Based on Sentinel 1. *Remote Sensing*, 11(12), p.1441.

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D., Pozza, L.E., Ugbaje, S.U., Jephcott, T.G., Paterson, S.E., Whelan, B.M. and Bishop, T.F. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, *20*(5), pp.1015-1029.

Filipponi, F. (2019). Sentinel-1 GRD Preprocessing Workflow. In *Multidisciplinary Digital Publishing Institute Proceedings*. Vol. 18, No. 1, p.11.

Finger, R., Swinton, S.M., El Benni, N. and Walter, A. (2019). Precision farming at the nexus of agricultural production and the environment.

Flickr. (2010). *Flickr - CIAT*. [online] Available at: <https://www.flickr.com/photos/ciat/4326133080/> [Accessed 8 September 2020].

Gan, Y., Stobbe, E.H. and Moes, J. (1992). Relative date of wheat seedling emergence and its impact on grain yield. *Crop Science*, *32*(5), pp.1275-1281.

Gandhi, N., Petkar, O. and Armstrong, L.J. (2016). Rice crop yield prediction using artificial neural networks. In *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, IEEE, pp.105-110.

Garbulsky, M.F., Peñuelas, J., Gamon, J., Inoue, Y. and Filella, I. (2011). The photochemical reflectance index (PRI) and the remote sensing of leaf, canopy and ecosystem radiation use efficiencies: A review and meta-analysis. *Remote sensing of environment*, *115*(2), pp.281-297.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gerstmann, H., Möller, M. and Gläßer, C. (2016). Optimization of spectral indices and long-term separability analysis for classification of cereal crops using multi-spectral

RapidEye imagery. *International journal of applied earth observation and geoinformation*, *52*, pp.115-125.

Gilardelli, C., Stella, T., Confalonieri, R., Ranghetti, L., Campos-Taberner, M., García-Haro, F.J. and Boschetti, M. (2019). Downscaling rice yield simulation at sub-field scale using remotely sensed LAI data. *European journal of agronomy*, *103*, pp.108-116.

González-Betancourt, M. and Mayorga-Ruíz, Z.L. (2018). Normalized difference vegetation index for rice management in El Espinal, Colombia. *Dyna*, *85*(205), pp.47-56.

Haagsma, M. (2015). *Crop Monitoring With Radar - Correlation Between SAR Polarimetric Response And Vegetation Indices*. Delft University of Technology, pp.1-127.

Halevy, A., Norvig, P. and Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), pp.8-12.

He, Z., Li, S., Wang, Y., Dai, L. and Lin, S. (2018). Monitoring rice phenology based on backscattering characteristics of multi-temporal RADARSAT-2 datasets. *Remote Sensing*, *10*(2), p.340.

Heinemann, A.B. and Sentelhas, P.C. (2011). Environmental group identification for upland rice production in central Brazil. *Scientia Agrícola*, *68*(5), pp.540-547.

Heinemann, A.B., Barrios-Perez, C., Ramirez-Villegas, J., Arango-Londoño, D., Bonilla-Findji, O., Medeiros, J.C. and Jarvis, A. (2015). Variation and impact of drought-stress patterns across upland rice target population of environments in Brazil. *Journal of experimental botany*, *66*(12), pp.3625-3638.

Houborg, R. and McCabe, M. (2016). High-resolution NDVI from Planet's constellation of earth observing nano-satellites: a new data source for precision agriculture. *Remote Sensing*, *8*(9), p.768.

Houborg, R. and McCabe, M. (2018). Daily Retrieval of NDVI and LAI at 3 m Resolution via the Fusion of CubeSat, Landsat, and MODIS Data. *Remote Sensing*, *10*(6), p.890.

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

Iizuka, K., Hayakawa, Y.S., Ogura, T., Nakata, Y., Kosugi, Y. and Yonehara, T. (2020). Integration of Multi-Sensor Data to Estimate Plot-Level Stem Volume Using Machine Learning Algorithms–Case Study of Evergreen Conifer Planted Forests in Japan. *Remote Sensing*, *12*(10), p.1649.

Iizumi, T., Luo, J.J., Challinor, A.J., Sakurai, G., Yokozawa, M., Sakuma, H., Brown, M.E. and Yamagata, T. (2014). Impacts of El Niño Southern Oscillation on the global yields of major crops. *Nature communications*, *5*(1), pp.1-7.

Iizumi, T., Shin, Y., Kim, W., Kim, M. and Choi, J. (2018). Global crop yield forecasting using seasonal climate information from a multi-model ensemble. *Climate Services*, *11*, pp.13-23.

Jaikla, R., Auephanwiriyakul, S. and Jintrawet, A. (2008). Rice yield prediction using a support vector regression method. In *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, IEEE, Vol. 1, pp.29-32.

Jensen, R.D. (1971). Effects of Soil Water Tension on the Emergence and Growth of Cotton Seedlings 1. *Agronomy Journal*, *63*(5), pp.766-768.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R. and Kim, S.H. (2016). Random forests for global and regional crop yield predictions. *PLoS One*, *11*(6), p.e0156571.

Ji, B., Sun, Y., Yang, S. and Wan, J. (2007). Artificial neural networks for rice yield prediction in mountainous regions. *The Journal of Agricultural Science*, *145*(3), p.249.

Jiang, J., Chen, S., Cao, S., Wu, H., Zhang, L. and Zhang, H. (2005). Leaf area index retrieval based on canopy reflectance and vegetation index in eastern China. *Journal of Geographical Sciences*, *15*(2), pp.247-254.

Jiménez, D., Delerce, S., Dorado, H., Cock, J., Muñoz, L.A., Agamez, A. and Jarvis, A. (2019). A scalable scheme to implement data-driven agriculture for small-scale farmers. *Global Food Security*, *23*, pp.256-266.

Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A. and Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and forest meteorology*, *218*, pp.74-84.

Jordan, C.F. (1969). Derivation of leaf-area index from quality of light on the forest floor. *Ecology*, *50*(4), pp.663-666.

Joshi, N., Baumann, M., Ehammer, A., Fensholt, R., Grogan, K., Hostert, P., Jepsen, M.R., Kuemmerle, T., Meyfroidt, P., Mitchard, E.T. and Reiche, J. (2016). A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. *Remote Sensing*, *8*(1), p.70.

Kamir, E., Waldner, F. and Hochman, Z. (2020). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, *160*, pp.124-135.

Kim, N. and Lee, Y.W. (2016). Machine learning approaches to corn yield estimation using satellite images and climate data: a case of Iowa State. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, *34*(4), pp.383-390.

Knipling, E.B. (1970). Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. *Remote sensing of environment*, *1*(3), pp.155-159.

Kogan, F. (2019). Vegetation health for insuring drought-related yield losses and food security enhancement. In *Remote Sensing for Food Security*, Springer, pp. 163-173.

Kogan, F., Guo, W. and Yang, W. (2019). Drought and food security prediction from NOAA new generation of operational satellites. *Geomatics, Natural Hazards and Risk*, *10*(1), pp.651-666.

Kuenzer, C. and Knauer, K. (2013). Remote sensing of rice crop areas. *International Journal of Remote Sensing*, *34*(6), pp.2101-2139.

Lane, A. and Jarvis, A. (2007). Changes in climate will modify the geography of crop suitability: agricultural biodiversity can help with adaptation. *International Crops Research Institute for the Semi-Arid Tropics*, 4(1), pp.1-12.

Leroux, C., Jones, H., Clenet, A., Dreux, B., Becu, M. and Tisseyre, B. (2018). A general method to filter out defective spatial observations from yield mapping datasets. *Precision Agriculture*, *19*(5), pp.789-808.

Li, S., Ganguly, S., Dungan, J.L., Wang, W. and Nemani, R.R. (2017a). Sentinel-2 MSI radiometric characterization and cross-calibration with Landsat-8 OLI. *Advances in Remote Sensing*, *6*(02), p.147.

Li, T., Angeles, O., Marcaida III, M., Manalo, E., Manalili, M.P., Radanielson, A. and Mohanty, S. (2017b). From ORYZA2000 to ORYZA (v3): An improved simulation model for rice in drought and nitrogen-deficient environments. *Agricultural and forest meteorology*, *237*, pp.246-256.

Li, X., Qian, Q., Fu, Z., Wang, Y., Xiong, G., Zeng, D., Wang, X., Liu, X., Teng, S., Hiroshi, F. and Yuan, M. (2003). Control of tillering in rice. *Nature*, *422*(6932), pp.618-621.

Liakos, K., Busato, P., Moshou, D., Pearson, S. and Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, 18(8), p.2674.

Lillesand, T., Kiefer, R. and Chipman, J. (2015). *Remote sensing and image interpretation*. 7th ed. Hoboken: John Wiley & Sons, pp.1-680.

Lin, W., Zhang, F.C., Jing, Y.S., Jiang, X.D., Yang, S.B. and HAN, X.M. (2014). Multi-temporal detection of rice phenological stages using canopy spectrum. *Rice Science*, *21*(2), pp.108-115.

Liu, H.Q. and Huete, A. (1995). A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. *IEEE transactions on geoscience and remote sensing*, *33*(2), pp.457-465.

Lobell, D.B., Thau, D., Seifert, C., Engle, E. and Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, pp.324-333.

López-Granados, F. (2011). Weed detection for site-specific weed management: mapping and real-time approaches. *Weed Research*, *51*(1), pp.1-11.

Mansaray, L.R., Wang, F., Kanu, A.S. and Yang, L. (2020). Evaluating Sentinel-1A datasets for rice leaf area index estimation based on machine learning regression models. *Geocarto International*, (just-accepted), pp.1-11.

Marshall, M., Thenkabail, P., Biggs, T. and Post, K. (2016). Hyperspectral narrowband and multispectral broadband indices for remote sensing of crop evapotranspiration and its components (transpiration and soil evaporation). *Agricultural and forest meteorology*, *218*, pp.122-134.

Mather, P. and Koch, M. (2011). *Computer processing of remotely sensed images*. 4th ed. Chichester, West Sussex, England: John Wiley & Sons.

Meinke, H. and Stone, R. (2005). Seasonal and Inter-Annual Climate Forecasting: The New Tool for Increasing Preparedness to Climate Variability and Change In Agricultural Planning And Operations. *Climatic Change*, 70(1-2), pp.221-253.

Mercier, A., Betbeder, J., Baudry, J., Denize, J., Leroux, V., Roger, J.L., Spicher, F. and Hubert-Moy, L. (2019). Evaluation of Sentinel-1 and-2 time series to derive crop phenology and biomass of wheat and rapeseed: northen France and Brittany case studies. In *Remote Sensing for Agriculture, Ecosystems, and Hydrology XXI*, International Society for Optics and Photonics, (Vol. 11149, p.1114903.

Meteoblue. (n.d.). *Weather History+*. [online] meteoblue. Available at: <https://www.meteoblue.com/en/historyplus> [Accessed 9 October 2020].

Mishra, S., Mishra, D. and Santra, G. (2016). Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper. *Indian Journal of Science and Technology*, 9(38).

Mohanty, M. and Painuli, D.K. (2004). Modeling rice seedling emergence and growth under tillage and residue management in a rice–wheat system on a Vertisol in Central India. *Soil and Tillage Research*, *76*(2), pp.167-174.

Moldenhauer, K.E.W.C. and Slaton, N. (2001). Rice growth and development. *Rice production handbook*, *192*, pp.7-14.

Monteith, J.L. (1977). Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *281*(980), pp.277-294.

Mosleh, M.K., Hassan, Q.K. and Chowdhury, E.H. (2015). Application of remote sensors in mapping rice area and forecasting its production: A review. *Sensors*, *15*(1), pp.769-791.

Mulla, D.J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, *114*(4), pp.358-371.

Munibah, K., Barus, B., Tjahjono, B., Wijayanti, R.S., Mufti, B. and Hongo, C. (2019). Utilization of Sentinel-2 imagery to identify a growth phase of rice plant in Cianjur Regency, West Java, Indonesia. In *Sixth International Symposium on LAPAN-IPB Satellite,* International Society for Optics and Photonics, 11372, p.1137203.

Nasrallah, A., Baghdadi, N., El Hajj, M., Darwish, T., Belhouchette, H., Faour, G., Darwich, S. and Mhawej, M. (2019). Sentinel-1 Data for Winter Wheat Phenology Monitoring and Mapping. *Remote Sensing*, *11*(19), p.2228.

Noureldin, N.A., Aboelghar, M.A., Saudy, H.S. and Ali, A.M. (2013). Rice yield forecasting models using satellite imagery in Egypt. *The Egyptian Journal of Remote Sensing and Space Science*, *16*(1), pp.125-131.

Oguntunde, P.G., Lischeid, G. and Dietrich, O. (2018). Relationship between rice yield and climate variables in southwest Nigeria using multiple linear regression and support vector machine analysis. *International journal of biometeorology*, *62*(3), pp.459-469.

Olson, R.S., Moore, J.H. (2016). Tpot: A tree-based pipeline optimization tool for automating machine learning. In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds.) Proceedings of the Workshop on Automatic Machine Learning. Proceedings of Machine Learning Research, vol. 64, pp.66-74.

Onojeghuo, A., Blackburn, G., Huang, J., Kindred, D. and Huang, W. (2018). Applications of satellite 'hyper-sensing' in Chinese agriculture: Challenges and opportunities. *International Journal of Applied Earth Observation and Geoinformation*, 64, pp.62-86.

Pachauri, R.K., Allen, M.R., Barros, V.R., Broome, J., Cramer, W., Christ, R., Church, J.A., Clarke, L., Dahe, Q., Dasgupta, P. and Dubash, N.K. (2014). *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change, IPCC,* p.151.

Panda, S.S., Ames, D.P. and Panigrahi, S. (2010). Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing*, *2*(3), pp.673-696.

Panek, E. and Gozdowski, D. (2020). Analysis of relationship between cereal yield and NDVI for selected regions of Central Europe based on MODIS satellite data. *Remote Sensing Applications: Society and Environment*, *17*, p.100286.

Patil, P., Biradar, P., Bhagawathi, A.U., Hejjegar. I.S. (2018). A Review on Leaf Area Index of Horticulture Crops and Its Importance. *International Journal of Current Microbiology and Applied Sciences,* 7(4), pp. 505–513.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.

Poveda, G., Jaramillo, A., Gil, M.M., Quiceno, N. and Mantilla, R.I. (2001). Seasonally in ENSO-related precipitation, river discharges, soil moisture, and vegetation index in Colombia. *Water resources research*, *37*(8), pp.2169-2178.

Pringle, M.J., McBratney, A.B., Whelan, B.M. and Taylor, J.A. (2003). A preliminary approach to assessing the opportunity for site-specific crop management in a field, using yield monitor data. *Agricultural Systems*, *76*(1), pp.273-292.

Quevedo Amaya, Y.M., Beltrán Medina, J.I. and Barragán Quijano, E. (2019). Identification of climatic and physiological variables associated with rice (Oryza sativa L.) yield under tropical conditions. *Revista Facultad Nacional de Agronomía Medellín*, *72*(1), pp.8699-8706.

Quevedo-Amaya, Y.M., Beltrán-Medina, J.I., Hoyos-Cartagena, J.Á., Calderón-Carvajal, J.E. and Barragán-Quijano, E. (2020). Selection of sowing date and biofertilization as alternatives to improve the yield and profitability of the F68 rice variety. *Agronomía Colombiana*, *38*(1).

Ramankutty, N., Foley, J., Norman, J. and McSweeney, K. (2002). The global distribution of cultivable lands: current patterns and sensitivity to possible climate change. *Global Ecology and Biogeography*, 11(5), pp.377-392.

Rasmussen, M.S. (1997). Operational yield forecast using AVHRR NDVI data: reduction of environmental and inter-annual variability. *International Journal of Remote Sensing*, *18*(5), pp.1059-1077.

Rees, W.G. (2013). *Physical principles of remote sensing*. Cambridge university press.

Rouse Jr, J.W., Haas, R.H., Schell, J.A. and Deering, D.W. (1974). Paper A 20. In *Third Earth Resources Technology Satellite-1 Symposium: The Proceedings of a Symposium Held by Goddard Space Flight Center at Washington, DC on December 10-14, 1973: Prepared at Goddard Space Flight Center* (Vol. 351, p. 309). Scientific and Technical Information Office, National Aeronautics and Space Administration.

Sakamoto, T., Shibayama, M., Kimura, A. and Takada, E. (2011). Assessment of digital camera-derived vegetation indices in quantitative monitoring of seasonal rice growth. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(6), pp.872-882.

Sarker, M.A.R., Alam, K. and Gow, J. (2012). Exploring the relationship between climate change and rice yield in Bangladesh: An analysis of time series data. *Agricultural Systems*, *112*, pp.11-16.

Satir, O. and Berberoglu, S. (2016). Crop yield prediction under soil salinity using satellite derived vegetation indices. *Field crops research*, *192*, pp.134-143.

Schneider, E.C. and Gupta, S.C. (1985). Corn emergence as influenced by soil temperature, matric potential, and aggregate size distribution. *Soil Science Society of America Journal*, *49*(2), pp.415-422.

Scihub.copernicus.eu. (n.d.). [online] Available at: https://scihub.copernicus.eu/dhus/#/home [Accessed 29 June 2020].

Sellers, P.J., Berry, J.A., Collatz, G.J., Field, C.B. and Hall, F.G. (1992). Canopy reflectance, photosynthesis, and transpiration. III. A reanalysis using improved leaf models and a new canopy integration scheme. *Remote sensing of environment*, *42*(3), pp.187-216.

Shah, F., Huang, J., Cui, K., Nie, L., Shah, T., Chen, C. and Wang, K. (2011). Impact of high-temperature stress on rice plant and its traits related to tolerance. *The Journal of Agricultural Science*, *149*(5), pp.545-556.

Shanahan, J.F., Schepers, J.S., Francis, D.D., Varvel, G.E., Wilhelm, W.W., Tringe, J.M., Schlemmer, M.R. and Major, D.J. (2001). Use of remote-sensing imagery to estimate corn grain yield. *Agronomy Journal*, *93*(3), pp.583-589.

Shen, M., Chen, J., Zhu, X., Tang, Y. and Chen, X. (2010). Do flowers affect biomass estimate accuracy from NDVI and EVI?. *International Journal of Remote Sensing*, *31*(8), pp.2139-2149.

Shihua, L., Jingtao, X., Ping, N., Jing, Z., Hongshu, W. and Jingxian, W. (2014). Monitoring paddy rice phenology using time series MODIS data over Jiangxi Province, China. *International Journal of Agricultural and Biological Engineering*, *7*(6), pp.28-36.

Shiu, Y.S. and Chuang, Y.C. (2019). Yield Estimation of Paddy Rice Based on Satellite Imagery: Comparison of Global and Local Regression Models. *Remote Sensing*, *11*(2), p.111.

Shivrain, V.K., Burgos, N.R., Gealy, D.R., Smith, K.L., Scott, R.C., Mauromoustakos, A. and Black, H. (2009). Red rice (Oryza sativa) emergence characteristics and influence on rice yield at different planting dates. *Weed Science*, *57*(1), pp.94-102.

Sivasankar, T., Kumar, D., Srivastava, H.S. and Patel, P. (2018). Advances in radar remote sensing of agricultural crops: a review. *Int. J. Adv. Sci. Eng. Inf. Technol*, *8*, p.1126.

Smola, A.J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*(3), pp.199-222.

SNAP (n.d.). *SNAP | STEP*. [online] Step.esa.int. Available at: <https://step.esa.int/main/toolboxes/snap/> [Accessed 30 June 2020].

Sotelo, S., Guevara, E., Llanos-Herrera, L., Agudelo, D., Esquivel, A., Rodriguez, J., Ordoñez, L., Mesa, J., Borja, L.A.M., Howland, F. and Amariles, S. (2020). Pronosticos AClimateColombia: A system for the provision of information for climate risk reduction in Colombia. *Computers and Electronics in Agriculture*, *174*, p.105486.

Spiegelhalter, D. (2019). *The Art Of Statistics: Learning From Data*. Penguin Books.

Su, Y.X., Xu, H. and Yan, L.J. (2017). Support vector machine-based open crop model (SBOCM): Case of rice production in China. *Saudi journal of biological sciences*, *24*(3), pp.537-547.

Tesfaye, A.A. and Awoke, B.G. (2020). Evaluation of the saturation property of vegetation indices derived from sentinel-2 in mixed crop-forest ecosystem. *Spatial Information Research*, pp.1-13.

Thippani S., Kumar SS., Senguttuvel P., Madhav MS. (2017). Correlation Analysis for Yield and Yield Components in Rice (*Oryza sativa* L.). International Journal of Pure & Applied Bioscience 5(4), pp.1412-1415.

Thorp, K., Wang, G., West, A., Moran, M., Bronson, K., White, J. and Mon, J. (2012). Estimating crop biophysical properties from remote sensing data by inverting linked

radiative transfer and ecophysiological models. *Remote Sensing of Environment*, 124, pp.224-233.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M. and Traver, I.N. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, pp.9-24.

Truckenbrodt, J., Freemantle, T., Williams, C., Jones, T., Small, D., Dubois, C., Thiel, C., Rossi, C., Syriou, A. and Giuliani, G. (2019). Towards Sentinel-1 SAR analysis-ready data: A best practices assessment on preparing backscatter data for the cube. *Data*, *4*(3), p.93.

Tucker, C.J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, *8*(2), pp.127-150.

U.N. (2019). World population prospects 2019: Highlights. *New York (US): United Nations Department for Economic and Social Affairs*.

Van Oort, P.A.J., Zhang, T., De Vries, M.E., Heinemann, A.B. and Meinke, H. (2011). Correlation between temperature and phenology prediction error in rice (Oryza sativa L.). *Agricultural and Forest Meteorology*, *151*(12), pp.1545-1555.

Verger, A., Baret, F. and Weiss, M. (2014b). Near real-time vegetation monitoring at global scale. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(8), pp.3473-3481.

Verger, A., Vigneau, N., Chéron, C., Gilliot, J.M., Comar, A. and Baret, F. (2014a). Green area index from an unmanned aerial system over wheat and rapeseed crops. *Remote Sensing of Environment*, *152*, pp.654-664.

Vicente-Serrano, S.M., Nieto, R., Gimeno, L., Azorin-Molina, C., Drumond, A., El Kenawy, A., Dominguez-Castro, F., Tomas-Burguera, M. and Peña-Gallardo, M. (2018). Recent changes of relative humidity: Regional connections with land and ocean processes.

Vreugdenhil, M., Wagner, W., Bauer-Marschallinger, B., Pfeil, I., Teubner, I., Rüdiger, C. and Strauss, P. (2018). Sensitivity of Sentinel-1 backscatter to vegetation dynamics: An Austrian case study. *Remote Sensing*, *10*(9), p.1396.

Wakamori, K., Ichikawa, D. and Oguri, N. (2017). Estimation of rice growth status, protein content and yield prediction using multi-satellite data. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE,* pp.5089-5092.

Wang, F., Wang, F., Zhang, Y., Hu, J., Huang, J. and Xie, J. (2019a). Rice yield estimation using parcel-level relative spectral variables from UAV-based hyperspectral imagery. *Frontiers in plant science*, *10*, p.453.

Wang, H., Cutforth, H., McCaig, T., McLeod, G., Brandt, K., Lemke, R., Goddard, T. and Sprout, C. (2009). Predicting the time to 50% seedling emergence in wheat using a Beta model. *NJAS-Wageningen Journal of Life Sciences*, *57*(1), pp.65-71.

Wang, J., Dai, Q., Shang, J., Jin, X., Sun, Q., Zhou, G. and Dai, Q. (2019b). Field-Scale Rice Yield Estimation Using Sentinel-1A Synthetic Aperture Radar (SAR) Data in Coastal Saline Region of Jiangsu Province, China. *Remote Sensing*, *11*(19), p.2274.

Wang, Q., Shi, W., Li, Z. and Atkinson, P.M. (2016). Fusion of Sentinel-2 images. *Remote sensing of environment*, *187*, pp.241-252.

Weiss, M., Jacob, F. and Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, 236, p.111402.

Wheeler, T. and von Braun, J. (2013). Climate Change Impacts on Global Food Security. *Science*, 341(6145), pp.508-513.

Whelan, B. and Taylor, J. (2013). *Precision agriculture for grain production systems*. Csiro publishing.

Wiegand, C.L., Richardson, A.J. and Kanemasu, E.T. (1979). Leaf Area Index Estimates for Wheat from LANDSAT and Their Implications for Evapotranspiration and Crop Modeling 1. *Agronomy Journal*, *71*(2), pp.336-342.

Woodhouse, I. (2005). *Introduction to Microwave Remote Sensing*. London: Taylor & Francis.

Wu, W., Wang, W., Meadows, M.E., Yao, X. and Peng, W. (2019). Cloud-based typhoon-derived paddy rice flooding and lodging detection using multi-temporal Sentinel-1&2. *Frontiers of Earth Science*, *13*(4), pp.682-694.

Xue, J. and Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*.

Yaghouti, H., Pazira, E., Amiri, E. and Masihabadi, M.H. (2019). The Feasibility of Using Vegetation Indices and Soil Texture to Predict Rice Yield. *Polish Journal of Environmental Studies*, *28*(4).

Yang, C., Everitt, J.H. and Bradford, J.M. (2006). Comparison of QuickBird satellite imagery and airborne imagery for mapping grain sorghum yield patterns. *Precision Agriculture*, *7*(1), pp.33-44.

Yang, C., Everitt, J.H. and Murden, D. (2011). Evaluating high resolution SPOT 5 satellite imagery for crop identification. *Computers and Electronics in Agriculture*, *75*(2), pp.347-354.

Yang, Z., Li, K., Shao, Y., Brisco, B. and Liu, L. (2016). Estimation of paddy rice variables with a modified water cloud model and improved polarimetric decomposition using multi-temporal RADARSAT-2 images. *Remote Sensing*, *8*(10), p.878.

Yawata, K., Yamamoto, T., Hashimoto, N., Ishida, R. and Yoshikawa, H. (2019). Mixed model estimation of rice yield based on NDVI and GNDVI using a satellite image. In *Remote Sensing for Agriculture, Ecosystems, and Hydrology XXI*, International Society for Optics and Photonics, *V*ol.11149, p.1114918.

Yommy, A.S., Liu, R. and Wu, S. (2015). SAR image despeckling using refined Lee filter. In *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, IEEE*. Vol. 2, pp.260-265.

Yonezawa, C., Negishi, M., Azuma, K., Watanabe, M., Ishitsuka, N., Ogawa, S. and Saito, G. (2012). Growth monitoring and classification of rice fields using multitemporal RADARSAT-2 full-polarimetric data. *International journal of remote sensing*, *33*(18), pp.5696-5711.

Yoshida, S. (1981). Climatic Environment and its influence. Fundamentals of rice crop science, p.65–109.

Young, A. and Verhulst, S. (2017). Aclimate Colombia: Open Data to Improve Agricultural Resiliency. Open Data's Impact. GovLab.

Yzarra Tito, W.J., Lopez Rios, F.M. (2011). Manual de Observaciones Fenólogicas. Servicio Nacional de Meteorolog´ıa e hidrolog´ıa, Lima, Peru.

Zabel, F., Putzenlechner, B. and Mauser, W. (2014). Global Agricultural Land Resources – A High Resolution Suitability Evaluation and Its Perspectives until 2100 under Climate Change Conditions. *PLoS ONE*, 9(9), p.e107522.

Zhang, C. and Kovacs, J.M. (2012). The application of small unmanned aerial systems for precision agriculture: a review. *Precision agriculture*, *13*(6), pp.693-712.

Zhang, H.K., Roy, D.P., Yan, L., Li, Z., Huang, H., Vermote, E., Skakun, S. and Roger, J.C. (2018). Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. *Remote sensing of environment*, *215*, pp.482-494.

Zhang, K., Ge, X., Shen, P., Li, W., Liu, X., Cao, Q., Zhu, Y., Cao, W. and Tian, Y. (2019b). Predicting rice grain yield based on dynamic changes in vegetation indexes during early to mid-growth stages. *Remote Sensing*, *11*(4), p.387.

Zhang, L., Traore, S., Ge, J., Li, Y., Wang, S., Zhu, G., Cui, Y. and Fipps, G. (2019a). Using boosted tree regression and artificial neural networks to forecast upland rice yield under climate change in Sahel. *Computers and Electronics in Agriculture*, *166*, p.105031.

Zhang, T., Su, J., Liu, C., Chen, W.H., Liu, H. and Liu, G. (2017). Band selection in Sentinel-2 satellite for agriculture applications. In *2017 23rd International Conference on Automation and Computing (ICAC)*, *IEEE*, pp.1-6.

Zhao, G., Miao, Y., Wang, H., Su, M., Fan, M., Zhang, F., Jiang, R., Zhang, Z., Liu, C., Liu, P. and Ma, D. (2013). A preliminary precision rice management system for increasing both grain yield and nitrogen use efficiency. *Field Crops Research*, *154*, pp.23-30.

Zheng, H., Cheng, T., Yao, X., Deng, X., Tian, Y., Cao, W. and Zhu, Y. (2016). Detection of rice phenology through time series analysis of ground-based spectral index data. *Field Crops Research*, *198*, pp.131-139.

Zhou, X., Zheng, H.B., Xu, X.Q., He, J.Y., Ge, X.K., Yao, X., Cheng, T., Zhu, Y., Cao, W.X. and Tian, Y.C. (2017). Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, pp.246-255.

Zorrilla, G., Martínez, C., Berrío Orozco, L.E., Corredor, E., Carmona, L. and Pulver, E. (2012). Improving rice production systems in Latin America and the Caribbean. Centro Internacional de Agricultura Tropical (CIAT).

Appendix

Appendix A: The identified corresponding Sentinel-1 and Sentinel-2 scenes captured on the same dates, for use in cloud cover mitigation.

Date	Sentinel-1	Sentinel-2
08/07/2017	S1B_IW_GRDH_1SDV_20170708T104211_20170708T104236_006395_00B3E1_4D33	S2B_MSIL1C_20170708T153109_N0205_R025_T18NVK
06/09/2017	S1B_IW_GRDH_1SDV_20170906T104214_20170906T104239_007270_00CD25_4F8F	S2B_MSIL1C_20170906T153109_N0205_R025_T18NVK
04/01/2018	S1B_IW_GRDH_1SDV_20180104T104213_20180104T104238_009020_0101D3_78E7	S2B_MSIL1C_20180104T152629_N0206_R025_T18NVK
31/10/2018	S1B_IW_GRDH_1SDV_20181031T104222_20181031T104247_013395_018C7A_5FB4	S2B_MSIL1C_20181031T152639_N0206_R025_T18NVK
30/12/2018	S1B_IW_GRDH_1SDV_20181230T104220_20181230T104245_014270_01A8AF_F455	S2B_MSIL1C_20181230T152639_N0207_R025_T18NVK
04/01/2019	S1B_IW_GRDH_1SDV_20190104T231309_20190104T231334_014351_01AB40_F03F	S2A_MSIL1C_20190104T152631_N0207_R025_T18NVK
05/03/2019	S1B_IW_GRDH_1SDV_20190305T231308_20190305T231333_015226)01C7C5_91ED	S2A_MSIL1C_20190305T152631_N0207_R025_T18NVK
27/08/2019	S1B_IW_GRDH_1SDV_20190827T104226_20190827T104251_017770_021713_CE93	S2B_MSIL1C_20190827T152649_N0208_R025_T18NVK