Advances in Space Research

Retrieval of Crop Biophysical and Biochemical Variables from AVIRIS-NG Hyperspectral Airborne data using Spectral Band Selection and Canonical Correlation Forests Regression in Diverse Agricultural Systems

--Manuscript Draft--

Manuscript Number:	
Article Type:	SI: AVIRIS-NG
Keywords:	Hyperspectral, Crop retrieval, PROSAIL, AVIRIS-NG, Biochemical, Biophysical, Spectral Band Selection, Machine Learning Regression, Canonical Correlation Forests
Corresponding Author:	jayachandra ravi, PhD (pursuing) INDIA
First Author:	jayachandra ravi, PhD (pursuing)
Order of Authors:	jayachandra ravi, PhD (pursuing)
	Rahul Nigam
	Bimal K Bhattacharya
	Devansh Desai
	Parul R Patel
Abstract:	Sustainable agricultural management reduces over-utilization of farm resources and reduces risk of negative impacts on environment. Monitoring crop growth and status under various conditions at various spatio-temporal resolutions is a key to assess yield stability, crop diversity, adaptability and response. The quantitative assessment of of crop biophysical and biochemical variables from remotely sensed data with help of spectroscopic methods provide discerning information regarding canopy foliar condition and eco-physiological processes. AVIRIS-NG airborne data offers high spatial and spectral resolution giving an unique advantage to retrieve such crop biophysical and biochemical (BP-BC) variables. The retrieval of crop variables from leaf-canopy radiative transfer models such as PROSAIL is a powerful method to derive the crop biophysical and biochemical variables and can be complemented with use of nonlinear non-parametric methods which can offer simplicity, fastness, reliability and competency. The hyperspectral band selection is a cost-effective method to overcome data redundancy in high dimensional hyperspectral data. The most sensitive bands specific to vegetation properties (including canopy effects) were determined based on using Gaussian Processes Regression for BP-BC variables from spectral signatures collected from diverse agricultural systems of India - Raichur (Karnataka) and Anand (Gujarat) districts representing diverse landscapes and heterogeneous crop canopies. The study utilized spectral band selection method as technique to overcome band redundancy from field hyperspectral dataset. A decision tree ensemble Canonical Correlation Forests (CCF) is capable to naturally represent data with correlated inputs and suitable for retrieval using chosen subset of bands of a BP-BC varaible. The retrieval of various BP-BC variables from AVIRIS-NG airborne dataset was performed using hybrid inversion of PROSAIL-D model by CCF regression. Validation of retrieval was done using in-situ ground observations collected over heterogeneous crop landscape and gave high correlation for most variables with respect to in situ observations: chlorophyll (R^2 =0.83), equivalent water thickness (R^2 =0.81), leaf area index (R^2 =0.76) and dry matter (R^2 =0.73). Study showed limitation of inversion based on radiative transfer model in retrieval in case of carotenoid (R^2 =0.46), anthocyanin (R^2 =0.525).
Suggested Reviewers:	K R Manjunath, PhD Scientist, SAC: Space Applications Centre krmanjunath@mailcity.com An expert in the field of crop remote sensing including hyperspectral sensors
	Nidamanuri Rama Rao, PhD Professor, IIST: Indian Institute of Space Science and Technology

	rao@iist.ac.in An expert in retrievals and mapping using hyperspectral remote sensing
	R N Sahoo, PhD Principal Scientist, IARI: Indian Agricultural Research Institute rnsahoo@iari.res.in A senior science in the area of hyperspectral remote sensing
	Melba Crawford, PhD Professor, Purdue University mcrawford@purdue.edu She has done extensive research in statistical pattern recognition for high dimensional data analysis, data fusion techniques for multisensor problems, multiresolution methods in image analysis, and knowledge transfer in data mining.

Retrieval of Crop Biophysical and Biochemical Variables in Diverse Agricultural System from AVIRIS-NG Hyperspectral Airborne data through Hybrid Inversion of PROSAIL using Canonical Correlation Forests

Jayachandra Ravi[a], Rahul Nigam[b], Bimal. K. Bhattacharya[c], Devansh Desai[d], Parul Patel[e]

[a]*Civil Engineering Department, Nirma University, Ahmedabad, India (email: jayachandra.ravi50@gmail.com)*

[b]*Agriculture and Land Eco-system Division, Biological and Planetary Sciences and Applications Group, Earth, Ocean, Atmosphere Planetary Sciences and Applications Area, Space Applications Centre (ISRO), Ahmedabad, India. (email: rahulnigam@sac.isro.gov.in)*

[c]*Agriculture and Land Eco-system Division, Biological and Planetary Sciences and Applications Group, Earth, Ocean, Atmosphere Planetary Sciences and Applications Area, Space Applications Centre (ISRO), Ahmedabad, India. (email: bkbhattacharya@sac.isro.gov.in)*

[d]*Institute of sciences, Silver oak University, Ahmedabad (email: ddesai10793@gmail.com)*

[e]*Civil Engineering Department, Nirma University, Ahmedabad, India (email: parul.patel@nirmauni.ac.in)*

*Corresponding author e mail: jayachandra.ravi50@gmail.com

Declarations Authors' contributions: All authors contributed to the study conception and design. Material preparation, data collection was performed by Rahul Nigam, Devansh Desai. The analysis and first draft of the manuscript was written by Jayachandra Ravi and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding: The research leading to these results received funding from the Space Applications Center, ISRO, Ahmedabad and Nirma University, Ahmedabad.

Conflicts of interests/Competing interests: The authors have no conflicts of interest to declare that are relevant to the content of this article.

Retrieval of Crop Biophysical and Biochemical Variables from AVIRIS-NG Hyperspectral Airborne data using Spectral Band Selection and Canonical Correlation Forests Regression in Diverse Agricultural Systems

Jayachandra Ravi[a], Rahul Nigam[b], Bimal K. Bhattacharya[b], Devansh Desai[c], Parul R. Patel[a]

[a]*Civil Engineering Department, Nirma University, Ahmedabad, India*

[b]*Agriculture and Land Eco-system Division, Biological and Planetary Sciences and Applications Group, Earth, Ocean, Atmosphere Planetary Sciences and Applications Area, Space Applications Centre (ISRO), Ahmedabad, India.*

[c]*Institute of sciences, Silver oak University, Ahmedabad*

**Corresponding author e mail: jayachandra.ravi50@gmail.com*

Abstract: Sustainable agricultural management reduces over-utilization of farm resources and reduces risk of negative impacts on environment. Monitoring crop growth and status under various conditions at various spatio-temporal resolutions is a key to assess yield stability, crop diversity, adaptability and response. The quantitative assessment of of crop biophysical and biochemical variables from remotely sensed data with help of spectroscopic methods provide discerning information regarding canopy foliar condition and eco-physiological processes. AVIRIS-NG airborne data offers high spatial and spectral resolution giving an unique advantage to retrieve such crop biophysical and biochemical (BP-BC) variables. The retrieval of crop variables from leaf-canopy radiative transfer models such as PROSAIL is a powerful method to derive the crop biophysical and biochemical variables and can be complemented with use of nonlinear non-parametric methods which can offer simplicity, fastness, reliability and competency. The hyperspectral band selection is a cost-effective method to overcome data redundancy in high dimensional hyperspectral data. The most sensitive bands specific to vegetation properties (including canopy effects) were determined based on using Gaussian Processes Regression for BP-BC variables from spectral signatures collected from diverse agricultural systems of India - Raichur (Karnataka) and Anand (Gujarat) districts of India representing diverse landscapes and heterogeneous crop canopies. The study utilized spectral band selection method as technique to overcome band redundancy from field hyperspectral dataset. A decision tree ensemble Canonical Correlation Forests (CCF) is capable to naturally represent data with correlated inputs and suitable for retrieval using chosen subset of bands of a BP-BC variable. The retrieval of various BP-BC variables from AVIRIS-NG airborne dataset was performed using hybrid inversion of PROSAIL-D model by CCF regression. Validation of retrieval was done using in-situ ground observations collected over heterogeneous crop landscape and gave high correlation for most variables with respect to in situ

observations: chlorophyll (R^2=0.83), equivalent water thickness (R^2=0.81), leaf area index (R^2=0.76) and dry matter (R^2=0.73). Study showed limitation of inversion based on radiative transfer model in retrieval in case of carotenoid (R^2=0.46), anthocyanin (R^2=0.525).

Keywords: Hyperspectral, Crop, PROSAIL, AVIRIS-NG, Biochemical, Biophysical, Spectral Band Selection, Machine Learning Regression, Canonical Correlation Forests

1. Introduction

The assessment of dynamic response of crop to fluctuating weather and management strategies need regular quantitative extraction of biophysical and biochemical (BP-BC) variables (Chloupek, Hrstkova and Schweigert, 2004; Weiss, Jacob and Duveiller, 2020) at farm to landscape scale at various temporal resolution. Remotely sensed spectral imaging can provide valuable insight of crop variables for applications in monitoring, detection or estimation depending on spatial resolution for macro and micromanagement for farm operations. Detection of pigment content and composition from remotely sensed data with help of spectroscopic methods provide judicious information regarding crop condition and able to address response of eco-physiological processes on foliage (Ustin *et al.*, 2009). Remote Sensing (RS) based imaging spectroscopy can capture spectral responses of plant functional traits and is based on link between spectral and functional signatures. The precise estimation of crop BP-BC variables is a vital input for effective application of remote sensing for precision agriculture to achieve the goal of sustainable growth (Tejada et al., 2003; Liaghat 2010). Their determination from an over-determined spectral signal is a challenge that could be solved using pattern detection and Machine Learning (ML) techniques (Schweiger *et al.*, 2016). These variables can be obtained non-destructively on a spatio-temporal scale using advanced remote sensors with high spectral resolution in optical and Infrared (IR) spectral regions. There is significant scope and potential of Hyperspectral (HS) RS in many applications including crop monitoring (Goel, et al. 2003) but which is proved from growing number of scientific publications on hyperspectral RS over years that use Machine Learning (ML) (Gewali, Monteiro and Saber 2018). Due to better spatial as well as spectral resolution along with flexibility of operation, Hyperspectral Airborne Remote Sensing (HARS) is better at time-crucial and time-specific precision, larger band selection than satellite–based systems (Goel *et al.*, 2003; Koponen *et al.*, 2007). The hyperspectral airborne sensors have potential to offer better relationships between remote sensing data and crop parameters (Goel *et al.*, 2003). Moreover, inter and intra pixel variation is resulted from more finer patches with higher Confidence Levels (CL) by HARS compared to smoothing out of small patches by satellite sensors of low spatial or spectral resolution (Mumby *et al.*, 1997). Nevertheless, satellite sensors provide a synoptic view with a wide swath and annual repeat coverage making it a more cost effective system than an airborne platform (Holmgren and Thuresson,

1998; Lucas *et al.*, 2007; Jha, Levy and Gao, 2008). But these airborne hyperspectral data provide insight to select appropriate spectral bands and their specifications for satellite configuration for crop variable retrieval.

Select hyperspectral bands carry higher sensitivity to a small variation of green and non-green pigments and overall leaf composition (Gitelson, Gritz † and Merzlyak, 2003). The information rich hyperspectral data comes with challenges such as information redundancy removal, optimal information identification (Bajcsy and Groves, 2004). Spectral band selection depends a lot on the input data properties and criterion defined to discriminate 'best' optimal feature subset of bands (Pal, 2006). Among methods of feature selection, there is no one single "best feature selection method" as it depends on different metrics like computational cost, time, accuracy, ratio of feature selection as well as various aspects of dataset like multi-class outputs, noise in data, band redundancy, irrelevant features (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2013). Identification of most informative spectral bands in Hyperspectral data relied on statistical methods like Partial Least Squares in conjunction with techniques like correlation coefficient analysis, elimination of non-informative variables, stepwise regression variable selection, exhaustive band combination (Centner *et al.*, 1996; Fung, Yan Ma and Siu, 2003; Cai, Li and Shao, 2008; Li *et al.*, 2014; Chen *et al.*, 2015; Kira *et al.*, 2016; Jin and Wang, 2019). Most works in band selection associated to Hyperspectral RS are related to classification problems, while relatively fewer works were meant for regression problems considering vegetation characteristic properties (Abdel-Rahman, Ahmed and Ismail, 2013; Verrelst, Malenovský, *et al.*, 2019). Among six types of Hyperspectral band (feature) selection methods for classification problems: Ranking-based, Searching-based, Clustering-based, Sparsity-based, Embedding Learning-based, Hybrid scheme-based, the Sparsity-based methods were found best performing in terms of accuracy, but ranking-based methods were more suitable for large hyperspectral datasets due to low-complexity (Sun and Du, 2019). Feature selection methods are also classified on the basis of search organization, sub setting and evaluation methods as: filter, wrapper and embedded methods (Khalid, Khalil and Nasreen, 2014). It has been shown that filter methods are faster and better suited to high dimensional datasets. The estimation of number of bands for band selection is a challenge as well because too-less number of bands will not allow enough spectral information to be preserved within selection while too-large number of bands causes band redundancy. The retrieval of BP-BC variables is most likely a non-linear inverse function of given observations (spectral) received by the remote sensing sensor. Most works have shown special interest in retrievals of C_{ab} and LAI owing to their easier techniques for field data collection (Berger *et al.*, 2018). Inversion model applied on RS images for estimation of vegetation properties assumes an inversion function to be nonlinear, smooth, and continuous. A hybrid inversion scheme achieves inversion of Canopy Radiative Transfer (CRT) model with use of input-output data of model simulations to train statistical regression models (Camps-Valls *et al.*, 2020). Various Machine

Learning Regression Algorithms (MLRAs) were made part of Automatic Radiative Transfer Models Operator (ARTMO) GUI that uses various Leaf-Canopy Radiative Transfer Models (Verrelst *et al.*, 2011). A Gaussian Processes Regression band analysis GPR-BAT (Verrelst *et al.*, 2016a) which works on basis on Sequential search strategy and backward elimination of bands for subset generation and starts with a full set of bands and subset is evaluated on information criterion, whereas stopping criteria is that of ranking and elimination of least contributing band is done sequentially.

Unlike classification problems, a regression estimation involves continuous values as outcome variables (Poldrack, Huckins and Varoquaux, 2020). Unlike parametric regression models in which model form is specified a priori, the nonparametric regression is essentially a data-driven method that is determined from dataset (Verrelst *et al.*, 2015; Mahmoud, 2021). Nonparametric regression models have emerged as a suitable interface that links the efficacy of standard statistical techniques with the detail and complexity of physically based approaches (Houborg and McCabe, 2018). Examples of nonlinear nonparametric regression methods include ML algorithms like Decision Trees, Neural Networks, Kernel based regression methods. In present work, a decision tree ensemble named Canonical Correlation Forests (CCF) is used for retrieval and tested for performance. CCFs consist of numerous binary decision trees and sequentially best splits in projection plane are chosen after application of Canonical Correlation Analysis (CCA) at each decision tree node while training in order to find best projection feature for correlation (Rainforth and Wood, 2015).

A CRT model simulates the scattering and absorption of radiation inside leaf canopies and provide an intrinsic connection between the plant biophysical characteristics and canopy reflectance. PROSAIL is a one-dimensional homogenous canopy radiative transfer model that couples two models, namely:1) PROSPECT (Jacquemoud and Baret, 1990) for simulation of leaf optical properties w.r.t leaf constituent composition; 2) Scattering by Arbitrary Inclined Leaves (SAIL) for computation of Top of Canopy (TOC) reflectance w.r.t canopy geometry and leaf distribution (Verhoef, 1984). PROSPECT model is based on plate model and computes reflectance and transmittance in spectral range of 400 nm to 2500 nm (ALLEN WA *et al.*, 1969). SAIL is based on four-stream RT modeling which involves two direct fluxes (incident solar flux and radiance in the viewing direction) and two diffuse fluxes (upward and downward hemispherical flux) (Verhoef *et al.*, 2007). The Global Sensitivity Analysis (GSA) of RT model PROSAIL using Sobol's total sensitivity indices uses a modified version of variance based sensitivity analysis of model output. identifies the sensitive spectral regions variance-based methods, which decomposes the variance of the model output into fractions that can be attributed to inputs or sets of inputs (Verrelst *et al.*, 2015; Verrelst, Rivera and Moreno, 2015). The decision tree ensembles such as Random Forest trained with PROSAIL simulations relating crop phenology using hyperspectral data are capable of better

predictions of LAI, Chlorophyll (Doktor *et al.*, 2014). Canonical Correlation Analysis (CCA) performs better in predicting LAI than Partial Least Squares Regression (PLSR) as it achieves maximum correlation between the spectral variables and LAI (Pu, 2012). The decision tree ensembles tend to perform one of the best among non-parametric methods (Verrelst *et al.*, 2015).

The objective of present work two pronged: (i) To test the efficacy of band selection method GPR-BAT which uses Sequential Backward Band Removal (SBBR) as a hyperspectral band selection method over field hyperspectral dataset. (ii) To evaluate the accuracy of PROSAIL-D CCF hybrid regression retrieval of BP-BC variables from AVIRIS-NG airborne dataset. It is to be noted that validation work of retrieval of various parameters in case of airborne RS within the time of airborne flights is extremely challenging. In this experiment, validation work involved measurements collected from various crop canopies in a number of sites within study area during time of flight and posed risk of diurnal variations of spectral measurements (Zarco-Tejada, Catalina, *et al.*, 2013).

2. Materials

2.1 Experimental sites

This study was conducted in two different agricultural belts nearly a month apart from each other. The two experimental sites are: Site (A) Anand, the central region of western state of Gujarat, India; Site (B) Raichur, the eastern region of Southern state of Karnataka, India. The former is located in Gujarat Plain & Hill (GPH) agro-climatic zone of India whereas the latter is located in Southern Plateau & Hill (SPH) zone in Krishna-Tungabhadra doab. Anand (22.5645° N, 72.9289° E) located in central part of Gujarat state has an average temperature of 27.2 °C and annual rainfall of 882 mm with predominantly sandy loamy soil. Raichur (16.2160° N, 77.3566° E) located in middle-eastern part of Karnataka state has an average temperature of 30°C and annual rainfall of 713 mm. Raichur had a black clayey soil with high fertility and moisture retention capacity but develop deep cracks on drying. The agrometeorological, soil and cropping properties are listed out in table 1. They include different crops within summer growing seasons in different growth stages. Crops and their approximate periods of sowing to harvest in as follows : Site (A) Anand: wheat (November to March), tobacco (November to March), Vegetables (sown in February or March), Fodder (sown in February or March), Sesame (February to June), Paddy (March to May), Pearl Millet (March to May), Maize (March to May); Site(B) Raichur: Sorghum (November to February), Pearl millet (February to May), Paddy (January to May), Groundnut (January to May), Bengalgram (February to May), Greengram (February to May), Redgram (February to May). At the time of airborne campaign on dates - 24th February (Site B - Raichur), 26th March 2018 (Site A - Anand) comprised of various crops cultivated plot level which allowed spatial heterogeneity.

<Insert Figure 1 and Table 1 here>

2.2 In-situ data

To assess airborne hyperspectral retrieval of quantitative estimates of BP-BC variables from the images, ground truth campaign was designed to have good spatial coverage of airborne hyperspectral pixels. The sampling measurements were taken in experimental plots using random stratified sampling strategy and with precise geographical coordinates to avoid location uncertainty. The spatial heterogeneity with well-marked sampling sites is adopted for field sampling done on a larger area compared to pixel size (Deguise *et al.*, 2015).

2.2.1 Ground-level Optical Measurements

The ground truth data for validation involved point hyperspectral reflectance signatures of crops was measured at the time of the flights using spectroradiometer (ASD FieldSpec Pro, Analytical Spectral Devices, Boulder, CO, USA) with a spectral range of 400-2500 nm. The instrument has spectral sampling of 1.4 nm in VNIR and 2 nm in the SWIR automatically interpolated to 1 nm. The Full Width Half Maximum (FWHM) are 3.5, 9.5 and 6.5 at 700, 1400 and 2100 nm. The instrument was attached to standard fore-optic with 25° field of view (FOV) through a permanent fibre optic cable.

<Insert Figure 2 here>

2.2.2 Measurement of Biophysical and Biochemical (BP-BC) variables

Non-destructive measurement of LAI and Chlorophyll in the plots using LAI-2000 (Li-COR Biosciences, Lincoln, NE, USA) and SPAD-502 (Konica Minolta Optics, Inc., Tokyo) respectively. The LAI measurements were taken in at least 4 to 5 subplots of size 1 m^2 (1 m × 1 m) within every plot and averaged (Atzberger *et al.*, 2013). The leaf total chlorophyll *a* and *b* content (C_{ab}, $\mu g.cm^{-2}$) were measured from leafs belonging top canopies of plants in subplots were averaged and converted into actual C_{ab} using empirical calibration functions (Atzberger *et al.*, 2013; Li *et al.*, 2020). The destructive plant sampling involved quantity estimates of leaf biochemical composition under laboratory controlled conditions for anthocyanin, carotenoid, dry matter and water contents. At leaf level, the pigment mass per leaf dry mass (g/g or %) is the ratio of pigment mass per leaf area ($g.cm^2$) and Leaf Dry Mass per Area (g/cm^2). At canopy level, since pigment content per canopy surface area ($\mu g.cm^{-2}$) is calculated as the product of pigment mass per leaf area ($\mu g.cm^2$) and LAI, the pigment content per canopy surface area ($\mu g.cm^{-2}$) is calculated as product of pigment mass per leaf dry mass (g/g or %), LAI and Leaf Dry Mass per Area ($\mu g/cm^2$) (Kattenborn *et al.*, 2019). To extract and measure pigments carotenoid (C_c), Chlorophyll a & b (Chl_a & Chl_b), their measurements are calculated using the extinction coefficients (Wellburn, 1994) and absorbance's A_{470},

A_{646}, A_{663} measured at wavelengths of 470 nm, 646 nm, and 663 nm using spectrophotometer as per procedure (Huang *et al.*, 2018) (Zarco-Tejada, Guillén-Climent, *et al.*, 2013) through following equations.

$$Chl_{a-conc}\ (mg/L) = 12.21 * A_{663} - 2.81 * A_{646} \quad (1)$$

$$Chl_{b-conc}\ (mg/L) = 20.13 * A_{646} - 5.03 * A_{663} \quad (2)$$

$$C_{ab-mass}\ (g/g) = [Chl_{ab-conc}\ (g/L) * VT\ (ml)]\ /\ [Leaf\ dry\ mass(g) * 1000] \quad (3)$$

where $C_{ab-conc}\ (g/g) = Chl_{a-conc}\ (mg/L) + Chl_{b-conc}\ (mg/L)$

$$C_{ab\text{-}area}\ (g.cm^{-2}) = C_{ab\text{-}mass}\ (g/g) * Leaf\ dry\ mass\ per\ unit\ leaf\ area\ (g/cm^2) \quad (4)$$

$$C_{c\ conc}\ (mg/L) = [1000 * A_{470} - 3.27 * Chl_a - 104 * Chl_b]/229 \quad (5)$$

$$C_{c-mass}\ (g/g) = [\ C_{c\ conc}\ (g/L) * VT\ (ml)]\ /\ [Leaf\ dry\ mass(g) * 1000] \quad (6)$$

$$C_{c\text{-}area}\ (g.cm^{-2}) = C_{c\text{-}mass}\ (g/g) * Leaf\ dry\ mass\ per\ unit\ leaf\ area\ (g/cm^2) \quad (7)$$

where A_x is the absorbance of the extract solution at wavelength x, VT (ml) is the volume of leaf pigment extract solution. $Chl_{ab\text{-}conc}$ and $C_{c\text{-}conc}$ is the concentration of Chlorophyll-ab and carotenoid per unit volume of solvent (water) respectively. $C_{ab\text{-}mass}$ and $C_{c\text{-}mass}$ is Chlorophyll-ab and Carotenoid mass per leaf dry mass respectively. $C_{ab\text{-}area}$ and $C_{c\text{-}area}$ is Chlorophyll-ab and Carotenoid mass per leaf area respectively. The procedure for quantification of Anthocyanin in leaf extracts is according to (Gitelson *et al.*, 2017; Gitelson and Solovchenko, 2018; Falcioni *et al.*, 2020) where absorbance measured at 530 nm absorption coefficient of 30 $mM^{-1}\ cm^{-1}$. The dry matter and water content are calculated as shown in equation 8 to 9 (Yilmaz, Hunt and Jackson, 2008).

$$C_m\ (g.cm^{-2}) = [Leaf\ dry\ mass\ (g)]/\ [leaf\ area\ (cm^2)] \quad (8)$$

$$C_{vwc}\ (kg.m^{-2}) = \eta * [Leaf\ fresh\ mass\ (g) - Leaf\ dry\ mass(g)] + [Stem\ fresh\ mass\ (g) - Stem\ dry\ mass(g)] \quad (9)$$

$$C_w\ (cm^{-1}) = [Leaf\ fresh\ mass\ (g) - Leaf\ dry\ mass(g)]/\ dw * leaf\ area\ (cm^2)] \quad (11)$$

where C_m is foliar dry matter content obtained from dry weight of leaves when oven heated at 800 °C for 2 days. C_{vwc} is vegetation water content obtained from fresh and dry mass of the leaves and stem. C_w is Equivalent Water Thickness. η is the plant density (number of plants) in square unit area. dw is the density of liquid water (0.001 $g.cm^{-3}$). The observed Canopy Water Content (CWC) expressed as weight of water per unit area of leaf surface area is obtained from Water Absorption Area Index (WAAI) calculated from spectral reflectance collected in-situ (Pasqualotto *et al.*, 2018).

$$WAAI_{optimized} = 180(1.812\ R_{911} + 0.271) - \int_{911}^{1271} R(\lambda) d\lambda \quad (12)$$

$$CWC(g/m^2) = 42.98\ exp^{0.061 WAAI_{optimized}} \quad (13)$$

3. Methods

The workflow of methodology is depicted in Fig. 3.

<Insert Figure 3 here>

3.1 Airborne Hyperspectral campaign - Image acquisition and preprocessing

The hyperspectral imaging campaign was organized in local time around 11 am on 26th March 2018 (Site A - Anand) and 24th February (Site B - Raichur), using National Aeronautics and Space Administration (NASA)'s Next Generation Airborne Visible Infrared Imaging Spectrometer (AVIRIS-NG) instrument which is a pushbroom scanner based mapping system that was flown on Indian Space Research Organization (ISRO)'s B200 aircraft platform (Bhattacharya *et al.*, 2019). The aircraft was flown at an altitude of nearly 4 km under favorable weather conditions of predominantly clear sky with minimal cloud coverage and no precipitation during any acquisitions. The instrument covers a spectral range of 380-2510 nm with a single Focal Plane Array (FPA), high signal-to-noise ratio (SNR) (>2000 @ 600 nm and >1000 @ 2200 nm) at high spectral sampling interval of 5nm (Chapman *et al.*, 2019). The sensor had 34° field-of-view (FOV) and 1 milliradian instantaneous field of view (IFOV) allowing a ground sampling distance of approximately 4 meters. The photodetectors formed a 640 × 480 pixel array with the teledyne imaging sensor having 640 cross-track spatial samples (perpendicular to the flight direction) and 425 spectral samples. AVIRIS-NG data uncertainties that emanate from optical (radiometric and spectral measurement errors) and electronic imperfections in the instrument can be addressed by steps of calibration as described in (Chapman *et al.*, 2019). In geometric correction, the precalibrated boresight coefficients, GNSS position data, IMU angular information, and surface digital elevation are utilized to orthorectify hyperspectral imagery (Wang *et al.*, 2021). The removal of high noise bands due to certain imperfections in calibration, water-vapor and carbon-dioxide effects, overlapping of wavelengths, high-frequency band noise, bad data quality caused number of bands to reduce to 372 radiometrically calibrated bands (Malhi *et al.*, 2020). The radiometric corrections allow processing of level-0 (L0) digital number (DN) values to level-1 (L1) radiance product. The resulting noise eliminated high-resolution hyperspectral image is capable of identification of pure vegetation pixels, extracting the pure canopy radiance and reflectance (Quemada, Gabriel and Zarco-Tejada, 2014). The atmospheric correction converts the measured radiances to level-2 (L2) surface reflectance product (Bhattacharya *et al.*, 2019). Table (2) shows the data description and characteristics used for the study. The level-2 surface reflectance image is available for download from the VEDAS geo-portal (https://vedas.sac.gov.in/aviris) of the Space Applications Centre (SAC) , ISRO and which has been used for retrieval of BP-BC variables by identifying most sensitive bands using band selection (BS) algorithm.

<Insert Table 2 here>

3.2 Sensitive band selection based on GPR based on recursive backward feature elimination

The Gaussian Processes regression has been used for band selection (Verrelst *et al.*, 2016b). The Gaussian processes can directly define and infer a distribution over latent functions a priori and be converted into a posterior over functions based on the observed functional values.

For regression, assuming an equation is in form of $Y_i = \mathrm{x}^{\mathrm{T}}\mathrm{w} + \varepsilon_i$ where y is a vector of response variable (variable to be predicted), x is vector of reflectance bands (explanatory variable) ranging from $x_1, \dots, x_n, x_*$ and w is the latent function

Assuming noise ε_i follow a zero-mean Gaussian distribution

$$\varepsilon_i \sim N_d(0,\sigma_n^2) \tag{17}$$

applying the likelihood, the probability density, when w is known and factored over input vector X, is

$\mathrm{p}(\mathrm{y}|\mathrm{X},\mathrm{w}) = N_d\left(\mathrm{x}^{\mathrm{T}}\mathrm{w}, \sigma_n^2 I\right)$ where σ_n^2 is the noise variance which is assumed to be independent

$$\mathrm{w} \sim N_d(0,\Sigma) \text{ where } \Sigma = \mathrm{k}(\mathrm{x_i},\mathrm{x_j}) \text{ is covariance function} \tag{18}$$

For the test data points (X_*,Y_*) with n observations, the prior joint distribution of $\mathrm{Y_i}$ and Y_* $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ y_* \end{bmatrix} \sim N_d(0,\Sigma)$

under the priors Eq. (17) and (18) where Σ is covariance matrix

For input given response variable $f(\mathrm{x})$, the posterior distribution for f_* is given by

$$f_*|\,(Y_1 = y_1, \dots, Y_n = y_n, x_1, \dots, x_n, x_t) \sim N_d(K_*^T K^{-1} y, K_{**} - K_*^T K^{-1} K_*) \tag{19}$$

where $f_* = f(x_*)$ is the prediction from test input data; K, K_*, K_{**}, K are training, training-testing, testing-training and testing kernel matrices respectively in covariance matrix Σ which can be decomposed as $\begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}$ and are functions of $x_1, \dots, x_n, x_*$. Covariance matrix for a zero-mean, noise independent Gaussian converts to $\hat{\Sigma} = \Sigma + \sigma^2 I$ for equation where $\varepsilon_i \sim N_d(0,\sigma^2)$. Substituting updated covariance matrix in Eq. (19), we get posterior distribution for Y_* as

$$Y_*|\,(Y_1 = y_1, \dots, Y_n = y_n, x_1, \dots, x_n) \sim N_d(K_*^T (K + \sigma^2 I)^{-1} y, K_{**} - K_*^T (K + \sigma^2 I)^{-1} K_*) \tag{20}$$

where the term $K_*^T (K + \sigma^2 I)^{-1} y$ is mean (i.e., best predicted value) and $K_{**} - K_*^T (K + \sigma^2 I)^{-1} K_*$ is the covariance (i.e., confidence measure) for the prediction .

From above the important role of covariance function is understood. The most popular choice for covariance function is a kernel function of a smooth form of squared exponential represented as

$$k(\mathrm{x_i}, \mathrm{x_j}) = \theta_1^2 \exp\left(-\frac{(\mathrm{x_i} - \mathrm{x_j})^2}{2\theta_2^2}\right) \quad (21)$$

Accounting for noise, the Eq. (21) changes to $\mathrm{k}(\mathrm{x_i}, \mathrm{x_j}) = \theta_1^2 \exp\left(-\frac{(\mathrm{x_i}-\mathrm{x_j})^2}{2\theta_2^2}\right) + \sigma_n^2 \delta_{ij}$ (22)

where $\theta_1^2$is the variance of the correlated noise component, θ_2 is its characteristic length-scale and σ_n^2 is the variance of the independent noise component. δ_{ij} is a Kronecker delta which equals unity if i = j and equals zero otherwise. The two hyper parameters θ_1 and θ_2 control function.. To train the GPR, using marginal likelihood, i.e., integral of likelihood times the prior of $y \sim N_d(0, K + \sigma^2 I)$ distribution with marginalization of values of w.

$$\log p(y|X) = \int p(y|w, X)\, p(w|X) d\, w \quad (23)$$

$$\log p\big(y|(Y_1 = y_1, \ldots, Y_n = y_n, x_1, \ldots, x_n)\big) \quad (24)$$

$$= -\frac{1}{2} y_*^T (K + \sigma^2 I)^{-1} y \; - \frac{1}{2} \log |\, K + \sigma^2 I| - \frac{n}{2}\, log 2\pi$$

Optimal values for hyper parameters and noise parameters can be estimated by maximizing the log marginal likelihood. Through the determined optimal hyper parameters, the mean and covariance of prediction in Eq. (20) can be calculated (Li and Huang, 2021).

The general procedure for feature selection involves four key steps: subset Generation, evaluation of subset, stopping criteria, result validation (Kumar, 2014). The band selection of variables in most cases if not all exist in the sensitive regions which is known from GSA of PROSAIL (Verrelst *et al.*, 2016c). A technique based on sequential search strategy and backward elimination of bands for regression is employed for subset generation. This kind of band selection starts with a full set of bands and subset is evaluated on information criterion, whereas stopping criteria is that of ranking and elimination of least contributing band is done sequentially. The Sequential Backward Band Removal (SBBR) algorithm identifies least contributing band is removed in every iteration after ten-fold Cross Validation (CV). The error is calculated for each such combination. The ten-fold CV is adopted to overcome and differentiate any low differences in sigma bands (σ_b) in any certain spectral region. Although, the best combination set of bands are chosen from this method with least Normalized RMSE, one still has to be mindful of high band correlation in outputs which might affect regression output. The most sensitive bands for C_{ab}, C_c, C_a, C_w, C_m, LAI were used for model inversion as observation or explanatory variables for training. In this process, water absorption bands in 1340–1480nm and 1770–1970 nm range are removed. Ground collected spectra showed relatively high amount of signal noise above 2200 nm which was also removed with exception unless deemed necessary as sensitive region for variable retrieval in some cases. The spectra were used as input for band selection of spectral subsets variable-wise as per their respective sensitive wavelength regions from PROSAIL-D GSA

(Verrelst and Rivera, 2016; Verrelst *et al.*, 2016c; Verrelst, Vicent, *et al.*, 2019). The areas where sensitive bands are found for each of variables are: C_{ab} 457-780 nm, C_c 482-562 nm, C_a 522-622 nm, C_w 1093-2089 nm, C_m 692-2089 nm, LAI 401-922 nm and 1488-2089 nm. The spectral region 1333-1483 nm, 1704-1979 nm is excluded as those bands were removed in correction of atmospheric attenuation due to gaseous absorptions in AVIRIS-NG image, the spectral region 2094-2500 nm is excluded due to low SNR in that range of ground collected spectra by spectroradiometer.

3.3 Simulation of PROSAIL radiative transfer model in forward mode

The PROSPECT-D + SAIL, which is henceforth referred to as PROSAIL-D couples PROSPECT-D (Féret *et al.*, 2017) which is leaf RT model describing directional-hemispherical reflectance and transmittance with 4SAIL canopy model offering an advantage of simplicity, robustness and extensive validation (Jacquemoud *et al.*, 2009; Zhang *et al.*, 2018).

$$\rho_l(\lambda) = R_{N,a} = \rho_\alpha + \frac{\tau_\alpha \tau_{90} R_{N-1,90}}{1-\rho_{90} R_{N-1,90}} \text{ and } \tau_l(\lambda) = T_{N,\alpha} = \frac{\tau_\alpha T_{N-1.90}}{1-\rho_{90} R_{N-1,90}} \quad (14)$$

$$k(\lambda) = K_e(\lambda) + \sum_i \frac{C_i . K_i(\lambda)}{N} \quad (15)$$

The PROSPECT-D assumes internal leaf is assumed to be composed of N parallel homogeneous layers as plates and simulates reflectance and transmittance within a leaf as a function of the leaf mesophyll structure, biochemistry component weights C_i (includes leaf chlorophyll content (C_{ab}), leaf carotenoid content (C_c), leaf anthocyanin content (C_{ab}), leaf water content (C_w), leaf dry matter content (C_m)) characterized by an absorption coefficient (K_i) of biochemistry component (Zhang *et al.*, 2018) and dependent on the refractive index (n), incident angle (α) and transmission coefficient (θ) (Jacquemoud and Baret, 1990) which are shown in Eq. (14) and (15).The symbols ρ_{90} and τ_{90} denote the reflectance and transmittance of every internal layer, and $R_{N-1,90}$ and $T_{N-1,90}$ are the total reflectance and transmittance of the internal N-1 layers. k_e is an absorption coefficient in case of albino leaf or dry flat leaves.

SAIL model (Verhoef, 1984) assumes an infinite horizontally homogeneous vertical layers of canopy and calculates bi-directional reflectance considering the variables which describe specular sunlight and canopy geometry. It is based on four-flux theory that describes the interactions among four fluxes: direct solar flux, downward diffuse flux, upward diffuse flux, and flux in observer direction (Yang, Verhoef and van der Tol, 2020). It is a numerically robust and computationally efficient (Sobrino, Jiménez-Muñoz and Verhoef, 2005) and may be represented as function of various variables associated with canopy and sunlight geometry (Verhoef and Bach, 2003) which is coupled with $\rho_l(\lambda)$ and $\tau_l(\lambda)$ from PROSPECT-D (Zhang *et al.*, 2018) as shown in Eq. (16):

$$\rho_c = 4SAIL(LAI, ALA, \rho_l(\lambda), \tau_l(\lambda), P_{soil}(\lambda), SZA, VZA, \varphi, Hspot, SKYL) \quad (16)$$

where ρ_c is the canopy reflectance. The input parameter space for PROSAIL-D forward simulations for model inversion varied according to the ranges in Table 3 and covers all kinds of species like monocotyledonous, dicotyledonous and senescent leaves (Verhoef and Bach, 2003). The sun and view geometry angles are fixed.

In total 11000 BP-BC variable combinations were generated equal to N (k + 2) (where N = number of samples and k = number of variables) using Saltelli periodic function (Saltelli, Tarantola and Chan, 1999) by the uniform pseudo random sampling (Zhang *et al.*, 2018) of the nine BP-BC variables. The model input variables produced were entered into the PROSAIL-D model for processing simulated directional reflectances over the range of 400–2500 nm at a 5 nm (to AVIRIS-NG) interval in forward mode in MATLAB (The MathWorks, Inc.). The increase in bandwidth of simulations from 1nm to 5nm does not affect variability of spectral response much and hence negates necessity to apply sensor response function (Chen *et al.*, 2014; Cundill, der van Werff and der van Meijde, 2015). The correlation coefficient (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) between variable to be predicted and explanatory variables from simulated data is checked for C_{ab} to justify the selection of CCF for BP-BC retrieval and band selection approach against use of vegetation indices (VIs) as explanatory variables (Liang *et al.*, 2015; Liu, Shi and Gao, 2018).

3.4 Canonical Correlation Forests for BP-BC retrieval

CCFs just like other decision tree ensembles are effective due to their scalability and need diminutive parameter tuning, but the two important factors that determine their performance are the accuracy of individual trees and their prediction accuracies (Rainforth and Wood, 2015). CCF engages careful selection of hyperplane splits based on Canonical Correlation Analysis (CCA) computed at every node leading to local incorporation of correlation between features there by resulting in reduction of correlation between trees.

$$[\phi, \Omega] = \mathrm{CCA}(x', y') \quad (17)$$

where ϕ and Ω are canonical coefficients corresponding to x' and y' respectively

$$U = X_{(w_j \cdot s_j)}\phi \quad (18)$$

Where space U is split in the space of X given by the projection ϕ_j corresponding to one of the columns of ϕ, and the split point s_j in the projected space. w_j is the index of data points present at node j.

$$\{\phi_j, S_j\} = \underset{\Phi \in \phi, s \in R}{\arg\max}\ G\left(Y_{(w_j,:)}, \Phi, s\right) \text{ ; where } G\left(Y_{(w_j,:)}, \Phi, s\right) \text{ is gain of split} \quad (19)$$

$$= \underset{\Phi \in \phi, s \in R}{\arg\max} \left(g\left(Y_{(\omega_{j,:})}\right) - \frac{N_{x_{j,l}}}{N_j} g\left(Y_{\left(\omega_{x_{j,l},:}\right)}\right) - \frac{N_{x_{j,l}}}{N_r} g\left(Y_{\left(\omega_{x_{j,l},:}\right)}\right) \right)$$

The measure of impurity which is also mean squared error split criterion for CCF regression is

$$g\left(Y_{(w_j)}\right) = \frac{1}{N_j} \sum_{n \in w_j} Y_{(n)}^2 - \left(\frac{1}{N_j} Y_{(n)}\right)^2 \quad (20)$$

This is crucial because often decorrelation methods employed in other decision tree ensembles compromises accuracy as a trade-off. To justify the selection of CCF the measures of R^2, RMSE, MAE is used for testing prediction of CCF over 70% observations of sample 11000 PROSAIL simulations dataset compared with few other ML regression methods: Least Squares Linear Regression (LSLR), PLSR, Random Forests (RF), Regression Tree (RT), Neural Network (NN), Support Vector Regression (SVR), Kernel Ridge Regression (KRR), Gaussian Processes Regression (GPR).

3.4.1 Statistical analysis of retrieval

The validation of PROSAIL inversion maps using CCF and other ML methods were performed with independent ground measurements of BP-BC variables at two diverse agricultural sites of India. The coefficient of determination (R^2) and root mean square error (RMSE) are calculated and analysed for the predicted values against the measured values of BP-BC variables.

$$R^2 = \frac{[\sum_{i=1}^{n}(Y_{predicted} - Y_{mean\_predicted}) \cdot \sum_{i=1}^{n}(Y_{measured} - Y_{mean\_measured})]^2}{\sum_{i=1}^{n}(Y_{predicted} - Y_{mean\_predicted})^2 \sum_{i=1}^{n}(Y_{measured} - Y_{mean\_measured})^2} \quad (21)$$

A higher coefficient of determination (R^2) is an indicator of a better goodness of fit for the observations.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{n=N}(Y_{measured} - Y_{predicted})^2}{N}} \quad (22)$$

4. Results and Discussion

4.1 Results of GPR band selection and impact of band selection on CCF PROSAIL-D inversion

The present work used band selection based on wrapper method of recursive backward elimination aimed at removing least contributing bands. The results show that GPR being a data-driven is dependent on the field dataset presented to it in training phase and affects band selection. The rationale of using k-fold cross-validation is overcome the fluctuation depending on training-validation partitioning. Table 4 and Fig. 5 shows the statistics for the most optimum combination of bands to be used for retrieval of BP-BC variables. For C_{ab} band combination 457, 467, 487, 552, 707, 762 nm (NRMSE = 4.72) outperformed other

combinations and lesser standard deviation. For chlorophyll, the NRMSE stabilizes after 6 bands with accuracy in calibration and the corresponding validation was very close which included bands in red edge and green region. The band combinations for Carotenoids retrieval showed stable accuracies after 4 bands (NRMSE = 17.65) but the accuracies between the calibration and validation widened. Anthocyanin retrieval followed an uneven and non-uniform pattern of accuracies where accuracies converged intermittently, but the NRMSE was least at 10 bands which were mostly from green region: 517, 522, 537, 542, 552, 562, 592, 602, 617 nm with NRMSE of 15.91. The accuracy of Equivalent Water Thickness varied intermitted across all band combinations although with narrow difference between calibration and validation accuracies. It stabilized near 9 bands with wavelengths comprising mostly from near infrared but also shortwave infrared (SWIR) region: 1123, 1128, 1163, 1188, 1659, 1674, 1984, 1093, 1138 nm with NRMSE of 14.14. The Dry Matter Content is influenced by bands mostly from red edge, near infrared to SWIR regions of optical spectrum confirming the GSA of PROSAIL that the Dry matter content dominates output in almost all part of from red to SWIR and predominant band wavelengths being from SWIR. The accuracies did not stabilize despite for dry matter unlike other BP-BC variables, yet the combination comprising 21 bands : 697, 702, 707, 742, 977, 992, 1113, 1198, 1208, 1243, 1263, 1268, 1503, 1508, 1513, 1553, 1573, 1659, 1684, 1984, 2089 nm with NRMSE of 7.67 was chosen. This is done despite the band combination comprising 69 bands and 95 bands gave least NRMSE of 6.6 and 6.6 respectively considering the requirement to choose tolerable number of bands for feature selection rather than the optimal model and this requires decrease of number of input features. Similarly, LAI band selection establishes the findings of LAI contribution throughout optical spectrum (Verrelst *et al.*, 2016a). The NRMSE for LAI reduced at 22 bands with bands comprising all regions of optical spectrum though most bands are between green to red regions: 411, 441, 587, 577, 582, 592, 602, 607, 632, 642, 647, 652, 692, 712, 727, 882, 1644, 1689, 1984, 1989, 2004, 2084 nm with NRMSE of 6.96 respectively. Coefficient of determination as shown in Fig. 5, R^2_{cv} of Cab and LAI are 0.9647 and 0.9215 respectively is highest among all others whereas R^2_{cv} of C_c and C_a is 0.4716 and 0.6182 respectively which is a moderate correlation comparatively among other variables. R^2_{cv} of C_w, C_m is 0.7449, 0.8669 respectively. When the band selection was first tested on simulation data, it is seen that nonlinear non parametric predictors especially KRR, GPR, CCF exhibit nearly lesser RMSE and higher R^2 - MAE ratio showing computationally superior performance when tested on simulated data. Comparative analysis between use of all bands, Indices, GPR-Band selection on PROSAIL simulated reflectance bands for retrieval show that Band selection has vastly improved the predictability with least computational cost, time compared all bands and at the same time prediction performance closer or at times even better.

The tree-based regressions performed superior in prediction with lower RMSE values and higher R^2 /MAE values. CCF showed the highest R^2 /MAE and lowest RMSE among all other ML regression types indicating its robustness in prediction than PLSR, LSLR, NN, SVR, KRR, GPR and other decision tree

methods of RF and RT. The results justify the selection of CCF as a reliable regression method for hybrid retrieval.

<Insert Table 4 here> <Insert Figure 4 here> <Insert Figure 5 here>

4.2 Results of hybrid PROSAIL-D inversion using CCF

The test image retrievals of Site (A) is shown in Fig. 6 includes two region of interests (ROIs). The ROI 1 in Fig. 6 comprise areas north east of Samarkha village with small densely distributed agricultural plots dominated by pockets of crops like maize, millets, vegetables and tobacco showing lower leaf level C_{ab}-leaf (0-25 µg.cm^{-2}) but have higher C_{ab} at canopy level due to higher LAI (3 to 4 m^2.m^{-2}), medium C_m (0.01-0.018 g.cm^{-2}), moderate C_c (10-20 µg.cm^{-2}) and lower C_a (0-10 µg.cm^{-2}) proving the larger presence of tree canopy and prominent vegetation around plots. The region of interest -2 in Fig. 6 show area north of Anand city with an agricultural tank surrounded by agricultural plots. The eastern part showed higher levels of C_{ab}-leaf (55-65 µg.cm^{-2}), C_w (0.02-0.03 cm) and lower level of LAI (1.5-2 m^2.m$^{-2)}$, C_c (12-17 µg.cm^{-2}), C_a (2-3 µg.cm^{-2}), C_m(0.008-0.01 g.cm^{-2}) indicating vegetation stage of summer crops: Paddy, fodder, Maize, Millets. The right part of ROI-2 showed lower levels of C_{ab}-leaf (55-65 µg.cm^{-2}), C_w (0.02-0.03 cm) and higher LAI (2.5-4 m^2.m$^{-2)}$, C_c (15-20 µg.cm^{-2}), C_a (6-12 µg.cm^{-2}), C_m (0.015-0.025 g.cm^{-2}) indicating senescent leaves in Ripening stage. The retrieval over Site B in Raichur (shown in Fig. 7) comprised the summer crops of Millets, Safflower, Bengalgram, Greengram, Redgram (pigeon pea) and other vegetable crops in vegetative stage. They have shown high C_{ab} (55-65 µg.cm^{-2}), medium C_c (20 -50 µg.cm^{-2}), lower C_a (10-15 µg.cm^{-2}). The low to medium variability of C_w (0.01-0.02 cm) is seen across the crops. The millets were seen to have comparatively higher C_m(0.01-0.018 g.cm^{-2}) compared to gram (0.006-0.01 g.cm^{-2}) and vegetable crops (0.008-0.012 g.cm^{-2}). variables retrieved match with site data correspondingly. LAI was seen higher in Safflower (2 to 2.3 m^2.m^{-2}) compared to millets (1.2-1.7 m^2.m^{-2}), gram (1-1.5 m^2.m^{-2}) crops. The coefficient of determination (R^2) for CCF hybrid retrieval of sites A and B AVIRIS-NG images for C_{ab}, C_c, C_a, C_w, C_m, LAI were 0.838, 0.4671, 0.625, 0.8159, 0.7343, 0.7644 respectively.

<Insert Table 5 here> <Insert Figure 6 here> <Insert Figure 7 here>

Important observation from the results point to the only moderate correlation of C_a and C_c prediction from PROSAIL-D inversion against observed values as a limitation of RTM based approach. PROSAIL inversion based retrieval did not disentangle spectral responses of C_a and C_c in heterogeneous species and are suitable to monocultures (Schiefer, Schmidtlein and Kattenborn, 2021). Given that that the relationship for the absorbance vs. content of anthocyanins does not maintain a linear relationship as the content grows higher, models based non-linear regression and not based on RTM based inversion tend to retrieve the leaf anthocyanin content with higher accuracy (Li and Huang, 2021). The reflectance based approach to retrieve

C_{ab}, C_c, C_a can utilize certain bands positioned outside the main absorption bands of the pigments (Gitelson and Solovchenko, 2018). The restriction of spectral ranges of C_a (522-622 nm) and C_c (482-562 nm) present work accounted for bands in those region and training with heterogeneous canopy configurations did affect the band selection proving that foliar absorbance based on concept of specific absorbance response would be able to overcome limitations in pigment retrieval typical for the reflectance based approaches (Gitelson and Solovchenko, 2018). The new insights are necessary of chlorophyll along with extra-plastidial (mainly vacuolar), such as anthocyanins and flavonoids, as well as other phenolic compounds green, blue, and ultraviolet/violet regions is necessary which have an overlap points in absorption spectrum (Falcioni *et al.*, 2020) (Gitelson *et al.*, 2017). The band selection for C_m could not give a stable pattern of accuracies and lesser number of features as optimal band combinations similar to other variables. Though C_m influences almost whole spectral range starting from red region (Verrelst and Rivera, 2016), the best band combination of wavelengths for C_m obtained are mostly from Shortwave Infrared (SWIR) region as also observed by (Wang *et al.*, 2011) proving that C_m is one of most important driver of reflectance in the SWIR region. The water content estimation in present work is in agreement with finding of most sensitive region for water content: the near-infrared region (Clevers, Kooistra and Schaepman, 2010). The smoothing of regions in spectra with noise using filters like Savitsky-Golay filter applying a polynomial fit within the window may allow consider more spectral region in SWIR for retrieval of C_m, C_w, LAI (Clevers, Kooistra and Schaepman, 2010). Different feature selection algorithms may be tested The problems are related to the structural differences of canopies and the effects of varying 'background effects' like soil color, moisture, shadows, the presence of other non-green landscape components along the effects of seasonality do influence the outcome of results (Hernández-Clemente *et al.*, 2019).

5. Conclusion

Airborne platforms that are flown at medium to high altitudes could be used as test beds for imaging technologies and can be used for validating satellite sensor imagery, as a prototype of satellite sensors yet to be launched or to just collect empirical data for development and testing new scientific algorithms (Myers and Miller., 2005). Quantitative vegetation variable extraction is fundamental to assess the dynamic response of vegetation to changing environmental conditions. Earth observation sensors in the optical domain enable the spatiotemporally explicit retrieval of plant BP-BC variables. This data stream has never been so rich as is foreseen with the new-generation imaging spectrometer missions. The HS space borne spectrometers of high spectral and spatial resolution like HySIS (India), FLEX (EU), DESIS (Germany), SHALOM (Israel), HyspIRI (US), ENMAP (Germany), CHIME (EU) offer excellent opportunities for research in this respect for agriculture crop monitoring. Likewise, low-altitude sensors of airborne platform can accurately map and be used in conjunction with satellite imagery to aid in the interpretation of the same

(Klemas, 2013). Most crop BP-BC variables were retrieved with better accuracy from AVIRIS-NG data. The band selection proves to be a cost-effective method to overcome data redundancy in high dimensional hyperspectral data. The GPR band selection used ranking and recursive elimination of features (bands) which fall under wrapper type techniques of feature selection helped identify the best subset hyperspectral bands from field hyperspectral dataset checked with k-fold cross validation. The chosen subset of characteristic reflectance bands displayed lower error in prediction of the BP-BC variable conveniently being assumed that displayed results depend on the crop types chosen from heterogeneous crop landscapes of sites: Raichur and Anand. The study showed the limitations of leaf-canopy radiative transfer model inversion approach in retrieval of carotenoid (C_c) and anthocyanin (C_a) despite robust hybrid regression using CCF. The study utilized CCFs for BP-BC retrieval which are capable to naturally represent data with correlated inputs and suitable for hyperspectral data with high band correlations. Besides, CCF can naturally accommodate multiple outputs and needs relatively lesser hyper parameter tuning as an advantage. The higher coefficient of determination was observed for BP-BC foliar retrievals from high spatial resolution hyperspectral remote sensing data from sensitive spectral regions using combination of GPR band selection and retrieval with CCF regression.

Acknowledgements

This work was supported by Space Applications Centre (SAC) of Indian Space Research Organization (ISRO) and uses data from NASA-ISRO partnered AVIRIS-NG airborne campaign programme. Authors also thank Nirma University, Ahmedabad for support.

References

Abdel-Rahman, E. M., Ahmed, F. B. and Ismail, R. (2013) 'Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data', *International Journal of Remote Sensing*, 34(2), pp. 712–728. doi: 10.1080/01431161.2012.713142.

ALLEN WA *et al.* (1969) 'Interaction of Isotropic Light With a Compact Plant Leaf', *J Opt Soc Amer*, 59(10), pp. 1376–1379. doi: 10.1364/josa.59.001376.

Atzberger, C. *et al.* (2013) 'Suitability and adaptation of PROSAIL radiative transfer model for hyperspectral grassland studies', *Remote Sensing Letters*, 4(1), pp. 56–65. doi: 10.1080/2150704X.2012.689115.

Bajcsy, P. and Groves, P. (2004) 'Methodology for hyperspectral band selection', *Photogrammetric Engineering and Remote Sensing*, 70(7), pp. 793–802. doi: 10.14358/PERS.70.7.793.

Berger, K. *et al.* (2018) 'Evaluation of the PROSAIL model capabilities for future hyperspectral model environments: A review study', *Remote Sensing*, 10(1). doi: 10.3390/rs10010085.

Bhattacharya, B. K. *et al.* (2019) 'An overview of AVIRIS-NG airborne hyperspectral science campaign over India', *Current Science*, 116(7), pp. 1082–1088. doi: 10.18520/cs/v116/i7/1082-1088.

Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2013) 'A review of feature selection methods on synthetic data', in *Knowledge and Information Systems*, pp. 483–519. doi: 10.1007/s10115-012-0487-8.
Cai, W., Li, Y. and Shao, X. (2008) 'A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra', *Chemometrics and Intelligent Laboratory Systems*, 90(2), pp. 188–194. doi: 10.1016/j.chemolab.2007.10.001.
Camps-Valls, G. *et al.* (2020) 'Statistical biophysical parameter retrieval and emulation with Gaussian processes', in *Data Handling in Science and Technology*, pp. 333–368. doi: 10.1016/B978-0-444-63977-6.00015-8.
Centner, V. *et al.* (1996) 'Elimination of Uninformative Variables for Multivariate Calibration', *Analytical Chemistry*, 68(21), pp. 3851–3858. doi: 10.1021/ac960321m.
Chapman, J. W. *et al.* (2019) 'Spectral and radiometric calibration of the Next Generation Airborne Visible Infrared Spectrometer (AVIRIS-NG)', *Remote Sensing*, 11(18). doi: 10.3390/rs11182129.
Chen, X. *et al.* (2014) 'Comparison of the sensor dependence of vegetation indices and vegetation water indices based on radiative transfer model', *Land Surface Remote Sensing II*, 9260, p. 92603G. doi: 10.1117/12.2069033.
Chen, Y. *et al.* (2015) 'Correlation coefficient optimization in partial least-squares regression with application to ATR-FTIR spectroscopic analysis', *Analytical Methods*. Royal Society of Chemistry, 7(14), pp. 5780–5786. doi: 10.1039/c5ay00441a.
Chloupek, O., Hrstkova, P. and Schweigert, P. (2004) 'Yield and its stability, crop diversity, adaptability and response to climate change, weather and fertilisation over 75 years in the Czech Republic in comparison to some European countries', *Field Crops Research*, 85(2–3), pp. 167–190. doi: 10.1016/S0378-4290(03)00162-X.
Clevers, J. G. P. W., Kooistra, L. and Schaepman, M. E. (2010) 'Estimating canopy water content using hyperspectral remote sensing data', *International Journal of Applied Earth Observation and Geoinformation*. Elsevier B.V., 12(2), pp. 119–125. doi: 10.1016/j.jag.2010.01.007.
Cundill, S. L., der van Werff, H. M. A. and der van Meijde, M. (2015) 'Adjusting spectral indices for spectral response function differences of very high spatial resolution sensors simulated from field spectra', *Sensors (Switzerland)*, 15(3), pp. 6221–6240. doi: 10.3390/s150306221.
Deguise, J. C. *et al.* (2015) 'Spatial High Resolution Crop Measurements with Airborne Hyperspectral Remote Sensing', pp. 1603–1608. doi: 10.2134/1999.precisionagproc4.c61b.
Doktor, D. *et al.* (2014) 'Extraction of plant physiological status from hyperspectral signatures using machine learning methods', *Remote Sensing*. MDPI AG, 6(12), pp. 12247–12274. doi: 10.3390/rs61212247.
Falcioni, R. *et al.* (2020) 'High resolution leaf spectral signature as a tool for foliar pigment estimation displaying potential for species differentiation', *Journal of Plant Physiology*. Elsevier, 249(March), p. 153161. doi: 10.1016/j.jplph.2020.153161.
Féret, J. B. *et al.* (2017) 'PROSPECT-D: Towards modeling leaf optical properties through a complete lifecycle', *Remote Sensing of Environment*. Elsevier Inc., 193, pp. 204–215. doi: 10.1016/j.rse.2017.03.004.
Fung, T., Yan Ma, H. F. and Siu, W. L. (2003) 'Band selection using hyperspectral data of subtropical tree species', *Geocarto International*, 18(4), pp. 3–11. doi: 10.1080/10106040308542284.
Gitelson, A. *et al.* (2017) 'In situ optical properties of foliar flavonoids: Implication for non-destructive estimation

of flavonoid content', *Journal of Plant Physiology*, 218(August), pp. 258–264. doi: 10.1016/j.jplph.2017.08.009.
Gitelson, A. A., Gritz †, Y. and Merzlyak, M. N. (2003) 'Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves', *Journal of Plant Physiology*, 160(3), pp. 271–282. doi: 10.1078/0176-1617-00887.
Gitelson, A. and Solovchenko, A. (2018) 'Non-invasive quantification of foliar pigments: Possibilities and limitations of reflectance- and absorbance-based approaches', *Journal of Photochemistry and Photobiology B: Biology*, 178(November 2017), pp. 537–544. doi: 10.1016/j.jphotobiol.2017.11.023.
Goel, P. K. *et al.* (2003) 'ESTIMATION OF CROP BIOPHYSICAL PARAMETERS THROUGH AIRBORNE AND FIELD HYPERSPECTRAL REMOTE SENSING', *Transactions of the ASAE*, 46(2000), pp. 1235–1246.
Hernández-Clemente, R. *et al.* (2019) 'Early Diagnosis of Vegetation Health From High-Resolution Hyperspectral and Thermal Imagery: Lessons Learned From Empirical Relationships and Radiative Transfer Modelling', *Current Forestry Reports*. Springer Science and Business Media LLC, 5(3), pp. 169–183. doi: 10.1007/s40725-019-00096-1.
Holmgren, P. and Thuresson, T. (1998) 'Satellite remote sensing for forestry planning—A review', *Scandinavian Journal of Forest Research*, 13(1–4), pp. 90–110. doi: 10.1080/02827589809382966.
Houborg, R. and McCabe, M. F. (2018) 'A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 135, pp. 173–188. doi: 10.1016/j.isprsjprs.2017.10.004.
Huang, W. *et al.* (2018) 'Monitoring Crop Carotenoids Concentration by Remote Sensing', *Progress in Carotenoid Research*, (September), p. 197. doi: 10.5772/intechopen.78239.
Jacquemoud, S. *et al.* (2009) 'PROSPECT + SAIL models: A review of use for vegetation characterization', *Remote Sensing of Environment*. Elsevier Inc., 113(SUPPL. 1). doi: 10.1016/j.rse.2008.01.026.
Jacquemoud, S. and Baret, F. (1990) 'PROSPECT: A model of leaf optical properties spectra', *Remote Sensing of Environment*, 34(2), pp. 75–91. doi: 10.1016/0034-4257(90)90100-Z.
Jha, M. N., Levy, J. and Gao, Y. (2008) 'Advances in remote sensing for oil spill disaster management: State-of-the-art sensors technology for oil spill surveillance', *Sensors*, 8(1), pp. 236–255. doi: 10.3390/s8010236.
Jin, J. and Wang, Q. (2019) 'Selection of Informative Spectral Bands for PLS Models to Estimate Foliar Chlorophyll Content Using Hyperspectral Reflectance', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 57(5), pp. 3064–3072. doi: 10.1109/TGRS.2018.2880193.
Kattenborn, T. *et al.* (2019) 'Advantages of retrieving pigment content [µg/cm2] versus concentration [%] from canopy reflectance', *Remote Sensing of Environment*. Elsevier, 230(May), p. 111195. doi: 10.1016/j.rse.2019.05.014.
Khalid, S., Khalil, T. and Nasreen, S. (2014) 'A survey of feature selection and feature extraction techniques in machine learning', *Proceedings of 2014 Science and Information Conference, SAI 2014*, pp. 372–378. doi: 10.1109/SAI.2014.6918213.
Kira, O. *et al.* (2016) 'Informative spectral bands for remote green LAI estimation in C3 and C4 crops', *Agricultural and Forest Meteorology*. Elsevier B.V., 218–219, pp. 243–249. doi: 10.1016/j.agrformet.2015.12.064.
Klemas, V. (2013) 'Airborne remote sensing of coastal features and processes: An overview', *Journal of Coastal*

Research, 29(2), pp. 239–255. doi: 10.2112/JCOASTRES-D-12-00107.1.
Koponen, S. *et al.* (2007) 'A case study of airborne and satellite remote sensing of a spring bloom event in the Gulf of Finland', *Continental Shelf Research*, 27(2), pp. 228–244. doi: 10.1016/j.csr.2006.10.006.
Kumar, V. (2014) 'Feature Selection: A literature Review', *The Smart Computing Review*, 4(3). doi: 10.6029/smartcr.2014.03.007.
Li, Ji *et al.* (2020) 'Seasonal variations in the relationship between sun-induced chlorophyll fluorescence and photosynthetic capacity from the leaf to canopy level in a rice crop', *Journal of Experimental Botany*, 71(22), pp. 7179–7197. doi: 10.1093/jxb/eraa408.
Li, X. *et al.* (2014) 'Exploring the best hyperspectral features for LAI estimation using partial least squares regression', *Remote Sensing*, 6(7), pp. 6221–6241. doi: 10.3390/rs6076221.
Li, Y. and Huang, J. (2021) 'Leaf anthocyanin content retrieval with partial least squares and gaussian process regression from spectral reflectance data', *Sensors*, 21(9). doi: 10.3390/s21093078.
Liang, L. *et al.* (2015) 'Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method', *Remote Sensing of Environment*. Elsevier Inc., 165, pp. 123–134. doi: 10.1016/j.rse.2015.04.032.
Liu, P., Shi, R. and Gao, W. (2018) 'Estimating leaf chlorophyll contents by combining multiple spectral indices with an artificial neural network', *Earth Science Informatics*. Earth Science Informatics, 11(1), pp. 147–156. doi: 10.1007/s12145-017-0319-1.
Lucas, R. *et al.* (2007) 'Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land cover mapping', *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(3), pp. 165–185. doi: 10.1016/j.isprsjprs.2007.03.003.
Mahmoud, H. F. F. (2021) 'Parametric Versus Semi and Nonparametric Regression Models', *International Journal of Statistics and Probability*, 10(2), p. 90. doi: 10.5539/ijsp.v10n2p90.
Malhi, R. K. M. *et al.* (2020) 'Band selection algorithms for foliar trait retrieval using AVIRIS-NG: a comparison of feature based attribute evaluators', *Geocarto International*. Taylor & Francis, 0(0), pp. 1–17. doi: 10.1080/10106049.2020.1870167.
Mumby, P. J. *et al.* (1997) 'Measurement of seagrass standing crop using satellite and digital airborne remote sensing', *Marine Ecology Progress Series*, 159, pp. 51–60. doi: 10.3354/meps159051.
Myers, J. S. and Miller., R. L. (2005) *Remote Sensing of Coastal Aquatic Environments*, *Remote Sensing and Digital Image Processing*. Edited by R. L. Miller., C. E. DELCASTILLO, and B. A. MCKEE. Springer Netherlands. doi: 10.1007/1-4020-3100-9_5.
Pal, M. (2006) 'Support vector machine-based feature selection for land cover classification: A case study with DAIS hyperspectral data', *International Journal of Remote Sensing*, 27(14), pp. 2877–2894. doi: 10.1080/01431160500242515.
Pasqualotto, N. *et al.* (2018) 'Retrieval of canopy water content of different crop types with two new hyperspectral indices: Water Absorption Area Index and Depth Water Index', *International Journal of Applied Earth Observation and Geoinformation*. Elsevier B.V., 67, pp. 69–78. doi: 10.1016/j.jag.2018.01.002.
Poldrack, R. A., Huckins, G. and Varoquaux, G. (2020) 'Establishment of Best Practices for Evidence for

Prediction: A Review', *JAMA Psychiatry*, 77(5), pp. 534–540. doi: 10.1001/jamapsychiatry.2019.3671.
Pu, R. (2012) 'Comparing canonical correlation analysis with partial least squares regression in estimating forest leaf area index with multitemporal landsat TM imagery', *GIScience and Remote Sensing*, 49(1), pp. 92–116. doi: 10.2747/1548-1603.49.1.92.
Quemada, M., Gabriel, J. L. and Zarco-Tejada, P. (2014) 'Airborne hyperspectral images and ground-level optical sensors as assessment tools for maize nitrogen fertilization', *Remote Sensing*, 6(4), pp. 2940–2962. doi: 10.3390/rs6042940.
Rainforth, T. and Wood, F. (2015) 'Canonical Correlation Forests', pp. 1–51. Available at: http://arxiv.org/abs/1507.05444.
Saltelli, A., Tarantola, S. and Chan, K. P. S. (1999) 'A quantitative model-independent method for global sensitivity analysis of model output', *Technometrics*, 41(1), pp. 39–56. doi: 10.1080/00401706.1999.10485594.
Schiefer, F., Schmidtlein, S. and Kattenborn, T. (2021) 'The retrieval of plant functional traits from canopy spectra through RTM-inversions and statistical models are both critically affected by plant phenology', *Ecological Indicators*. Elsevier Ltd, 121(June 2020), p. 107062. doi: 10.1016/j.ecolind.2020.107062.
Schweiger, A. K. *et al.* (2016) 'How to predict plant functional types using imaging spectroscopy: linking vegetation community traits, plant functional types and spectral response'. doi: 10.1111/2041-210X.12642.
Sobrino, J. A., Jiménez-Muñoz, J. C. and Verhoef, W. (2005) 'Canopy directional emissivity: Comparison between models', *Remote Sensing of Environment*. Elsevier, 99(3), pp. 304–314. doi: 10.1016/J.RSE.2005.09.005.
Sun, W. and Du, Q. (2019) 'Hyperspectral band selection: A review', *IEEE Geoscience and Remote Sensing Magazine*. IEEE, 7(2), pp. 118–139. doi: 10.1109/MGRS.2019.2911100.
Ustin, S. L. *et al.* (2009) 'Retrieval of foliar information about plant pigment systems from high resolution spectroscopy', *Remote Sensing of Environment*, 113, pp. S67–S77. doi: 10.1016/j.rse.2008.10.019.
Verhoef, W. (1984) 'Light scattering by leaf layers with application to canopy reflectance modeling: The SAIL model', *Remote Sensing of Environment*, 16(2), pp. 125–141. doi: 10.1016/0034-4257(84)90057-9.
Verhoef, W. *et al.* (2007) 'Unified optical-thermal four-stream radiative transfer theory for homogeneous vegetation canopies', *IEEE Transactions on Geoscience and Remote Sensing*, 45(6), pp. 1808–1822. doi: 10.1109/TGRS.2007.895844.
Verhoef, W. and Bach, H. (2003) 'Simulation of hyperspectral and directional radiance images using coupled biophysical and atmospheric radiative transfer models', *Remote Sensing of Environment*, 87(1), pp. 23–41. doi: 10.1016/S0034-4257(03)00143-3.
Verrelst, J. *et al.* (2011) 'ARTMO: an Automated Radiative Transfer Models Operator toolbox for automated retrieval of biophysical parameters through model inversion', *Proceedings of EARSeL 7th SIG-Imaging Spectroscopy Workshop, Edinburgh, UK*, pp. 11–13.
Verrelst, J. *et al.* (2015) 'Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties - A review', *ISPRS Journal of Photogrammetry and Remote Sensing*. Elsevier B.V., pp. 273–290. doi: 10.1016/j.isprsjprs.2015.05.005.
Verrelst, J. *et al.* (2016a) 'Spectral band selection for vegetation properties retrieval using Gaussian processes

regression', *International Journal of Applied Earth Observation and Geoinformation*. Elsevier B.V., 52, pp. 554–567. doi: 10.1016/j.jag.2016.07.016.
Verrelst, J. *et al.* (2016b) 'Spectral band selection for vegetation properties retrieval using Gaussian processes regression', *International Journal of Applied Earth Observation and Geoinformation*. Elsevier B.V., 52, pp. 554–567. doi: 10.1016/j.jag.2016.07.016.
Verrelst, J. *et al.* (2016c) 'Spectral band selection for vegetation properties retrieval using Gaussian processes regression', *International Journal of Applied Earth Observation and Geoinformation*. Elsevier B.V., 52, pp. 554–567. doi: 10.1016/j.jag.2016.07.016.
Verrelst, J., Vicent, J., *et al.* (2019) 'Global Sensitivity Analysis of Leaf-Canopy-Atmosphere RTMs: Implications for Biophysical Variables Retrieval from Top-of-Atmosphere Radiance Data', *Remote Sensing*. MDPI AG, 11(16), p. 1923. doi: 10.3390/rs11161923.
Verrelst, J., Malenovský, Z., *et al.* (2019) 'Quantifying Vegetation Biophysical Variables from Imaging Spectroscopy Data: A Review on Retrieval Methods', *Surveys in Geophysics*. Springer Netherlands, 40(3), pp. 589–629. doi: 10.1007/s10712-018-9478-y.
Verrelst, J. and Rivera, J. P. (2016) 'A Global Sensitivity Analysis Toolbox to Quantify Drivers of Vegetation Radiative Transfer Models', in *Sensitivity Analysis in Earth Observation Modelling*. Elsevier Inc., pp. 319–339. doi: 10.1016/B978-0-12-803011-0.00016-1.
Verrelst, J., Rivera, J. P. and Moreno, J. (2015) 'ARTMO'S GLOBAL SENSITIVITY ANALYSIS (GSA) TOOLBOX TO QUANTIFY DRIVING VARIABLES OF LEAF AND CANOPY RADIATIVE TRANSFER MODELS', in *9th EARSeL Imaging Spectroscopy Workshop*, pp. 1–11. doi: 10.12760/02-2015-2-01.
Wang, L. *et al.* (2011) 'Estimating dry matter content from spectral reflectance for green leaves of different species', *International Journal of Remote Sensing*, 32(22), pp. 7097–7109. doi: 10.1080/01431161.2010.494641.
Wang, S. *et al.* (2021) 'Airborne hyperspectral imaging of nitrogen deficiency on crop traits and yield of maize by machine learning and radiative transfer modeling', *International Journal of Applied Earth Observation and Geoinformation*. Elsevier B.V., 105(November), p. 102617. doi: 10.1016/j.jag.2021.102617.
Weiss, M., Jacob, F. and Duveiller, G. (2020) 'Remote sensing for agricultural applications: A meta-review', *Remote Sensing of Environment*. Elsevier, 236(August 2019), p. 111402. doi: 10.1016/j.rse.2019.111402.
Wellburn, A. R. (1994) 'The spectral determination of chlorophylls a and b, as well as total carotenoids, using various solvents with spectrophotometers of different resolutions.', *Journal of Plant Physiology*, pp. 307–313.
Yang, P., Verhoef, W. and van der Tol, C. (2020) 'Unified four-stream radiative transfer theory in the optical-thermal domain with consideration of fluorescence for multi-layer vegetation canopies', *Remote Sensing*, 12(23), pp. 1–19. doi: 10.3390/rs12233914.
Yilmaz, M. T., Hunt, E. R. and Jackson, T. J. (2008) 'Remote sensing of vegetation water content from equivalent water thickness using satellite imagery', *Remote Sensing of Environment*, 112(5), pp. 2514–2522. doi: 10.1016/j.rse.2007.11.014.
Zarco-Tejada, P. J., Guillén-Climent, M. L., *et al.* (2013) 'Estimating leaf carotenoid content in vineyards using high resolution hyperspectral imagery acquired from an unmanned aerial vehicle (UAV)', *Agricultural and Forest*

Meteorology, 171–172, pp. 281–294. doi: 10.1016/j.agrformet.2012.12.013.

Zarco-Tejada, P. J., Catalina, A., *et al.* (2013) 'Relationships between net photosynthesis and steady-state chlorophyll fluorescence retrieved from airborne hyperspectral imagery', *Remote Sensing of Environment*, 136, pp. 247–258. doi: 10.1016/j.rse.2013.05.011.

Zhang, X. *et al.* (2018) 'Potential investigation of linking PROSAIL with the Ross-Li BRDF model for vegetation characterization', *Remote Sensing*, 10(3). doi: 10.3390/rs10030437.

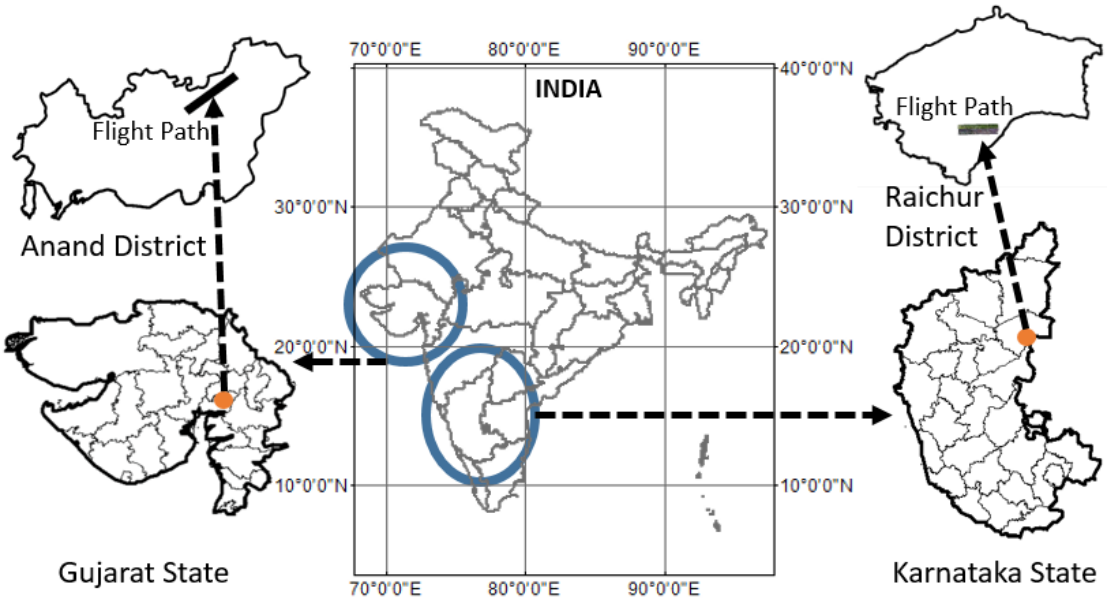


Figure 1. Map showing the study area and flight paths of AVIRIS-NG aerial survey in 2018 conducted in India for two study sites.

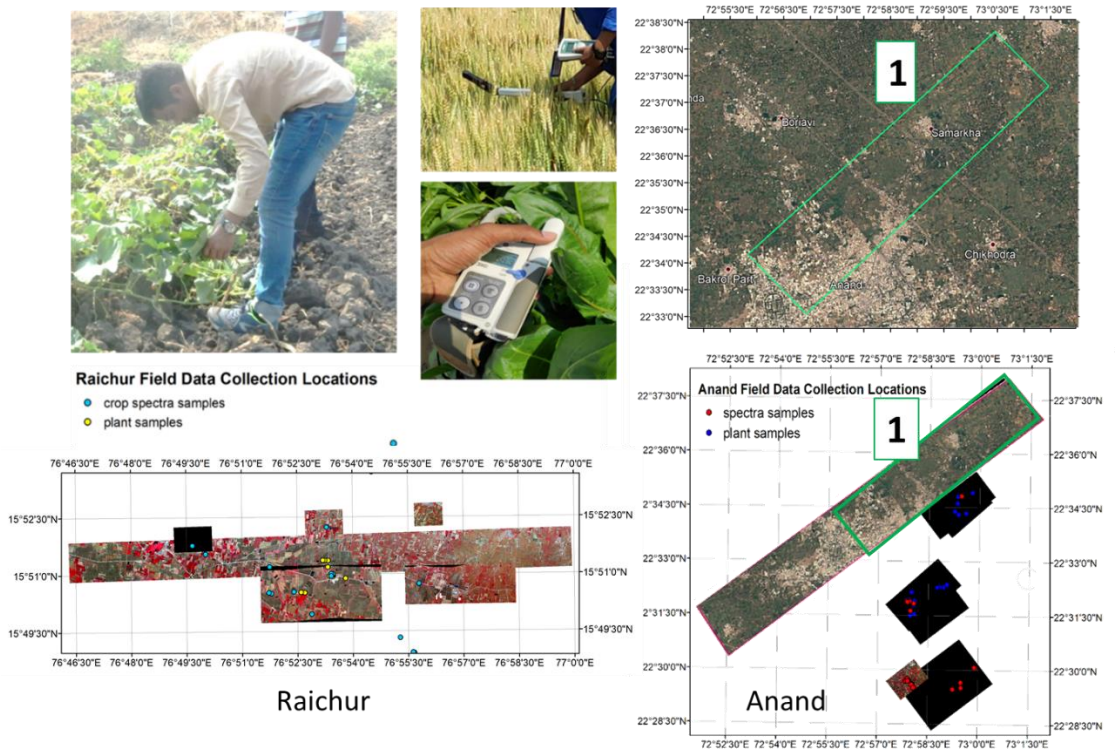


Figure 2 : Field data collection at two diverse agriculture system Raichur and Anand (a) Destructive ground sampling (top left) (b) LAI measurement (top-center) (c) Chlorophyll Index measurement(mid-center) (d) aerial coverage of Site A - Anand (bottom-right) (e) aerial coverage of Site B – Raichur (bottom-left)

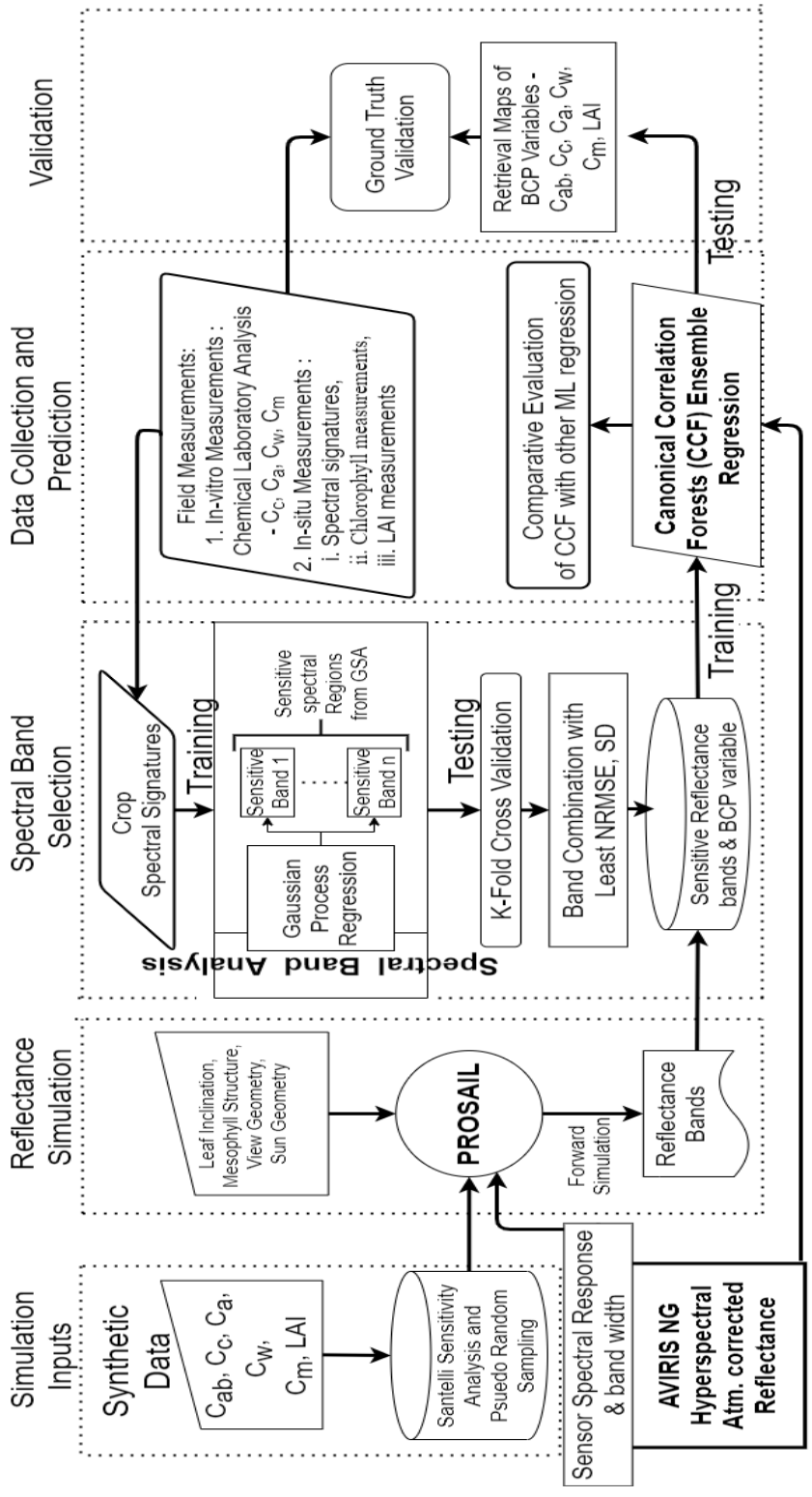


Figure 3. Flowchart of methodology

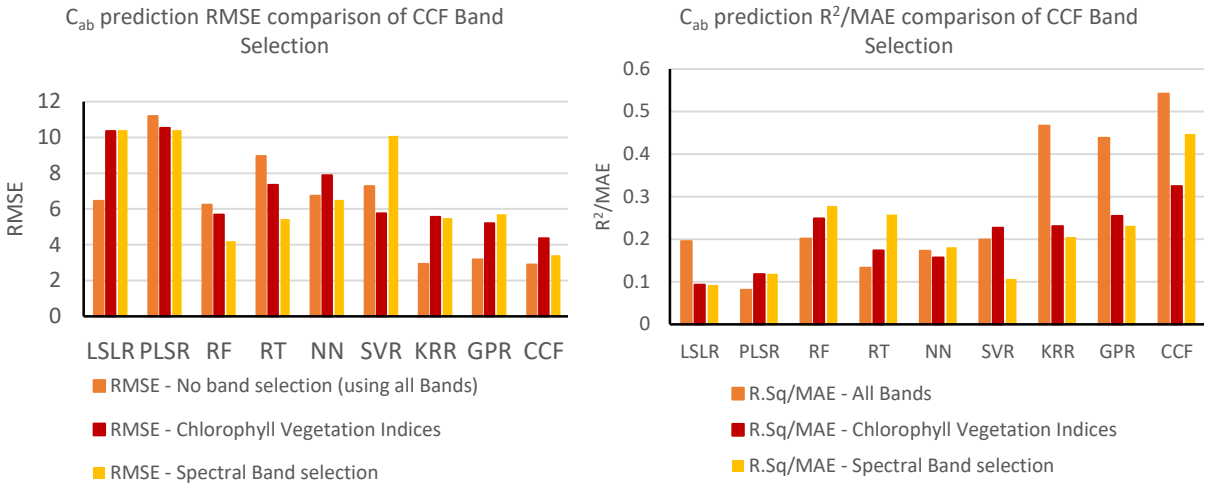
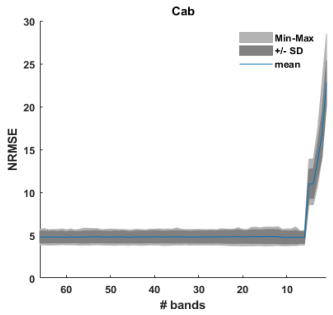
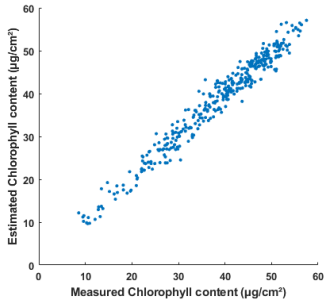
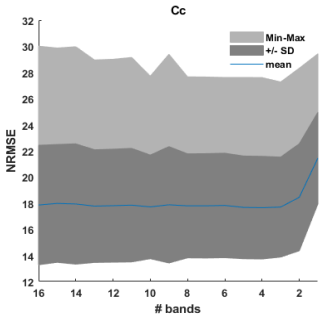
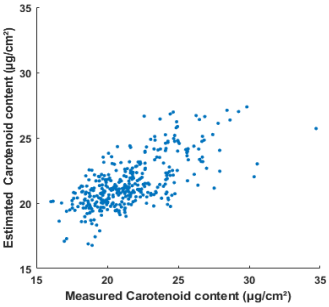
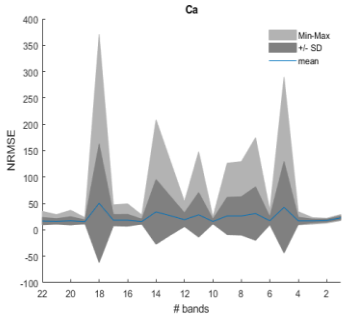
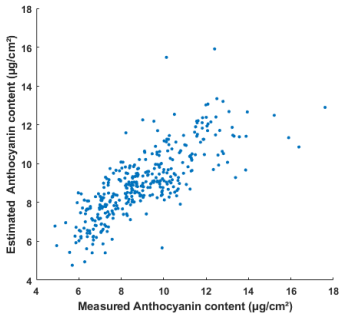
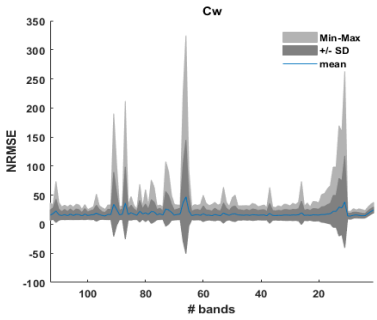
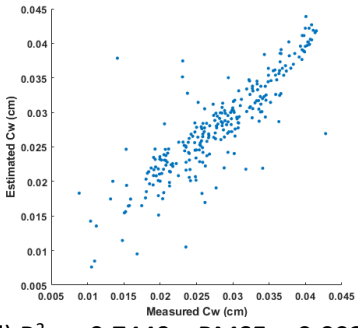
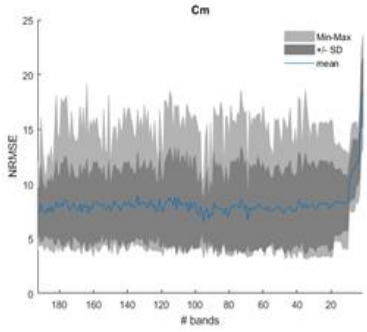
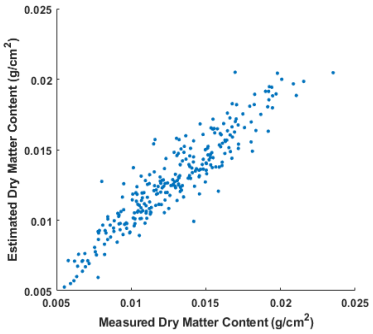


Figure 4. Statistics of (a) RMSE, (b) R^2/ MAE ratio for CCF prediction using band selection against (i) No band selection (ii) C_{ab} vegetation Indices; in comparison with prediction using LSLR, PLSR, RF, RT, NN, SVR, KRR, GPR over simulations for C_{ab}

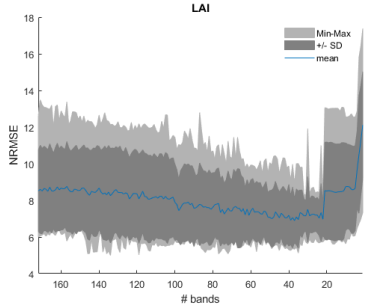
NRMSE of band combinations	Cross-validation of sampling
	
(a)C_{ab} – 6 bands	(a) R^2_{cv} = 0.9647 $RMSE_{cv}$ = 2.0821
	
(b) C_c – 4 bands	(b) R^2_{cv} = 0.4716 $RMSE_{cv}$ = 1.9838
	
(c)C_a – 10 bands	(c) R^2_{cv}=0.6182 $RMSE_{cv}$=1.2665
	
(d) C_w – 9 bands	(d) R^2_{cv} = 0.7449 $RMSE_{cv}$=0.0038



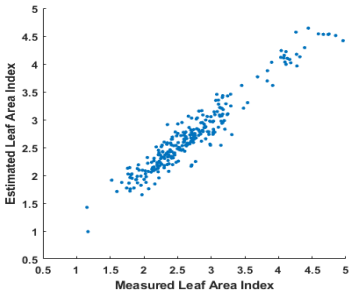
(e) C_m – 21 bands (or 95 bands)



(e) R^2_{cv} = 0.8669 $RMSE_{cv}$ = 0.0013



(f) LAI- 22 bands



(f) R^2_{cv}=0.9215 $RMSE_{cv}$ = 0.2405

Figure 5. (1) Normalized RMSE for band combinations; (2) Cross-validation statistics of CCF hybrid regression with selected bands of all BP-BC variables over field dataset

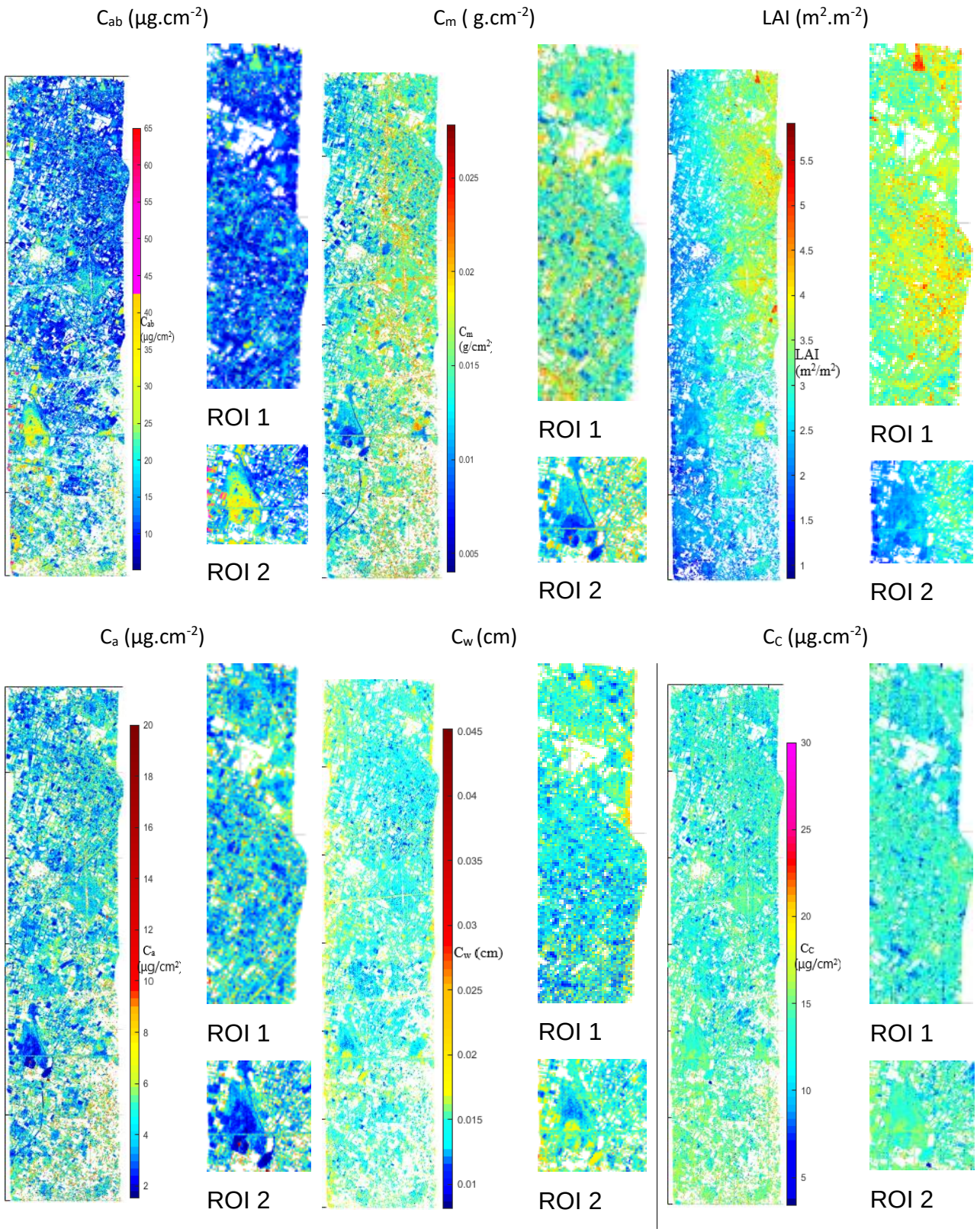


Figure 6: Retrieval of BP-BC variables by PROSAIL inversion using hybrid CCF regression for Site A

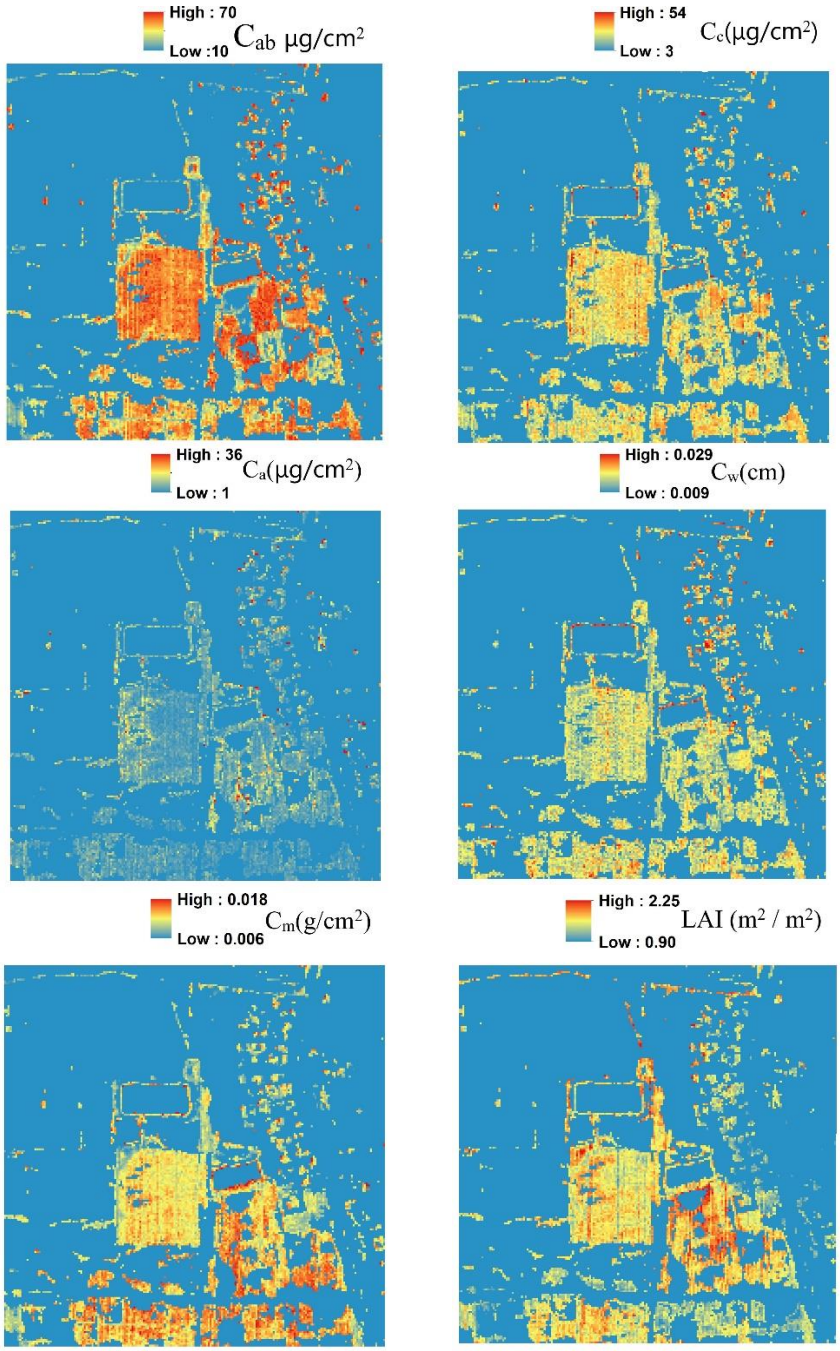


Figure 7: Retrieval of BP-BC variables by PROSAIL inversion using hybrid CCF regression for a ROI in Site B

Table 1. Basic agro-climatic, soil and crop information on the study site.

Characteristics	Raichur (Karnataka)	Anand (Gujarat)
Agro-climatic Zone	Southern Plateau and Hills Region (SPHR	Gujarat Plains and Hills Region (GPHR)
Climate	Hot Semi-arid ecoregion	Hot semi-arid ecoregion
Soil Characteristics		
Topography	Gently sloping Interfluves and Deccan Platea	Gently sloping Alluvial Plain and Central (Malwa) Highland
Drainage	Moderately deep, Well drained	Deep, Well drained
Texture	Clayey soils to loamy soils with low AWC	Fine to coarse-loamy soils
Type	Pellusterts, Chromusterts – Medium Black Soils (Vertisols)	Udifluents –Younger Alluvial Soils (Entisols)
Average Rainfall	713 mm	882 mm
Average PET	1950 mm	1550 mm
Prominent Crops	Pearl Millet (Bajra), Sorghum (Jowar), groundnut, Cotton, Red gram, Green gram, Bengal gram, Safflower, Vegetables (Bottle gourd, Brinjal, Ridge gourd)	Wheat, Bajra, Banana, Beans, Brinjal, Cabbage, Castor, Cauliflower, Fodder, Maize, Lemon, Tobacco, Onion, Pumpkin, Tomato

Table 2. Description of AVIRIS NG Datasets used for study.

Dataset name	Date of Flight	District	(Latitude, Longitude)	Sampling	Average Tile Area	Spectral Bands	View Angle	Spectral Range
AVIRIS NG Hyperspectral Reflectance (Atm. corrected)	24-02-2018	Raichur (Karnataka)	(15° 51' 28" N, 76° 52' 40" E)	~4 m @ 5 km altitude	6 km × 49.4 km (744 × 6179 pixels)	425 bands with 5 nm ± 0.5 nm width	36 ± 2 degrees	381-2500 Nanometer
AVIRIS NG Hyperspectral Reflectance (Atm. corrected)	26-03-2018	Anand (Gujarat)	(22° 34' 26" N, 72° 56' 49" E)	~4 m @ 5 km altitude	5.5 km × 40 km (693 × 4864 pixels)	425 bands with 5 nm ± 0.5 nm width	36 ± 2 degrees	381-2500 Nanometer

Table 3. Input parameters of PROSAIL Model and their ranges.

Leaf parameter	Symbol (Unit)	Range of parameter
Leaf Mesophyll Structure Parameter	N	1-2.5
Chlorophyll Content (in μg/cm^2)	C_{ab}	1-75
Carotenoid Content (in μg/cm^2)	C_c	1-70
Anthocyanin Content (in μg/cm^2)	C_a	1-50
Brown Pigment Content (arbitrary unit)	C_b	0.01-1
Equivalent Water Thickness (in cm)	C_w	0.004-0.05
Dry Matter Content (in g/cm^2)	C_m	0.002-0.03
Leaf Area Index (m^2 / m^2)	LAI	1-7
Leaf Inclination Angle	LIA	0-90°

Table 4. Statistics of spectral bands selection of AVIRIS-NG in terms of NRMSE and SD for optimum band combination and selection

No.of Bands	NRMSE	SD	Wavelengths	No.of Bands	NRMSE	SD	Wavelengths
Chlorophyll				Carotenoid			
10	4.7467	0.7579	457, 467, 472, 487, 552, 557, 597, 707, 712, 762	10	17.711	3.9669	482, 487, 517, 522, 527, 532, 537, 542, 547, 557
9	4.7469	0.7444	457, 467, 472, 487, 552, 597, 707, 712, 762	9	17.871	4.4532	482, 487, 522, 527, 532, 537, 542, 547, 557
8	4.7702	0.7853	457, 467, 487, 552, 597, 707, 712, 762	8	17.792	3.9804	482, 487, 522, 532, 537, 542, 547, 557
7	4.7416	0.6769	457, 467, 487, 552, 707, 712, 762	7	17.793	3.9944	482, 487, 522, 537, 542, 547, 557
6	**4.7262**	**0.6755**	**457, 467, 487, 552, 707, 762**	6	17.822	3.9979	482, 487, 522, 542, 547, 557
5	11.0011	1.7148	457, 467, 487, 552, 707	5	17.674	3.9311	487, 522, 542, 547, 557
4	10.9962	1.72	457, 487, 552, 707	**4**	**17.652**	**3.9291**	**487, 522, 542, 557**
3	13.7567	1.979	457, 487, 552	3	17.697	3.8255	487, 522, 557
2	16.7895	2.662	457, 487	2	18.446	4.0962	487, 522
1	22.8132	2.5759	457	1	21.450	3.5063	487
Equivalent Water Thickness				Anthocyanin			
10	14.2968	4.6151	1123, 1128, 1163, 1188, 1644, 1659, 1674, 1984, 1093, 1138	**10**	**15.915**	**4.4148**	**517, 522, 537, 542, 547, 552, 562, 592, 602, 617**
9	**14.1463**	**4.7284**	**1123, 1128, 1163, 1188, 1659, 1674, 1984, 1093, 1138**	9	26.419	35.350	517, 522, 537, 547, 552, 562, 592, 602, 617
8	15.6354	5.2653	1123, 1128, 1163, 1188, 1659, 1674, 1984, 1093	8	26.376	36.362	517, 522, 537, 547, 552, 562, 592, 617
7	16.0675	4.6898	1128, 1163, 1188, 1659, 1674, 1984, 1093	7	30.930	50.710	517, 522, 537, 547, 552, 592, 617
6	15.3126	4.5827	1128, 1163, 1659, 1674, 1984, 1093	6	17.227	7.5152	517, 522, 537, 547, 552, 592
5	14.6485	4.1879	1128, 1659, 1674, 1984, 1093	5	42.854	86.8	517, 522, 537, 547, 552
4	14.7039	3.6619	1128, 1659, 1984, 1093	4	17.064	6.9286	517, 522, 537, 547
3	18.3311	4.7427	1659, 1984, 1093	3	16.941	3.6308	517, 522, 537
2	21.8795	5.0181	1984, 1093	2	17.564	2.774	517, 537
1	24.6129	5.2656	1093	1	22.931	3.5585	537
Leaf Area Index				Dry Matter Content			
22	6.966	0.8812	411, 441, 537, 577, 582, 592, 602, 607, 632, 642, 647, 652, 692, 712, 727, 882, 1644, 1689, 1984, 1989, 2004, 2084	21	7.6792	3.9345	697, 702, 707, 742, 977, 992, 1113, 1198, 1208, 1243, 1263, 1268, 1503, 1508, 1513, 1553, 1573, 1659, 1684, 1984, 2089
21	8.5319	2.641	411, 441, 537, 577, 582, 592, 602, 607, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 1984, 1989, 2004, 2084	20	8.0364	4.0264	697, 702, 707, 742, 977, 992, 1113, 1198, 1208, 1243, 1263, 1268, 1503, 1508, 1513, 1553, 1659, 1684, 1984, 2089
20	8.5073	2.675	411, 441, 537, 577, 582, 592, 602, 607, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 1984, 1989, 2084	19	8.0004	4.0037	697, 702, 707, 742, 977, 992, 1113, 1208, 1243, 1263, 1268, 1503, 1508, 1513, 1553, 1659, 1684, 1984, 2089
19	8.5298	2.616	411, 441, 537, 577, 582, 592, 602, 607, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 1984, 2084.00	18	8.3385	3.0371	697, 702, 707, 742, 977, 992, 1113, 1208, 1243, 1263, 1268, 1503, 1508, 1513, 1553, 1659, 1684, 2089
18	8.5124	2.6621	411, 441, 537, 577, 582, 592, 602, 607, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 2084	17	8.3127	2.9997	697, 702, 707, 742, 977, 992, 1113, 1208, 1263, 1268, 1503, 1508, 1513, 1553, 1659, 1684, 2089
17	8.4877	2.6592	411, 441, 537, 582, 592, 602, 607, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 2084	16	8.2961	2.9827	697, 702, 707, 742, 977, 992, 1113, 1208, 1263, 1268, 1503, 1513, 1553, 1659, 1684, 2089
16	8.4435	2.6544	411, 441, 537, 592, 602, 607, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 2084	15	8.3541	3.0824	697, 702, 707, 742, 977, 992, 1113, 1208, 1263, 1268, 1503, 1513, 1553, 1659, 2089
15	8.4379	2.6362	411, 441, 537, 602, 607, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 2084	14	8.3389	2.9355	697, 702, 707, 742, 977, 1113, 1208, 1263, 1268, 1503, 1513, 1553, 1659, 2089
14	8.4619	2.6177	411, 441, 537, 607, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 2084	13	8.2666	2.5833	697, 702, 707, 977, 1113, 1208, 1263, 1268, 1503, 1513, 1553, 1659, 2089
13	8.4674	2.6156	411, 441, 537, 632, 642, 647, 652, 692, 712, 727, 1644, 1689, 2084	12	8.2536	2.5565	697, 702, 707, 977, 1113, 1208, 1263, 1268, 1513, 1553, 1659, 2089

Table 5. Validation statistics of retrieved crop BP-BC variables using CCF

Variable	Symbol	R^2	RMSE
Chlorophyll	C_{ab}	0.838	6.161 μg/cm^2
Carotenoid	C_c	0.4671	14.370 μg/cm^2
Anthocyanin	C_a	0.525	12.924 μg/cm^2
Equivalent Water Thickness	C_w	0.8159	0.002 cm
Dry Matter Content	C_m	0.7343	0.003 g/cm^2
Leaf Area Index	LAI	0.7644	0.350 m^2 / m^2