Final Report 10 YEARS OF SMOS – PASSIVE MICROWAVE VEGETATION OPACITY STUDY (PM-VO-S) OSMOSE

Arnaud Mialon, Simon Boitard, Nemesio Rodríguez-Fernández, Julio César Salazar-Neira, Beatrice Berthelot, Daria Malik, Mike Schwank, Yiwen Zhou, Frédéric Baret, Hongliang Ma, et al.

▶ **To cite this version:**

Arnaud Mialon, Simon Boitard, Nemesio Rodríguez-Fernández, Julio César Salazar-Neira, Beatrice Berthelot, et al.. Final Report 10 YEARS OF SMOS – PASSIVE MICROWAVE VEGETATION OPACITY STUDY (PM-VO-S) OSMOSE. SO-TN-CB-GS-0119, CESBIO. 2024, pp.SO-TN-CB-GS-0119. ⟨hal-04840129⟩

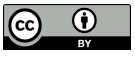
HAL Id: hal-04840129

https://hal.science/hal-04840129v1

Submitted on 16 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

10 YEARS OF SMOS – PASSIVE MICROWAVE VEGETATION OPACITY STUDY (PM-VO-S)

OSMOSE

Contract No. 4000137990/22/NL/IA
Ref. number: SO-TN-CB-GS-0119

Final Report

September 26, 2024

Prepared by:

CESBIO, Magellium, WSL, EMMAH, GlobEO

Version History

Issue	Date	Details	Editor
1.0	26 Sept. 2024	version sent to ESA	Arnaud Mialon
0.1	21 Feb. 2024	Initial document template	Arnaud Mialon

Contacts: Arnaud Mialon (arnaud.mialon@univ-tlse3.fr) or Béatrice Berthelot (beatrice.berthelot@magellium.fr).

Authors

Partener	Author Name
CESBIO	Arnaud Mialon S. Boitard N. Rodriguez-Fernandez J.C. Salazar
Magellium	B. Berthelot D. Malik
WSL	M. Schwank Y. Zhou
EMMAH	F. Baret H. Ma M. Weiss
GlobEO	S. Mermoz

Acknowledgements

Contents

ACRONYMS

2S	Two Stream Radiative Transfer Model
AMSRE	Advanced Microwave Scanning Radiometer-EOS
AMSR2	Advanced Microwave Scanning Radiometer 2
ADF	Auxiliary Data File
ATBD	Algorithm Theoretical Baseline Document
CATDS	Centre Aval de Traitement des Données SMOS
CESBIO	Centre d'Etudes Spatiales de la Biosphère
DQX	Data Quality Index
EASE	Equal-Area Scalable Earth Grid
ECMWF	European Centre for Medium-range Weather Forecasting
FAPAR	Fraction of Absorbed Photosynthetically Active Radiation
IGBP	International Geosphere-Biosphere Programme
LAI	Leaf Area Index
L-MEB	L - band Microwave Emission of the Biosphere
MD	Cardioid Model
MEMLV	Microwave Emission Model for Layered Vegetation
NDVI	Normalized Difference Vegetation Index
RFI	Radiometric Frequency Interference
RTM	Radiative Transfer Model
SM	Soil Moisture
SMOS	Soil Moisture and Ocean Salinity
Tau/τ	Vegetation Optical thickness
TB	Brightness Temperature
VOD	Vegetation Optical Depth
WP	Work Package

Part I: Project description

Chapter 1

Introduction

This technical note is the final report of the 10 YEARS OF SMOS – PASSIVE MICROWAVE VEGETATION OPACITY STUDY (PM-VO-S), which the consortium (CESBIO, Magellium, INRAE, WSL, GlobEO) carried out from June 2022 to September 2024, in the context of the ESA contract No. 4000137990/22/NL/IA. This final report describes the context and the activities done during these two years.

The first sections give the context of the project (Chapter 2 and Chapter 3). Chapter 4 then introduces an innovative model to describe better two essential microwave parameters representing the impact of vegetation that are τ and ω. Chapter 5 describes the method to derive the multi-frequency VOD, and an evaluation of the VOD. Chapter 6 details the Above Ground Biomass estimates obtained from the VOD. It also provides a validation of the derived AGB with in situ data. Chapter 7 starts the optical part, by giving the context and state of the art of the vegetation indices derived from optical data. Chapter 7 then describes a method of a new dataset which estimates the water content of the leaves components of the vegetation. Chapter 8 details the datacube 2 and the methodology to produce homogenized datasets. The analysis and the complementarity of all the indices to study the vegetation are analysed in Chapter 9. Chapter 12 concludes the project, and we lastly provide details of the project in terms of deliverables and meeting Chapter 10, outreach and communication Chapter 11.

Chapter 2

Scientific background

The thermal emission arising from the Earth surface at microwave frequencies depends on the soil characteristics such as its temperature, roughness and soil moisture content, which control the soil emissivity (Ulaby, 1976). In the presence of vegetation, part of the soil emission is absorbed and scattered. These effects can be parameterized using radiative transfer models such as the τ-ω model (*Mo et al.* (1982), *Ulaby and Wilson* (1985), *Ferrazzoli and Guerriero* (1996), *Wigneron et al.* (2007a), *Liu et al.* (2011)), where τ is the optical depth and ω is the single-scattering albedo. τ was shown to be linked to the vegetation water content (VWC, kg m^{-2}) (*Kirdiashev et al.* (1979), *Mo et al.* (1982), *Jackson and Schmugge* (1991)) and to other vegetation properties such as the leaf area index (*Jackson and Schmugge* (1991), *Van de Griend and Wigneron* (2004), *Wigneron et al.* (2007a)). Therefore, τ is known as vegetation optical depth (VOD). VOD is also a function of the vegetation structure, which determines its dependence on the incidence angle and on the polarization of the radiation (*Ulaby and Wilson* (1985), *Wigneron et al.* (1995), *Hornbuckle et al.* (2003), *Schwank et al.* (2005)). Passive microwave radiometry is therefore a promising tool for monitoring vegetation on a global scale.

VOD samples the vegetation canopy, including woody vegetation, which uses root zone soil moisture (*Andela et al.* (2013)). VOD was used to study deforestation in South America (*Van Marle et al.* (2016)) and Africa (*Brandt et al.* (2018)). It has also been used to reveal teleconnections linking the state of the vegetation in Australia and El Niño Southern Oscillation (*Liu et al.* (2007)). In addition, *Liu et al.* (2015) showed the high potential of microwave VOD (X and C bands) to monitor the AGB dynamics on a large scale. Combining passive microwave VOD and optical data such as the NDVI, contributes to a more robust assessment of the vegetation characteristics (*Liu et al.* (2011)). The mentioned studies used VOD derived from different radiometers operating at different frequencies (*Liu et al.* (2011)) such as SSM/I at 19GHz (K-band), TRMM-TMI at 10.7 GHz (X-band) and the Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E) at 10.7 and 6.9 GHz (C-band). It is worth noting that the VOD is a radiative parameter, and so intrinsically depends on the frequency of the electromagnetic radiation. VOD retrieved at different frequencies provide complementary information and should not be rescaled as if they were a single quantity as done by some studies (*Liu et al.* (2015); *Moesinger et al.* (2020a)). It is recommended that a specific VOD dataset is noted as B-VOD, where B stands for the microwave band (X-VOD, C-VOD, etc.). In principle, the lower the frequency, the lower the VOD for a given level of VWC (*Wigneron et al.* (1995, 2004); *Ferrazzoli and Guerriero* (1996)). Consequently, L-band (1.4 GHz, 21 cm) observations, which are less attenuated through the vegetation canopy, are capable of sampling the vegetation layer up to higher biomass values compared to higher-frequency observations.
SMOS retrieval algorithm (*Kerr et al.* (2012a, 2020)) is based on the L-MEB (L-band Model Emission of the Biosphere) transfer radiative model to compute the emission of the surface and its interactions with a vegetation layer (through the Tau Omega model). The full-polarization and multi-angular capabilities of SMOS allow the simultaneous retrieval of the soil moisture content and L-VOD. *Lawrence et al.* (2014a); *Grant et al.* (2016) compared SMOS L-VOD to X-VOD and C-VOD measured by AMSR-E and to visible/infrared vegetation indices. In crop zones, such as the MODIS vegetation indices, L-VOD increases during the growing season and decreases during senescence (*Lawrence et al.* (2014a)). On a global scale, L-VOD is less correlated

to optical/visible vegetation indices than X/C-VOD, suggesting that L-VOD can add more complementary information with respect to optical/infrared indices than X/C-VOD (*Grant et al.* (2016)). For instance, *Rahmoune et al.* (2014) found a significant correlation between L-VOD and tree height estimates. *Vittucci et al.* (2016) also discussed this relationship and compared it to the one estimated with X/C-VOD, which shows higher values for low tree-height than SMOS L-VOD, as expected. *Vittucci et al.* (2016) also showed a close to linear relationship between L-VOD and AGB at 20 selected points over Peru, Columbia and Panama. During the SMOS+Veg ESA project, *Rodríguez-Fernández et al.* (2018) showed that SMOS L-VOD could be used as a proxy to Above Ground Biomass (AGB estimates), by showing its higher sensitivity to AGB than optical indices and VOD derived at higher passive microwave frequencies. This work was extended globally by *Mialon et al.* (2020). *Rodríguez-Fernández et al.* (2018) showed the possibility of computing simple L-VOD to AGB relationships. Using this approach, subsequent studies such as those of *Brandt et al.* (2018) and *Fan et al.* (2019) were able to study the evolution of carbon stocks for the last ten years. The contribution of passive radiometry to above-ground carbon stocks monitoring arises from the fact that despite a low spatial resolution, they give access to the temporal evolution thanks to a global coverage with high revisit.
Considering that L-VOD is sensitive to both biomass and its water content, L-VOD was averaged on a monthly and yearly basis to smooth out the daily variability. However, considering the duration of the SMOS mission, it is possible to study the trends of AGB over the fifteen years of existing data. This is of significant interest, knowing that current AGB maps derived from radar and optical data are computed as yearly estimates and are not available regularly.

Estimating AGB from the VOD is however not an easy task, as the assumptions to derive VOD and the uncertainties associated with the AGB(VOD) relationship imply large uncertainties in terms of derived AGB. The present project takes advantage of the neural network approach to directly estimate AGB from passive microwave observations. The VODs and AGB from VODs are used in this project to evaluate the strengths of the new proposed approach. Considering that optical indices are linked to the upper part of the green vegetation, which differs from what is observed at microwave frequencies, the complementarity of these indices is evaluated to fully characterize the vegetation layer.
Even if SMOS provides the longest consistent time series of L-band VOD, many applications would require a simultaneous use of VOD at other frequencies. However, currently existing multi-frequency VOD data sets (VODCA, *Moesinger et al.* (2020a), SMOS-IC, AMSR-E/2, *Wigneron et al.* (2021); *Wang et al.* (2021a)) have been derived using different algorithms, or extending algorithms specifically designed for one frequency such as the L-Band Microwave Emission of the Biosphere, to other frequencies. Some of them include a rescaling of VOD obtained from different frequencies using CDF-matching (*Moesinger et al.* (2020a)), ignoring that VOD is intrinsically frequency dependent. Current analysis of multifrequency VOD data is a complex task because it is not clear that the observed behavior arises from artifacts due to non-harmonized data sets or from a geophysical origin (see for instance some examples in Sect. 2.1.3.4 from the ESA Land Carbon Constellation study). Therefore, generating a vegetation data set based on low frequency microwave brightness temperatures is needed, and should cover the 1.4 to 37 GHz frequency range using a generic algorithm well adapted to all the frequencies.
The acquisition of vegetation parameter datasets generated from measurements in the optical domain and processed to a datacube consistent with the data cube containing the microwave data is also necessary. First, because there is a large amount of scientific literature studying vegetation with optical indices, and second, because the information contained is complementary to that in VOD, in particular for L-VOD. Using these microwave and optical data cubes, it is possible to investigate the mechanisms driving the VOD at different frequencies and their relationships with the hydric state of the vegetation. The synergies of multi-frequency VODs and optical indices, also allow us to to perform better estimations of the Above Ground Biomass.

Chapter 3

OSMOSE Project overview

3.1 OSMOSE project objectives

Considering the context described in 2, the OSMOSE has the following objectives, presented in Table 3.1 which also shows the work packages covering each objective.

Objective	Description	WP
010	Generate a vegetation data set based on low frequency microwave brightness temperatures covering the 1.4 to 37 GHz frequency range	1-4
015	Acquire and analyse a set of data records based on measurements obtained in the optical domain, preferably including Sentinel-3 OLCI and / or Proba-V measurements	6-9
020	Exploit the complementary information of the passive microwave and optical data sets. Spatial and temporal variabilities on the local scale, i.e. individual pixels, shall be included in the analyses. Mismatches in spatial support shall be discussed and minimized. Generate Above Ground Biomass from multi-frequency VOD data complemented by optical indices.	10 5
025	Provide a roadmap for a continuous generation of vegetation data products based on optical measurements and low frequency passive microwave measurements. Outline potential application areas for the synergistic use of the observations.	11

Figure 3.1: Objectives of the project and the corresponding work packages (WP)

3.2 OSMOSE project tasks

The OSMOSE project was organized into twelve work packages to address these objectives, detailed in the next chapters. Overviews of the organization and activities are displayed in Figures 3.2, 3.3 and 3.4. These WP are entitled:

- **WP 0**: is the management and the website
- **WP 1**: Theoretical Analysis of multi-frequency brightness temperatures and VOD
- **WP 2**: Multi-mission Multi-frequency Passive Microwave VOD Data Sets– Algorithm Design Definition and Development
- **WP 3**: Multi-mission Passive Microwave VOD Data Sets – Input Data Preparation
- **WP 4**: Multi-mission Passive Microwave VOD Data Sets – Data Set Generation and Analysis
- **WP 5**: Derivation of Above Ground Biomass
- **WP 6**: Optical Vegetation Phenology / Indices – Literature Review
- **WP 7**: Multi-mission Phenology / Indices Data Set – Pre-processing Algorithm Design Definition and Development
- **WP 8**: Multi-mission Vegetation Phenology / Indices Data Set – InputData Set Preparation
- **WP 9**: Multi-mission Optical Vegetation Data Set – Data Cube Generation and Analyses
- **WP 10**: Comparative Data Analysis
- **WP 11**: Roadmap for Operational / continuous Data Generation and Exploitation

3.3 Project organization, workflow

Matthias Drusch -ESA- is the technical officer of this project, led by Arnaud Mialon (CESBIO) with the support of Beatrice Berthelot (Magellium). The different tasks are linked to each other so that the outputs of the activities benefit the other tasks.

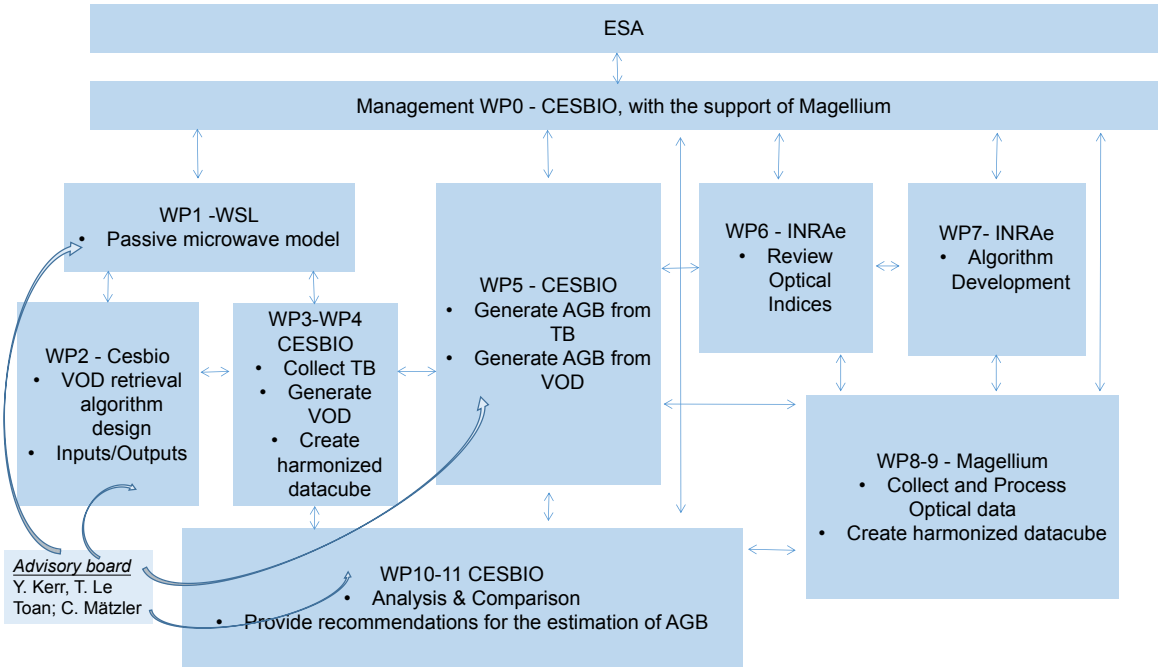


Figure 3.2: Overview of the projet.

Each task is divided into activities as shown in Figure 3.3.

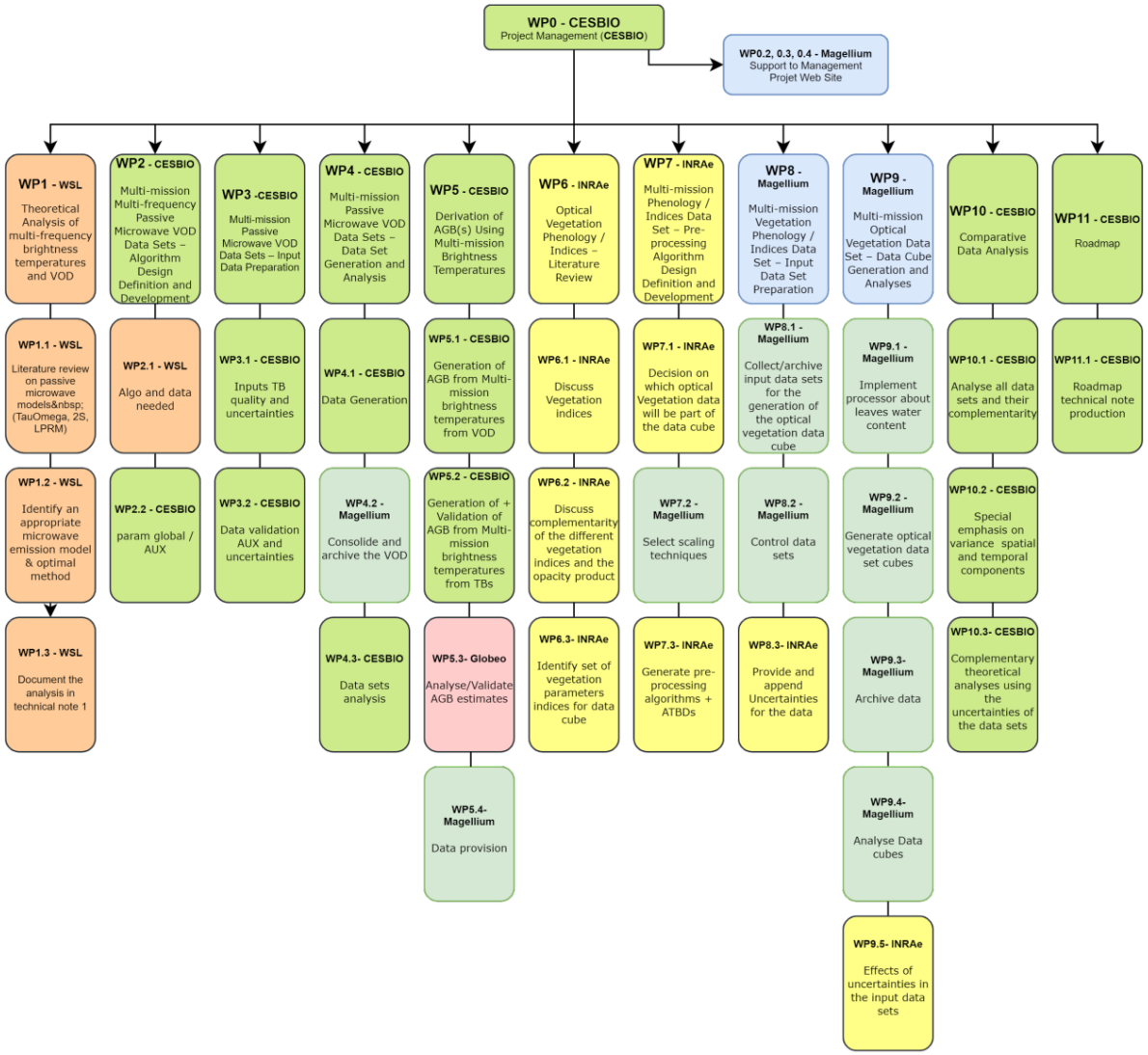


Figure 3.3: Organization of tasks.

3.4 Project schedule

The PM-VOS (OSMOSE) project was planned for two years starting in June 2022. Figure 3.4 shows the overall project schedule.

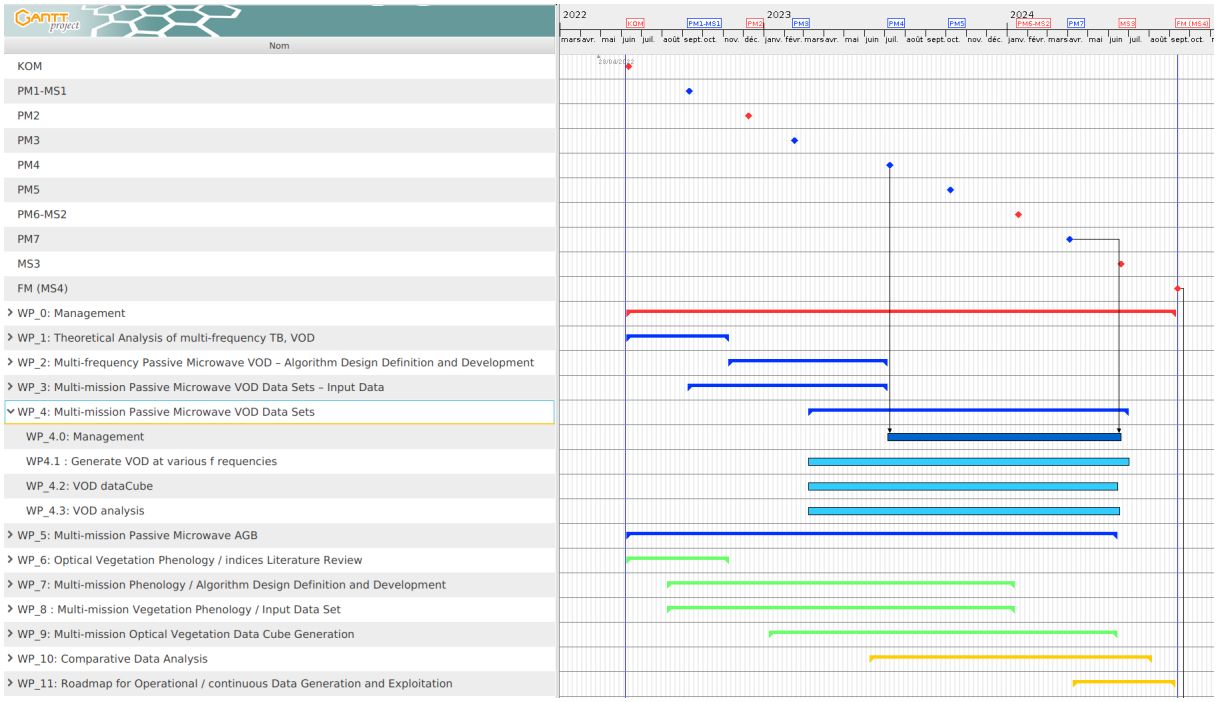


Figure 3.4: Gantt chart of the OSMOSE project, for the period June 2022-September 2024.

Chapter 4

Microwave Theoretical Analysis (WP 1)

4.1 Introduction

The primary aim of WP1 is to develop a physics-based microwave emission model (MEM) that elucidates the dependence of effective radiative transfer properties on frequency and vegetation state parameters. The developed model is named as the Microwave Emission Model for Layered Vegetation (MEMLV). It is applied to coniferous forest as an exemplary study, covering a wide frequency range from P- to Ka-band (0.4 GHz – 37 GHz). The simulation can be used to support and guide spaceborne multi-frequency VOD retrieval in WPs 2-4. This is achieved by enhancing the parameterizations of the retrieval algorithm and deepening the understanding of the physics principles behind VOD retrieval.

4.2 Methodology

Physics-based forward models describe the fundamental mechanisms of the interaction between electromagnetic waves and vegetation canopies *Fung and Chen* (2010); *Chauhan et al.* (1994); *Schwank et al.* (2021), providing valuable insights for VOD retrieval. The developed MEMLV is a compound model that consists of three major modelling components: (i) the single-layer discrete model (SL-DSM) for calculating the single scattering albedo and VOD of a single-layer of vegetation canopies; (ii) a new Tree Structure Model (TSM) for representing the forest architecture in multi-layer structure and defining the characteristics of tree constituents in each layer, based on the concept of radio realism; (iii) the Two-Stream microwave emission model (2S-MEM) for combining SL-DSM and TSM to calculate the brightness temperature (BT) of forested ground. Finally, based on the simulated BT, an optimization process is employed to obtain the effective scattering albedo, ω_{eff}, and the effective VOD, τ_{eff} of an equivalent single-layer canopy above ground. This process is a key to bridge the physics-based forward modelling with retrieval algorithms.

4.2.1 Single-layer Discrete Scatter Model

The discrete scatter model (DSM) *Lang and Sidhu* (1983); *Chauhan et al.* (1994); *Zhou et al.* (2020, 2023) employs a coherent approach to simulate the radiative transfer properties of canopies. It offers great flexibility in adopting various arrangements of vegetation constituents by representing vegetation as an effective medium with embedded discrete scatterers. The DSM has been employed in modelling the emission of a single layer canopy with below ground *Kurum et al.* (2012); *Kurum* (2013). However, this structure may not be sufficient to represent realistic coniferous forests.

In this WP, forest architectures are modeled using a multilayer structure to achieve a more realistic representation. Each layer contains tree constituents that have distinct number densities and size distributions. The single-layer DSM (SL-DSM) *Chauhan et al.* (1994); *Zhou et al.* (2020) (simplified DSM without considering the ground effect) is used to calculate the single scattering albedo and VOD of each layer.

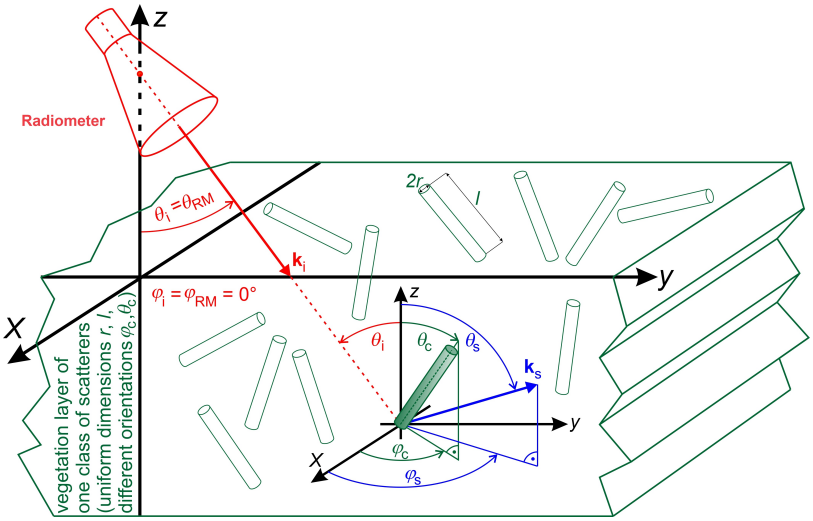


Figure 4.1: Sketch of the geometry and coordinate system used in the SL-DSM based on an example of a vegetation layer that includes only one single class of scatterers.

The essential part of the SL-DSM is the scatterer's scattering amplitude, F, which is defined as the ratio between the amplitude of the far-field scattered electric field and the amplitude of the incident electric field. Assuming a cylindrical scatterer with one end located at the origin of a Cartesian coordinate, a wave with an amplitude E_0^p and direction along the unit propagation vector $\mathbf{k}_\mathrm{i}$ is incident on the scatterer, as shown in Fig. 4.1. The electric field of the incident wave is:

$$E_\mathrm{i}^p(r) = E_0^p \exp(-\mathrm{j}k_0 \mathbf{k}_\mathrm{i} \cdot \mathbf{r}), \tag{4.1}$$

where $k_0 = 2\pi/\lambda$ represents the free-space wavenumber; λ is wavelength; $\mathbf{r}$ is the vector from the origin to the field point. The variables p and q denote the orthogonal polarizations of the incident and scattered waves, respectively, which can be either horizontal (H) or vertical (V).

The incident wave is then scattered in all directions with different amplitudes. Along a specific direction $\mathbf{k}_\mathrm{s}$, the scattered electric field can be written as *Seker and Schneider* (1988):

$$E_\mathrm{s}^q(\mathbf{r}) = F^{(p,q)}(\mathbf{k}_\mathrm{s}, \mathbf{k}_\mathrm{i}) \cdot E_0^p \frac{\exp(-\mathrm{j}k_0 \mathbf{k}_\mathrm{s} \cdot \mathbf{r})}{|\mathbf{r}|}, \tag{4.2}$$

where $F^{(p,q)}(\mathbf{k}_\mathrm{s}, \mathbf{k}_\mathrm{i})$ is the scattering amplitude specified by p, q, $\mathbf{k}_\mathrm{s}$, and $\mathbf{k}_\mathrm{i}$. If $\mathbf{k}_\mathrm{i}$ and $\mathbf{k}_\mathrm{s}$ are respectively defined by the associated zenith and azimuthal angles $(\theta_\mathrm{i}, \phi_\mathrm{i})$ and $(\theta_\mathrm{s}, \phi_\mathrm{s})$, and the orientation of an individual cylindrical scatterer is defined by $(\theta_\mathrm{c}, \phi_\mathrm{c})$, then its scattering amplitude $F^{(p,q)}(\mathbf{k}_\mathrm{s}, \mathbf{k}_\mathrm{i}, f)$ at frequency f can be expressed as $F^{(p,q)}(\theta_\mathrm{s}, \phi_\mathrm{s}; \theta_\mathrm{i}, \phi_\mathrm{i}; \theta_\mathrm{c}, \phi_\mathrm{c}; f)$. Fig. 4.1 depicts the geometry of a cloud of a specific class of scatterers in a layer. Hereinafter, a "class" of scatterers is defined for scatterers having the same dimensions (radius and length) and the same orientational probability density functions (PDFs) in a medium.

The directional patterns of the scattering amplitude are further investigated for a cylinder having radius $r = 1$ cm, length $l = 26$ cm, $m_\mathrm{d} = 0.5$ kg·kg^{-1}, $S = 5$ ppt, $T = 20°$C, as an example. The cylinder is vertically oriented, and the incident wave is along the direction $\{\theta_\mathrm{i}, \phi_\mathrm{i}\} = \{40°, 0°\}$. The polarizations of the incident and scattered waves are assumed as $\{p, q\} = \{\mathrm{H},\mathrm{H}\}$, respectively. Fig. 4.2 shows the 3D patterns of the scattering amplitudes at L-, C-, X- and Ka-band together with their projections. The brown vertical lines represent the cylinders, the red and blue arrows represent the directions of the incidence and forward scattering, respectively. The magnitude of the forward scattering amplitude, $|FF|$, is also given in the figure as a quantitative reference for the scale of plots. It is seen that at L-band, the specular- and forward scattering have comparable amplitudes. Small lobes are also present in the directions of the back- and bistatic scattering. At C-band, the forward scattering amplitude FF becomes greater, and it dominates the patterns at higher frequencies (X- and Ka-band). This trend is due to the transition in scattering mechanism from Rayleigh scattering to Mie scattering as frequency increases *Robitaille* (2007).

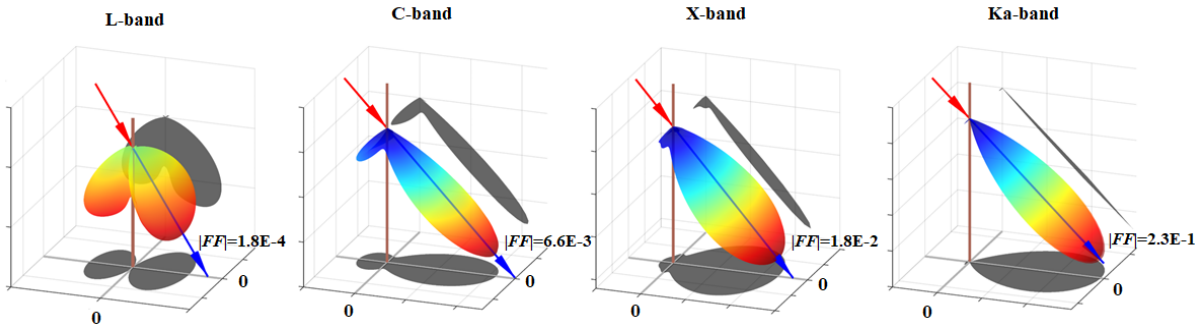


Figure 4.2: Three-dimensional spatial patterns of scattering amplitudes $F^{(H,H)}$ simulated for vertically oriented cylinders ($\theta_\mathsf{c} = 0°$) illuminated by a wave incident at $\theta_\mathsf{i} = 40°$, $\phi_\mathsf{i} = 0°$. From left to right: L-band ($f = 1.4$ GHz), C-band ($f = 6.8$ GHz), X-band ($f = 11$ GHz), and Ka-band ($f = 37$ GHz). $|FF|$ indicates the absolute values of forward scattering amplitudes.

Assuming the zenith and azimuthal orientational PDFs are P_{θ_c} and P_{ϕ_c} for a class of scatterers in the medium, respectively, the averaged scattering amplitude for this class of scatterers is:

$$\langle F^{(p,q)}(\theta_\mathsf{s}, \phi_\mathsf{s}; \theta_\mathsf{i}, \phi_\mathsf{i}; f)\rangle = \int_{\theta_\mathsf{c}=0}^{\pi} \int_{\phi_\mathsf{c}=0}^{2\pi} F^{(p,q)}(\theta_\mathsf{s}, \phi_\mathsf{s}; \theta_\mathsf{i}, \phi_\mathsf{i}; \theta_\mathsf{c}, \phi_\mathsf{c}; f) P_{\theta_\mathsf{c}} P_{\phi_\mathsf{c}}\, d\theta_\mathsf{c}\, d\phi_\mathsf{c}, \tag{4.3}$$

where angular brackets $\langle\cdot\rangle$ denote ensemble average over the angular statistics for a certain class of scatterers. Consequently, the averaged scattering cross-section $\hat{\kappa}_\mathsf{s}^p$ and extinction cross-section $\hat{\kappa}_\mathsf{e}^p$ for this class of scatterers are *Zhou et al.* (2023); *Kurum* (2013):

$$\hat{\kappa}_\mathsf{s}^p(f) = \sum_{q=\{\mathsf{H},\mathsf{V}\}} \int_{\theta_\mathsf{s}=0}^{\pi} \int_{\phi_\mathsf{s}=0}^{2\pi} \langle |F^{(p,q)}(\theta_\mathsf{s}, \phi_\mathsf{s}; \theta_\mathsf{i}, \phi_\mathsf{i}; f)|^2 \rangle \sin\theta_\mathsf{s}\, d\theta_\mathsf{s}\, d\phi_\mathsf{s} \tag{4.4}$$

and

$$\hat{\kappa}_\mathsf{e}^p(f) = \frac{4\pi}{k_0} \cdot \mathrm{Im}\big\{\langle FF^{(p,p)}\rangle\big\} = \frac{4\pi}{k_0} \cdot \mathrm{Im}\big\{\langle F^{(p,p)}(\theta_{\mathsf{s,f}}, \phi_{\mathsf{s,f}}; \theta_\mathsf{i}, \phi_\mathsf{i}; f)\rangle\big\}, \tag{4.5}$$

where $FF^{(p,p)}$ is the forward scattering amplitude along the forward direction defined by $\theta_{\mathsf{s,f}} = \pi - \theta_\mathsf{i}$ and $\phi_{\mathsf{s,f}} = \pi + \phi_\mathsf{i}$. Hereinafter, the symbol $\hat{\cdot}$ is used for the radiative transfer properties of a class of scatterers with the same dimensions and orientational PDFs.

In general, the fundamental relationships shown in (4.3) – (4.5) can be applied to scatterers having any arbitrary shapes. Furthermore, these equations assume the direction of the incident wave is fixed while the scattered wave is in all directions. In passive radiometry, however, the waves incident on a scatterer are from all directions, while a radiometer attached to an ideal pencil beam antenna only senses the scattered wave in one fixed zenith direction, denoted by θ_{RM}. These two conditions are essentially the same, as can be explained by the reciprocity of Maxwell's equations. Therefore, it is inherent that $\theta_\mathsf{i} = \theta_{\mathsf{RM}}$, as shown in Fig. 4.1.

To simulate a realistic vegetation layer, a composite of scatterer classes representing different vegetation constituents must be considered. Assuming the canopy consists of $i = 1, \ldots, M$ classes of cylindrical scatterers with different radius r_i and length l_i and orientational PDFs, the scattering- and extinction coefficients $\kappa_\mathsf{s}^p, \kappa_\mathsf{e}^p$ of the resultant canopy layer can be obtained by *Zhou et al.* (2023):

$$\kappa_\mathsf{s}^p = \sum_{i=1}^{M} \rho_i \cdot \hat{\kappa}_{\mathsf{s},i}^p \quad \text{and} \quad \kappa_\mathsf{e}^p = \sum_{i=1}^{M} \rho_i \cdot \hat{\kappa}_{\mathsf{e},i}^p, \tag{4.6}$$

where ρ_i is the number density of the i-th class of vegetation constituents presented in the layer, and $\hat{\kappa}_{\mathsf{s},i}^p$, $\hat{\kappa}_{\mathsf{e},i}^p$ are its associated averaged scattering- and extinction cross-sections.

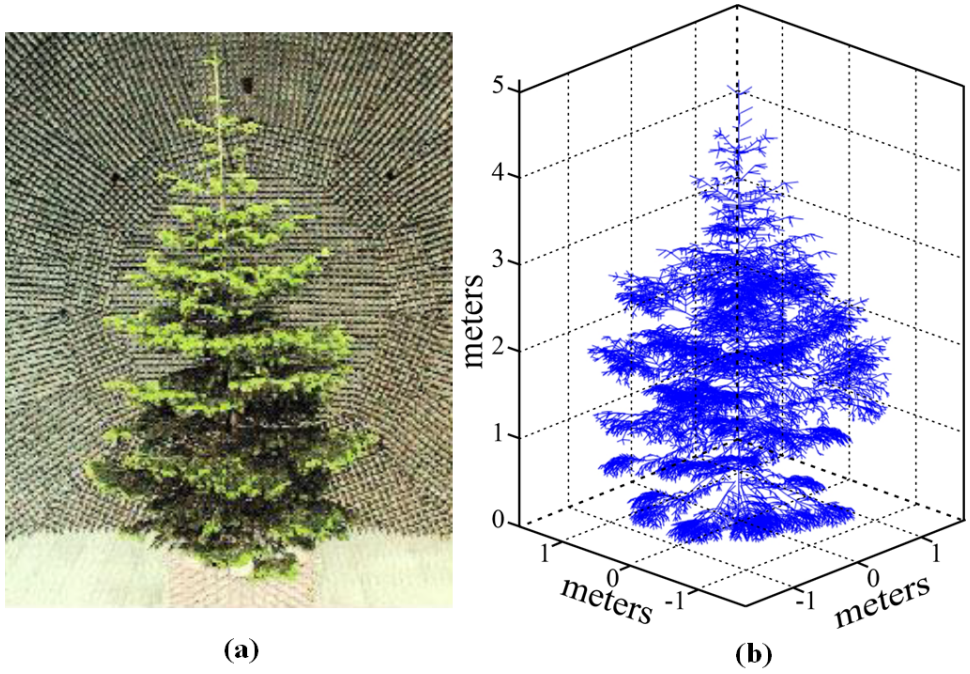


Figure 4.3: (a) Photo of the JRC tree in the anechoic chamber *Fortuny and Sieber* (1999). (b) Vectorization of the tree architecture *Zhao* (2013)

Finally, the single scattering albedo (ω^p) and the VOD (τ^p) of a single layer with a thickness of d, composed of composite scatterers, are given by *Kurum* (2013):

$$\omega^p = \frac{\kappa_{\mathsf{s}}^p}{\kappa_{\mathsf{e}}^p} \quad \text{and} \quad \tau^p = d \cdot \kappa_{\mathsf{e}}^p, \tag{4.7}$$

The scattering and absorption behavior of discrete scatterers over frequencies forms one of the fundamental elements of the MEMLV. As natural canopies consist of scatterers with different sizes, number densities and orientations PDFs, it is important to have a realistic tree structure to support the modelling of VOD and scattering albedo for coniferous forests. This will be discussed in the next section.

4.2.2 Tree Structure Model (TSM)

The development and parameterization of the TSM are inspired by an experiment conducted for a 5m-high Caucasian fir tree at the Joint Research Centre (JRC), Ispra, Italy *Lang et al.* (1998); *Fortuny and Sieber* (1999). The sample tree, referred to as the JRC tree, was placed in an anechoic chamber for radar measurements, and the photograph of the tree in the chamber *Fortuny and Sieber* (1999) is presented in Fig. 4.3(a). The structure of the JRC tree was measured, and a specific tree vectorization technique *Landry et al.* (1997) was employed to reconstruct the tree architecture, including the dimensions and numbers of the tree constituents, as shown in Fig. 4.3(b) *Zhao* (2013). Furthermore, the dielectric constants of the JRC tree constituents were measured and documented in *Zhao* (2013). These experimental data for the JRC tree are used as guidelines for parameterizing the TSM presented in the next section.

The TSM allows to discretely describe a variety of tree architectures based on simple equations. It assumes an individual tree is rotationally symmetric, and the discretization involves three aspects:

(i) A tree is stratified by $j = 0, 1, \ldots, N$ horizontal layers from the bottom to the top. If the height of the lower bound of tree crowns at which the first branch emerges from the trunk is H_{low}, and each layer in tree crowns has a uniform thickness of d, then the mean height above the ground of the j-th layer, h_j, can be represented as:

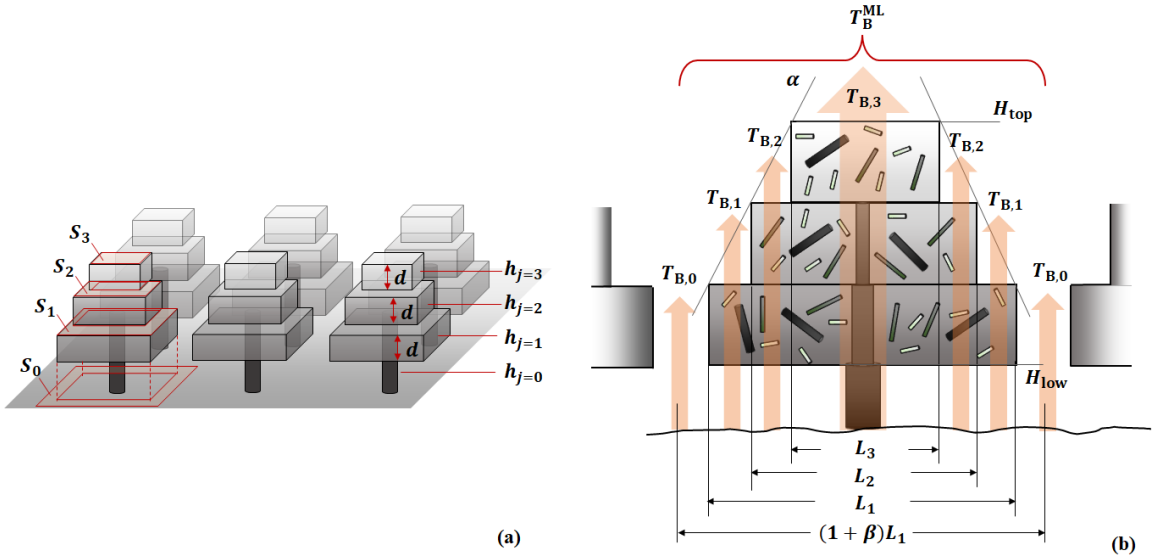


Figure 4.4: Illustration of the TSM. (a) Panoramic forest structure with marked layer heights h_j, thickness d and surface area S_j (b) Side view of a tree with marked macro-parameters H_{low}, H_{top}, α, β, L_1, together with $T_{\text{B},j}$ emitted from stacks of j layers and the total upwelling T_{B}^{ML}.

$$h_j = \begin{cases} \frac{H_{\text{low}}}{2}, & \text{for } j = 0 \\ H_{\text{low}} + \frac{(j-1)}{2} \cdot d, & \text{for } j = 1, 2, \ldots, N \end{cases} \tag{4.8}$$

where $h_{j=0}$ is the mean height of the trunk section that is below the tree crown.

(ii) The dimensions of tree constituents are discretized by categorizing them into a limited number of classes, each represented by cylinders with specified radii r_i and lengths l_i $(i = 1, 2, \ldots, M)$. The index i ascends as the size of the scatterer decreases.

(iii) The tree constituents are categorized into three types: trunk, branches and needles. Each type can comprise multiple classes of cylinders. The numbers of cylinder classes for trunk, branches, and needles are denoted as M_{t}, M_{b}, and M_{n}, respectively.

According to these discretization rules, four major components of the TSM are introduced below:

(i)**Tree macro structure:** The cross-sectional area of tree crowns is assumed to reduce as h_j increases based on the structure of the JRC tree and observations of most natural trees in coniferous forests. Fig. 4.4 illustrates an exemplary three-layer tree crown structure, where each crown layer is represented by a square cuboid. The edge lengths of these cuboids are assumed to decrease linearly with a crown-shape parameter, α, as h_j increases. If the edge length of the bottom cuboid is denoted as L_1, then the edge length of the j-th layer is given by:

$$L_j = L_1 - \alpha \cdot (h_j - h_1) \tag{4.9}$$

The TSM also includes the trunk layer, $j = 0$, located below the bottom layer of the crown, as illustrated in Fig. 4.4. In addition, the trunk is assumed to be also present within the tree crown layers, and its effect is simulated alongside branches and needles.

To realistically simulate the radiative transfer properties of coniferous forests, the canopy gaps (i.e. the area between trees without tree crown coverage) have been considered. The coefficient β is introduced as the canopy gap fraction, representing the average distance between the centers of neighboring trees in forests as $(\beta + 1) \cdot L_1$, as shown in Fig. 4.4(b).

The parameters α, β, and L_1 characterize the macro vertical structure and spatial arrangement of multi-layer trees in forests to model the upwelling BT, $T_{\mathrm{B}}^{\mathrm{ML}}$. As illustrated in Fig. 4.4, $T_{\mathrm{B}}^{\mathrm{ML}}$ is simulated as the weighted sum of BT emitted from the stack of j canopy layers above ground, $T_{\mathrm{B},j}$. Therefore, $T_{\mathrm{B}}^{\mathrm{ML}}$ can be written as:

$$T_{\mathrm{B}}^{\mathrm{ML}} = \frac{\sum_{j=0}^{N} T_{\mathrm{B},j}\, S_j}{L_1^2(1+\beta)^2}, \tag{4.10}$$

with

$$S_j = \begin{cases} L_1^2(1+\beta)^2 - L_1^2, & j = 0 \\ L_j^2 - L_{j+1}^2, & 0 < j < N \\ L_N^2, & j = N \end{cases} \tag{4.11}$$

where $S_j/[L_1^2(1+\beta)^2]$ represents the surface area fraction of the stack of j canopy layers. Specifically, S_0 is the surface area of canopy gaps, as shown in Fig. 4.4(a).

(ii) **Trunks**: The bottom section of the trunk (cylinder class $i = 1$) has the greatest radius and is considered as the sole constituent in layer $j = 0$. Starting from the first layer of the tree crown ($j = 1$), the trunk section has decreasing radii as h_j increases, mixing with branches and needles. On the uppermost crown layer ($j = N$), it is assumed that the trunk is not present. Therefore, there are a total of $M_{\mathrm{t}} = N$ classes of trunk sections in the forest canopy, and their radii can be written as:

$$r_i = r_1 - A \cdot (i-1) \cdot d, \quad \text{with } i = 1, \ldots, N, \tag{4.12}$$

where A is a parameter adjusting the changing rate of the trunk radius as the aboveground height h_j increases. The trunk sections are assumed to be mostly vertical.

(iii) **Branches**: The cylinder class index for branches is $i = N+1, \ldots, N+M_{\mathrm{b}}$ since $i = 1, \ldots, N$ accounts for the trunk. It is assumed that both the mean radius, $\bar{r}(h_j)$, and the radius-distribution PDF, $P_r(r, \bar{r}(h_j))$, of branches are dependent on h_j. Since $P_r(r, \bar{r}(h_j))$ is a PDF, the normalization condition $\int_0^{+\infty} P_r(r, \bar{r}(h_j))\, dr = 1$ must hold to ensure the probability of finding a branch with any possible radius is one (100%) at a given h_j. As often used in the statistical representation of randomly distributed positive quantities, a log-normal function is used to represent $P_r(r, \bar{r}(h_j))$:

$$P_r(r, \bar{r}(h_j)) = \frac{1}{r \cdot \sigma(h_j) \cdot \sqrt{2\pi}} \cdot \exp\left\{ -\frac{1}{2} \left(\frac{\ln(r/\bar{r}(h_j))}{\sigma(h_j)} \right)^2 \right\}, \tag{4.13}$$

where the height-dependent parameter, $\sigma(h_j)$, determines the shape of $P_r(r, \bar{r}(h_j))$, allowing for different radius distributions of branches at different h_j. As a log-normal function, $P_r(r, \bar{r}(h_j))$ peaks near $r = \bar{r}(h_j)$, takes the value zero at $r = 0\,\mathrm{cm}$, and converges to zero as $r \to \infty$. The smaller $\sigma(h_j)$ is, the steeper the peak of $P_r(r, \bar{r}(h_j))$ is, indicating an increased likelihood of the presence of a branch with $r \approx \bar{r}(h_j)$. For the i^{th} class of cylindrical scatterers with a radius of r_i, its number density $\rho_i(h_j)$ at height h_j can be obtained by integrating $P_r(r, \bar{r}(h_j))$ over the radius range $r_i - \Delta r_i \le r \le r_i + \Delta r_i$:

$$\rho_i(h_j) = B \cdot \int_{r_i - \Delta r_i}^{r_i + \Delta r_i} P_r(r, \bar{r}(h_j))\, dr, \tag{4.14}$$

where $i = N+1, \ldots, N+M_{\mathrm{b}}$, B is a parameter used to adjust the number density of all branches, and Δr_i is the half-bin width of the integration range.

For natural coniferous trees, as h_j increases, the crown is more likely to be dominated by small branches. Therefore, $\bar{r}(h_j)$ and $\sigma(h_j)$ are represented by functions that decrease with increasing h_j:

$$\bar{r}(h_j) = C - D \cdot h_j^{n_0}, \tag{4.15}$$

with

$$C = \frac{H_{\text{top}}^{n_0} \cdot R_{\text{low}} - H_{\text{low}}^{n_0} \cdot R_{\text{top}}}{H_{\text{top}}^{n_0} - H_{\text{low}}^{n_0}},$$

$$D = \frac{R_{\text{low}} - R_{\text{top}}}{H_{\text{top}}^{n_0} - H_{\text{low}}^{n_0}},$$

and

$$\sigma(h_j) = E - F \cdot \frac{h_j - H_{\text{low}}}{H_{\text{top}} - H_{\text{low}}}, \tag{4.16}$$

where R_{low} and R_{top} denote the mean radius of branches present in the lower-bound, H_{low}, and upper-bound, H_{top}, of the tree crown, respectively. The parameter n_0 allows the TSM to describe various patterns of how the mean radius $\bar{r}(h_j)$ of branches decreases with increasing h_j. In (4.16), E and F are parameters that adjust how $\sigma(h_j)$ changes with h_j.

The TSM also includes the statistical description of the branches' orientations expressed by the zenith and azimuthal angles $\theta_{\text{c,b}}$ and $\phi_{\text{c,b}}$ (Fig. 4.1). Similar to other constituents, branches are assumed to have an isotropic distribution $P_{\phi_{\text{c,b}}}$ in azimuth:

$$P_{\phi_{\text{c,b}}} = \frac{1}{2\pi} \quad \text{for} \quad 0 \leq \phi_{\text{c,b}} \leq 2\pi \tag{4.17}$$

The zenith orientation PDF, $P_{\theta_{\text{c,b}}}$, is assumed to be uniform within a limited range of zenith angles θ_{c}. Furthermore, it is assumed that the smaller the branch is, the wider the range of its zenith orientation. This reflects that larger branches are more likely to orient along a particular zenith direction, whereas smaller branches tend to have more random zenith orientations. Therefore, $P_{\theta_{\text{c,b}}}$ is represented by:

$$P_{\theta_{\text{c,b}}} = \begin{cases} (\theta_{\text{b-max}} - \theta_{\text{b-min}})^{-1}, & \text{if } \theta_{\text{b-min}} \leq \theta_{\text{c,b}} \leq \theta_{\text{b-max}} \\ 0, & \text{otherwise} \end{cases} \tag{4.18}$$

with $\theta_{\text{b-min}} = \pi/2 - \pi/2 \cdot (G \cdot r + 1)^{-1}$ and $\theta_{\text{b-max}} = \pi/2 + \pi/2 \cdot (G \cdot r + 1)^{-1}$, where G is the parameter to adjust the shape of the radius-dependence in $P_{\theta_{\text{c,b}}}$. For cylinders with $r \to 0\,\text{cm}$, the limits of their zenith angle approach $\theta_{\text{b-min}} = 0$ and $\theta_{\text{b-max}} = \pi$, corresponding to isotropic zenith orientation.

(iv) **Needles**: Since the needles of mature coniferous trees have relatively constant dimensions, they are represented by a single class of cylinders. This results in their cylinder class index $i = M$. The orientational distribution of the needles is considered as isotropic. The number density of needles, $\rho_{i=M}$, is strongly associated with the number density of branches. For pine trees, often the needles only grow on small branches. Therefore, the branches are further classified into small branches and large branches. If the total number of classes for small branches is $M_{\text{b,s}}$, the number density of needles can be represented as:

$$\rho_M(h_j) = \sum_{i=M-1-M_{\text{b,s}}}^{M-1} \left(K \cdot \rho_i(h_j) \cdot l_i\right), \tag{4.19}$$

where K represents the average number of needles growing per unit length on small branches.

Equations (4.8) to (4.19) define the TSM following the concept of radio-realism. It employs a minimal number of observable parameters to describe a coniferous forest from the perspective of its macro-geometric structure to its fine distribution of tree constituents. Although here the TSM is developed and used for coniferous forests, its formulation makes it flexible to adapt to other types of forests and vegetation.

The TSM is parameterized such that resulting numbers of tree constituents closely match the in-situ measurements of the JRC tree (Section 4.2.2). The JRC tree has a height of ∼5 m and a diameter of ∼3.4 m *Zhao* (2013). To achieve a similar crown shape, the TSM is parameterized by $H_{\text{top}} = 5.5$ m, $H_{\text{low}} = 0.5$ m, $L_1 = 3.4$ m, and $\alpha = 0.6$. The number of layers and their thicknesses are assumed as $N = 5$ and $d = 1$ m, respectively. With these macro parameters, the volume of each layer can be calculated.

Table 4.1: Comparison of the cylinder classes $i = 1, 2, \ldots, M = 15$ categorized in TSM with corresponding JRC-Tree experiments

Cylinder classes		**r (cm)**	**l (cm)**	**number** W_i	**r - JRC (cm)**	**number JRC**
Trunk	i=1	9	100	1	9 (DBH) 3 (avg)	1
	i=2	7				
	i=3	5				
	i=4	3				
	i=5	1				
Large branch	i=6	1.0	30	19	0.9∼1.0	23
	i=7	0.8	26	52	0.7∼0.9	63
	i=8	0.6	22	170	0.5∼0.7	173
	i=9	0.45	19	231	0.4∼0.5	165
Small Branch	i=10	0.35	16	540	0.3∼0.4	621
	i=11	0.25	13	1526	0.2∼0.3	1199
	i=12	0.17	10	2484	0.1∼0.2	4854
	i=13	0.12	8	3620		
	i=14	0.07	6	9786	0∼0.1	11760
Needle	i=15	0.03	2.5	2.8E6	0.02	2.8E6

For trunk sections, it is assumed their radii $r_i = \{9, 7, 5, 3, 1\}$ cm for $i = \{1, 2, 3, 4, 5\}$, associated with layers $j = \{0, 1, 2, 3, 4\}$ counted from bottom to top, as shown in Table 4.1. The maximal radius, $r_1 = 9$ cm, is the same as the diameter-at-breast-height (DBH) of the JRC tree, while the average of r_i for trunk sections is slightly larger than the average trunk radius of the JRC tree *Zhao* (2013). The trunk sections are assumed to be mostly vertical with a uniform orientation PDF within the range $0° \leq \theta_c \leq 5°$ and $0° \leq \phi_c \leq 360°$.

The branches of the JRC tree generally have radii from 0 to 1 cm. Accordingly, the TSM represents branches by nine different classes ($i = 6, \ldots, 14$) of cylinders with radii r_i from 0.07 cm to 1 cm, as shown in Table 4.1. Furthermore, branches with $r_i \geq 0.45$ cm are considered as large branches ($i = 6, \ldots, 9$), while the rest are small branches ($i = 10, \ldots, 14$).

According to the parameterization given in Appendix A, the number densities of these classes of branches are computed for each layer using (4.13) and (4.14). Based on the layer volumes, the total number of the i^{th} class of scatterers, W_i, present in the entire tree can be obtained. The resulting W_i are in good agreement with the in-situ measurements of the JRC tree, as shown in Table 4.1.

The in-situ measurements for the JRC tree also included branch lengths *Zhao* (2013). The results showed that the branch lengths in general range from 0 to 30 cm and are inversely proportional to the number of branches, following a similar distribution of the radius, see Table 4.1. However, the reported lengths and radii of the branches were decoupled, meaning for a branch with a specific radius r_i, its corresponding average length l_i was not documented. Thus, the TSM assumes l_i in accordance with the elastic similarity hypothesis *McMahon and Kronauer* (1976):

$$l_i = Q \cdot r_i^{2/3} \quad \text{with} \quad i = 6, 7, \ldots, 14. \tag{4.20}$$

The parameter Q is found to be approximately 33 to confine l_i within the documented range of branch lengths (0 – 30 cm). The corresponding l_i and r_i for branches are given in Table 4.1.

For needles, the TSM is parameterized such that the total number of needles $W_{i=15}$ matches the corresponding count 2.8×10^6 available from the in-situ measurements of the JRC tree (Table 4.1). The radius and length of needles are chosen as $r_{i=15} = 0.03$ cm and $l_{i=15} = 2.5$ cm based on *Fortuny and Sieber* (1999), which is confirmed by our own measurements made for a pine tree.

The dielectric constant of tree constituents, $\varepsilon_c = \varepsilon_c' - \mathrm{j} \cdot \varepsilon_c''$, is modelled based on *Matzler* (1994):

$$\varepsilon_c(S, T, m_d, f) = 0.522 \cdot (1 - 1.32 \cdot m_d) \cdot \varepsilon_{sw}(S, T, f) + 0.51 + 3.84 \cdot m_d, \tag{4.21}$$

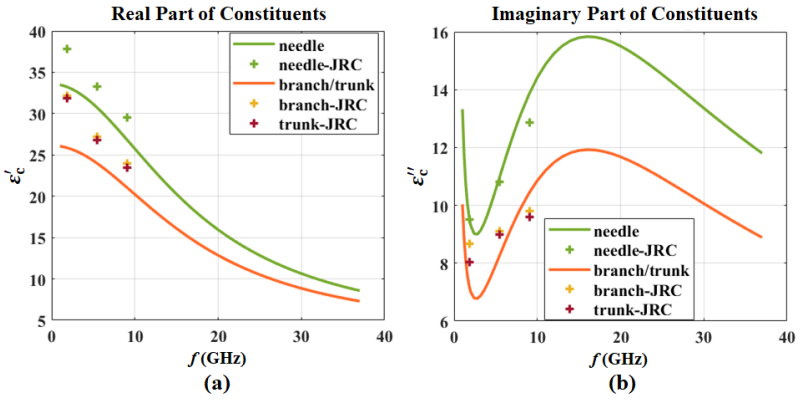


Figure 4.5: Comparison between simulated frequency dependence $\varepsilon_{\mathsf{c}}(f) = \varepsilon'_{\mathsf{c}}(f) - \mathrm{j} \cdot \varepsilon''_{\mathsf{c}}(f)$ of different types of tree constituents and corresponding permittivities measured at $f = \{1.9, 5.5, 9.1\}$ GHz for the JRC tree.

where m_{d} (kg · kg^{-1}) is the dry-matter fraction and $\varepsilon_{\mathsf{sw}}$ is the dielectric constant of sap water. It is assumed that $\varepsilon_{\mathsf{sw}}$ can be represented by the dielectric constant of saline water, which is obtained by the Meissner-Wentz model *Meissner et al.* (2014) as a function of salinity S, temperature T and frequency f.

Equation (4.21) is validated with the dielectric measurements made for the JRC tree at L- (1.9 GHz), C- (5.5 GHz), and X- (9.1 GHz) bands (Table 4.2 of *Zhao* (2013)). The m_{d} and S are adjusted (with $T = 20°$C) to match the modeled ε_{c} with the measurements. This yields $m_{\mathsf{d}} = 0.15\,\mathrm{kg} \cdot \mathrm{kg}^{-1}$ and $S = 10$ ppt for needles; $m_{\mathsf{d}} = 0.30\,\mathrm{kg}{\cdot}\mathrm{kg}^{-1}$ and $S = 10$ ppt for branches and trunks. Figure 4.5 demonstrates good agreement between the modelled ε_{c} and measured dielectric constants for the needles, branches, and trunks at the frequencies of the measurements.

4.2.3 Multi-layer Two Stream Model

According to the SL-DSM and TSM introduced in Section 4.2, the cross-sections $\hat{\kappa}^p_{\mathsf{s},i}$, $\hat{\kappa}^p_{\mathsf{e},i}$ and number densities ρ_i of the $i^{\mathsf{th}}(i = 1, 2, \ldots, M)$ scatterer in the $j^{\mathsf{th}}(j = 1, 2, \ldots, N)$ layer can be obtained. Then the VOD and single scattering albedo for the j^{th} layer, $\tau^p_j(\theta_{\mathsf{RM}}, f)$ and $\omega^p_j(\theta_{\mathsf{RM}}, f)$, are simulated for two polarizations $p = \{\mathsf{H},\mathsf{V}\}$, four observation angles $\theta_{\mathsf{RM}} = \{0°, 20°, 40°, 60°\}$, and frequencies over $0.4\,\mathrm{GHz} \leq f \leq 37\,\mathrm{GHz}$. Then they are fed into the ML-2S-MEM to simulate upwelling BT of multi-layer coniferous forests, $T_{\mathsf{B}}^{p,\mathsf{ML}}(\theta_{\mathsf{RM}}, f)$, with the following primary input parameters:

(i) Transmissivities $t^p_j(\theta_{\mathsf{RM}}, f)$ and volume reflectivities $r^p_j(\theta_{\mathsf{RM}}, f)$ of vegetation layers $j = 0, \ldots, N$. They are computed from the aforementioned ω^p_j and τ^p_j using equation (18) in *Schwank et al.* (2018).

(ii) Interface reflectivities s^p_j between vegetation layers. They are assumed as $s^p_j = 0$, meaning that the reflection between vegetation layers are neglected. In addition, the change of the wave propagation direction due to refraction at these interfaces are also neglected. These assumptions can be made without restrictions because tree crowns inherently form a continuous medium.

(iii) Interface reflectivity $s^p_{\mathsf{G}}(\theta_{\mathsf{RM}}, f)$ of the ground. It is computed from the effective ground dielectric constant ε_{G} using Fresnel's equations with surface roughness correction. The dielectric mixing model *Mironov et al.* (2009) is used to express ε_{G} by means of clay fraction (set to 0.16), ground temperature (set to $T_{\mathsf{G}} = 293\,\mathrm{K}$), and volumetric soil moisture (set to $SM = 0.3\,\mathrm{m}^3 \cdot \mathrm{m}^{-3}$). The HQN model *Wang and Choudhury* (1981) is used for roughness correction with the same empirical parameters $\{h_{\mathsf{G}}, q_{\mathsf{G}}, n^{\mathsf{V}}_{\mathsf{G}}, n^{\mathsf{H}}_{\mathsf{G}}\} = \{0.3, 0, 0, 2\}$ as used in SMOS retrievals over forested areas *Wigneron et al.* (2017).

(iv) Down-welling sky radiance $T_{\mathsf{sky}}(\theta_{\mathsf{RM}})$. It is estimated with the empirical approach *Pellarin et al.* (2003) that requires the observation angle θ_{RM}, air temperature 2-m above ground (set to $T_{\mathsf{air}} = 293\,\mathrm{K}$), and elevation above sea level set to $0.4\,\mathrm{km}$). Although the sky radiance model was developed for L-band ($f = 1.4\,\mathrm{GHz}$), it is applied in this study to estimate T_{sky} for microwave frequencies up to Ka-band ($f = 37\,\mathrm{GHz}$). The error introduced by this assumption is minimal as the reflection at the upper bound of the canopy is generally small.

(v) Temperatures of the vegetation layers j. They are assumed as constant $T_j = 293\,\mathrm{K}$.

With the aforementioned input parameters, the BT emitted from the stack of j canopy layers above ground, $T^p_{\mathrm{B},j}(\theta_{\mathrm{RM}}, f)$, can be computed. Using (4.10) and (4.11), the $T_{\mathrm{B}}^{p,\mathrm{ML}}(\theta_{\mathrm{RM}}, f)$ can be obtained by summing $T^p_{\mathrm{B},j}(\theta_{\mathrm{RM}}, f)$ weighted by the corresponding surface area fraction.

4.2.4 Optimization

Due to the large number of input parameters in the ML-2S-MEM, it is difficult to implement it directly in retrieval algorithms. The previous study *Li et al.* (2020) has demonstrated the feasibility of VOD retrieval using the single-layer (SL) 2S-MEM, which is the simplest form of ML-2S-MEM that requires much fewer input parameters by considering only one vegetation layer and the below ground (see Appendix B of *Li et al.* (2020)). Thus, the goal is to link the ML-2S-MEM with SL-2S-MEM so that the results derived from this research can facilitate retrieval practice. This is achieved by finding the effective VOD, τ_{eff}, and the effective scattering albedo, ω_{eff}, to minimize the cost function (CF) between $T_{\mathrm{B}}^{p,\mathrm{ML}}$ and the BT emitted from the equivalent single vegetation layer with beneath ground, $T_{\mathrm{B}}^{p,\mathrm{SL}}$:

$$\mathrm{CF} = \sum_{\{p,\theta_{\mathrm{RM}}\}} \left[T_{\mathrm{B}}^{p,\mathrm{ML}}(\theta_{\mathrm{RM}}, f) - T_{\mathrm{B}}^{p,\mathrm{SL}}(\tau_{\mathrm{eff}}, \omega_{\mathrm{eff}}, \theta_{\mathrm{RM}}, f) \right]^2, \tag{4.22}$$

where $\theta_{\mathrm{RM}} = \{0^\circ, 20^\circ, 40^\circ, 60^\circ\}$ and $p = \{\mathrm{H,V}\}$.

Formally (4.22) corresponds to a two-parameter retrieval of $\{\tau_{\mathrm{eff}}(f), \omega_{\mathrm{eff}}(f)\}$ derived from the modeled $T_{\mathrm{B}}^{p,\mathrm{ML}}(\theta_{\mathrm{RM}}, f)$ considered as "synthetic measurements" under ideal conditions (i.e., without noise and interference). The optimization process assumes τ_{eff} and ω_{eff} are at nadir and are only dependent on f, while their dependence on p and θ_{RM} is not considered. This assumption not only simplifies the representation of τ_{eff} and ω_{eff}, but also aligns with the current widely used {SM, VOD} retrieval algorithms *Li et al.* (2022); *Fernandez-Moran et al.* (2017) that only retrieves one VOD value at nadir for a pixel at a certain frequency according to a pre-defined scattering albedo. It should be noted that, however, the MEMLV can certainly include the polarization- and angular-dependence of τ_{eff} and ω_{eff} in the optimization process for future studies.

4.3 Results and Sensitivity Analysis

This section presents the results simulated using the MEMLV. First the results based on the original TSM parameterized by the JRC tree is presented. Then sensitivity analyses are shown for further evaluation of the impact of different parameters.

4.3.1 Effective scattering albedo and VOD based on the JRC structure

Figure 4.6(a) shows that as frequency goes higher, τ_{eff} increases most rapidly from L- to S-band (up to $f \approx 3$ GHz), then it increases almost linearly up to $f \approx 15$ GHz, finally it stays consistent from $f \approx 15$ GHz to $f \approx 37$ GHz. The τ_{eff} is in general greater than or similar to τ_j up to $f \approx 10$ GHz, then τ_{eff} becomes smaller than τ_j (especially for $f > 20$ GHz) due to the saturation effect. To the authors' knowledge, this is the first time that the saturation at high frequency has been quantitatively simulated using radiative transfer modeling.

To better understand this saturation effect, it is important to distinguish the physical meaning of τ_j from τ_{eff}. The τ_j corresponds to the definition of the extinction of a medium with a certain thickness d, as expressed in (4.7). If the medium has a fixed extinction coefficient κ_{e} (i.e. constant number density of scatterers), τ_j is proportional to d and can become infinitely large when $d \to \infty$. In contrast, τ_{eff} is obtained via an optimization process that minimizes the difference between $T_{\mathrm{B}}^{p,\mathrm{ML}}$ and $T_{\mathrm{B}}^{p,\mathrm{SL}}$, as shown in (4.22). Above a certain frequency, when an increase in τ_j can no longer effectively change $T_{\mathrm{B}}^{p,\mathrm{ML}}$, the saturation of τ_{eff} occurs. Another reason for τ_{eff} being smaller than τ_j is that the simulation of τ_{eff} considers canopy gaps and tree crown shapes, while τ_j does not take these aspects into account.

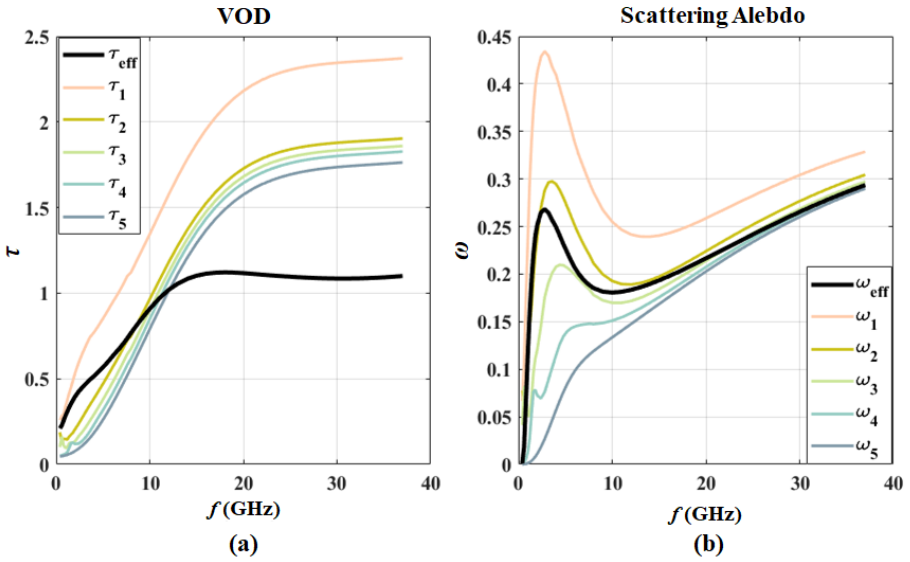


Figure 4.6: Frequency spectra of the effective- and layer- radiative transfer properties simulated for a forest stand consisting of JRC trees. (a) VOD τ_{eff} and τ_j; (b) scattering albedo ω_{eff} and ω_j.

Figure 4.6(b) shows that $\omega_{\text{eff}} \approx 0$ at P-band. It then increases rapidly, reaching a peak at S-band ($f \approx 3$ GHz). Following this peak, ω_{eff} decreases until the X-band ($f \approx 10$ GHz) and then steadily increases again at higher frequency. This peak is partially due to the frequency spectrum of the tree constituents' dielectric constant ε_c simulated with (8). As shown in Fig. 4.5, the imaginary part ε_c'' has a minimum at approximately the same frequency as the peak of ω_{eff}. Accordingly, absorption contributes less to the total extinction, resulting in an increased ω_{eff} at S-band. For $f > 10$ GHz, the increasing trend of ω_{eff} is due to the dominating contribution from needles, whose single scattering albedo increases monotonically with frequency.

The single scattering albedos of individual layers, ω_j, shown in Fig. 4.6(b) have a relatively large dispersion when $f < 10$ GHz. The ω_j of the bottom crown layer ($j = 1$) has the greatest value, particularly when $f < 10$ GHz, because it has the largest fraction of large scatterers (trunks and large branches). As the layer gets higher in the canopy (increasing j), the fraction of the small scatterers (small branches and needles) increases, resulting in a reduced value of ω_j. For $f < 10$ GHz, ω_{eff} demonstrates a combined effect of ω_j from all layers due to their higher transmissivity at this frequency range. For 10 GHz $< f < 20$ GHz, ω_{eff} is closer to the ω_j of higher layers (e.g. $j = 3, 4, 5$) due to the reduced sensing depth. For $f > 20$ GHz, all ω_j become closer to each other, and ω_{eff} is almost the same as ω_j of the uppermost layer ($j = 5$) due to the weak transmissivity.

4.3.2 Sensitivity analysis

As explained in Section 4.2.2, the TSM was developed based on a discretization of the measured architecture of the 5m-high JRC tree (Fig. 4.3). The discretization considers $N = 5$ layers with thickness $d = 1$ m (Fig. 4.4). To investigate the impact of the tree height H, d was scaled by a parameter γ. Here, it is defined as $\gamma = \{1, 2, 3, 4, 5\}$, resulting in the respective $d = \{1, 2, 3, 4, 5\}$m and $H = \{5, 10, 15, 20, 25\}$m.

According to (4.7), height scaling is reflected in scaling of τ_j^p of each layer j. To better simulate the realistic structure of trees in coniferous forests, the self-shading effect *Kramer* (2012) due to light competition has been considered. As H increases, the lower layers of trees become more shaded by the upper layers of trees, resulting in a lower number density of tree constituents, ρ_i. Since τ_j^p is also dependent on ρ_i according to (4.6) and (4.7), the scaling factor of τ_j^p for lower layers would be smaller than γ. Therefore, the layer-VOD scaling factor, γ_τ, is defined as,

$$\gamma_\tau = \begin{cases} \gamma, & j = 0 \\ \gamma - (\gamma - 1) \cdot (1 - j/N), & j = 1, 2, ..., N \end{cases} \tag{4.23}$$

For the trunk layer ($j = 0$), the γ_τ is the same as γ, as the self-shading effect has almost no effect on the trunk. When $\gamma = 1$, γ_τ is always equal to γ for self-consistency , as observed in the original 5m JRC tree. As γ increases, the VOD of the top layer ($j = N$) always has an equivalent increase, since the self-shading effect is minimal for this layer. When the layer becomes lower, γ_τ decreases but still remains greater than 1. For instance, when the tree height is 25m (scaled by $\gamma = 5$), the VOD of the bottom layer ($j = 1$) increases to only 1.8 times that of the original 5m tree due to the self-shading effect.

It is assumed that the height scaling has no impact on the scatterers' size distribution for each layer, resulting in that single scattering albedo ω_j^p remain unchanged. This assumption significantly simplify the whole process.

The crown-shape parameter, α, also affects ω_{eff} and τ_{eff}. As shown in (4.9) and Fig. 4.4, a tree with a lower α would have a more cube-like crown, whereas with a higher α, it would have a more pyramid-like crown. Because changing α in the TSM also changes the volume of each layer and the total number of all constituents, further adjustment has been made to B and n_0 in (4.14) and (4.15) to obtain comparable results. These adjustments are only effective when $\alpha \neq 0.6$ to remain the self-consistency of the original TSM. When α decreases, the adjustments made in B and n_0 lower the overall number density and slightly change the size-distribution of constituents, making the number of constituents contained within an individual tree similar to the JRC tree.

The resulting τ_{eff} and ω_{eff} for trees with $H = \{5, 10, 15, 20, 25\}$m and $\alpha = \{0.1, 0.6\}$ are plotted in Fig. 4.7. The insets show the τ_{eff} at P- and L-band (P-τ_{eff} and L-τ_{eff}), and ω_{eff} at L-band (L-ω_{eff}). In general, τ_{eff} increases and ω_{eff} decreases as H increases. There is a clear saturation for τ_{eff} at $f > 15$ GHz and/or $H > 10$ m due to the limited sensing depth at high frequency. Likewise, the H-dependence of ω_{eff} gets weaker as frequency goes higher. Both the L-τ_{eff} and P-τ_{eff} have strong dynamics as H changes. When H changes from 5m to 25m, P-τ_{eff} almost doubled, while L-τ_{eff} increases by 50% – 70%. These quantities are dependent on how the VOD is scaled. For most cases, the P-ω_{eff} is negligible.

As α decreases, the air volume between tree crowns, and the number density of constituents within the volume of tree crowns decreases. In other words, the constituents are more evenly and sparsely distributed among the canopies. This causes a general increase for τ_{eff}, as shown by the comparison between Fig. 4.7(a) and (b). Regarding the scattering albedo, it is interesting to see from Fig. 4.7(d) that L-ω_{eff} reduces significantly as H increases. For this case ($\alpha = 0.1$), the tree crown has a cube-like shape and most of the large scatterers in lower layers are obscured by above canopy layers that contain predominantly small scatterers. When H increases, the effect from these large scatterers in lower layers become weaker due to the increased thickness of upper layers. Since large scatterers have relatively high single scattering albedo at low frequencies, this obscuring effect causes an obvious decrease in L-ω_{eff} as H increases. This effect is not very obvious for the case with a more pyramid-like crown ($\alpha = 0.6$) shown in Fig. 4.7(c). For this case, regardless of the canopy height, there is always a relatively large area for lower layers exposing directly to the radiometer due to the pyramid-like shape of crown. Therefore, the scattering effect from those large scatterers in lower layers persistently affect ω_{eff} when H increases.

Besides H and α, the canopy-gap fraction β poses another important macroscopic parameter. Fig. 4.8 shows ω_{eff} and τ_{eff} for a forest consisting of scaled JRC trees with $H = 15$ m when β varies from 0.05 to 0.25. It is seen that the τ_{eff} in general decreases significantly as β increases because of the reduced area covered by trees. Compared to τ_{eff}, ω_{eff} demonstrates greater robustness to variations in β, with only a slight decrease in the magnitude of ω_{eff} as β increases.

4.3.3 Impact of Tree Constituents and Their Dry-matter Fraction

It is known that the VOD correlated with VWC. The impact of dry-matter fraction and sap-water salinity on τ_{eff} and ω_{eff} is evaluated and plotted in Fig. 4.9 (a) and (c). The evaluation is made for a 15m-high forest for the "normal" and "dry" cases. The "normal" case follows the parameterizations used in Section 4.2.2: $m_{\text{d}} = 0.15\,\text{kg} \cdot \text{kg}^{-1}$ and $S = 10$ ppt for needles, and $m_{\text{d}} = 0.3\,\text{kg} \cdot \text{kg}^{-1}$ and $S = 10$ ppt for branches and trunks. The "dry" case assumes $m_{\text{d}} = 0.3\,\text{kg} \cdot \text{kg}^{-1}$ and $S = 10$ ppt for needles, $m_{\text{d}} = 0.4\,\text{kg} \cdot \text{kg}^{-1}$ and $S = 5$ ppt for branches, and $m_{\text{d}} = 0.7\,\text{kg} \cdot \text{kg}^{-1}$ and $S = 5$ ppt for trunks.

The effects of m_{d} on τ_{eff} and ω_{eff} are shown in Fig. 4.9(a) and (c), respectively. The dry case has a lower τ_{eff} than the normal case up to $f \approx 13$ GHz. Specifically, the P-τ_{eff} and L-τ_{eff} for the dry case are about 35%

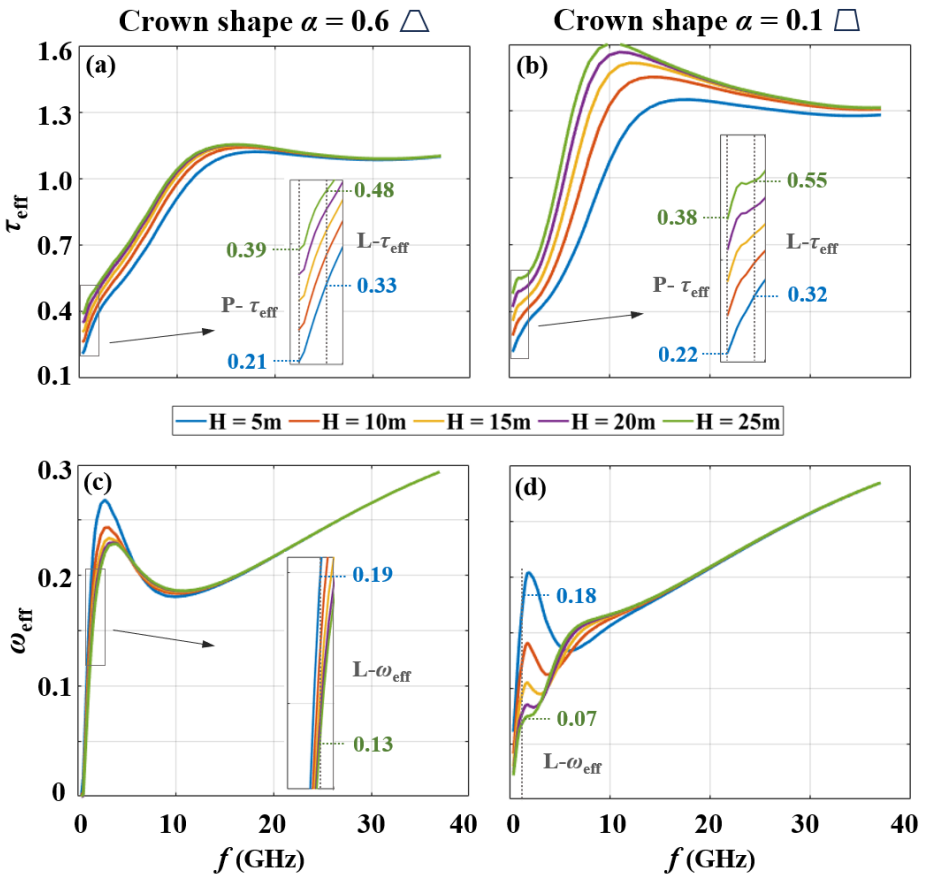


Figure 4.7: Frequency spectra of the effective radiative transfer properties of coniferous forests with scaled heights $H = \{5, 10, 15, 20, 25\}$m and crown-shape parameter $\alpha = \{0.1, 0.6\}$. (a) τ_{eff}- and (c) ω_{eff} for $\alpha = 0.6$; (b) τ_{eff}- and (d) ω_{eff} for $\alpha = 0.1$.

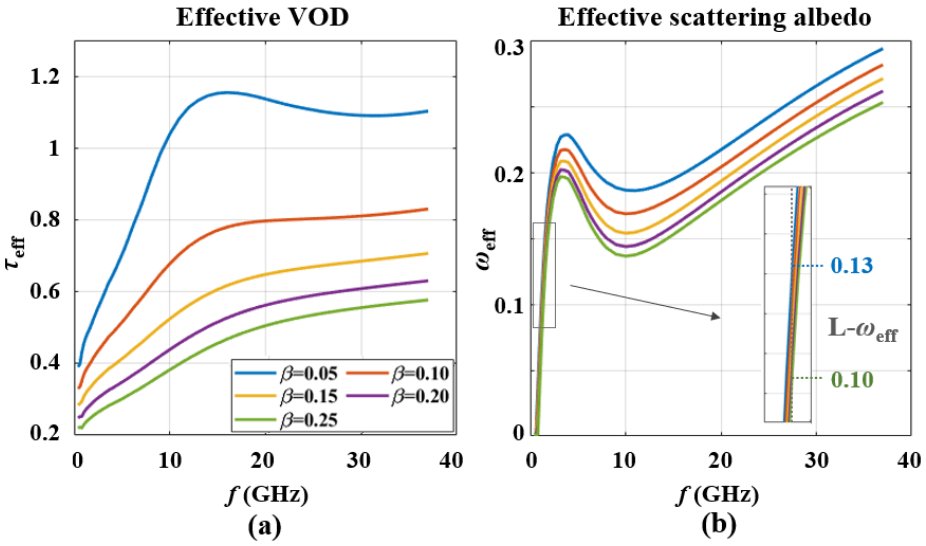


Figure 4.8: Frequency spectra of the effective radiative transfer properties for coniferous forests with height $H = 15$ m and canopy-gap fractions $\beta = \{0.05, 0.1, 0.15, 0.2, 0.25\}$ (a) for τ_{eff}; (b) for ω_{eff}, where the representation of line colors is indicated in the legend of (a).

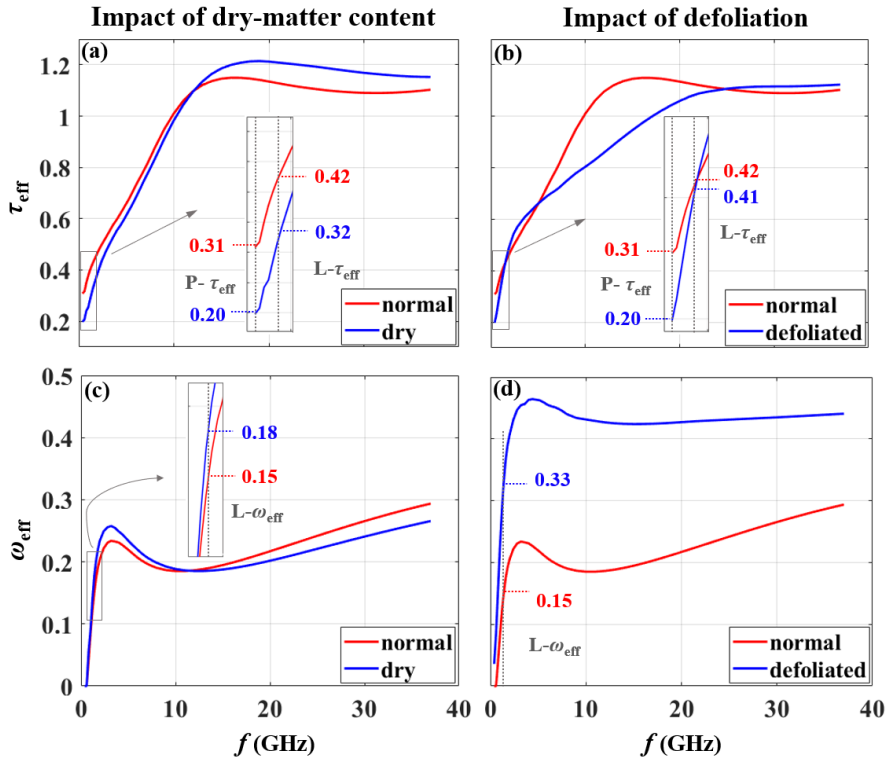


Figure 4.9: Frequency spectra of the effective radiative transfer properties for coniferous forests with height $H = 15$ m under different conditions. (a) τ_{eff}- and (c) ω_{eff} for "normal" and "dry" canopy; (b) τ_{eff}- and (d) ω_{eff} for "normal" and "defoliated" canopy.

and 25% lower than for the normal case, respectively. Compared with the normal case, the τ_{eff} of the dry case reaches saturation at higher frequency. After reaching saturation, τ_{eff} for the dry case is slightly larger ($< 10\%$) than for the normal case. The comparison between ω_{eff} for the normal and dry cases follows the opposite trend. For the dry case, ω_{eff} is slightly higher at low frequency but lower at high frequency. At low frequency, when m_{d} increases, the single scattering albedo $\hat{\omega}^p$ of branches and trunks increases, leading to an increase in ω_{eff} for the dry case. As frequency increases, the radiative transfer contribution from needles becomes dominated. The $\hat{\omega}^p$ of dry needles has lower increasing rate along f. This explains why ω_{eff} could become lower for the dry case at very high frequencies.

The effects of defoliation on τ_{eff} and ω_{eff} are shown in Fig. 4.9(b) and (d), respectively. For this analysis, the normal case is the same as before, while the "defoliated" case assumes the number density of needles as 0. The results show that both the P-τ_{eff} and L-τ_{eff} for the defoliated case are smaller than for the normal case. The impact of leaves on L-τ_{eff} is minor, agreeing with other studies *Guglielmetti et al.* (2007); *Santi et al.* (2009). From X- to Ku-band, τ_{eff} for the normal case is significantly higher than for the defoliated case, highlighting the suitability of these bands for detecting the defoliation states of forests. It is interesting to see that τ_{eff} of the defoliated case reaches saturation at much higher frequency (≈ 25 GHz) compared with the normal case (≈ 13 GHz). Since needles in general have small single scattering albedo, ω_{eff} for the defoliated case is higher than for the normal case, particularly at low frequencies.

4.3.4 Summary

Simulations demonstrate the dependence of τ_{eff} and ω_{eff} on the macro-parameters of coniferous forests (e.g., canopy height and -gap fraction), the structure of individual trees (e.g., crown shape), and the tree constituents (e.g., foliation stages and dry-matter fraction). The impacts of these factors are summarized in Table 4.2, where the sign "↑" indicates increasing, "↓" means decreasing, and "∼" suggests staying relatively consistent. Table 4.2 offers a qualitative estimation of the dependence of τ_{eff} and ω_{eff} on different ancillary parameters to

Table 4.2: Summary of the impact from different parameters on the effective VOD and albedo

	$\omega_{\text{eff}}(f)$			$\tau_{\text{eff}}(f)$		
f (GHz)	1-5	5-15	15-37	1-5	5-15	15-37
Tree crown slope α ↑	↑	↑	∼	↓	↑	↑
Tree height H ↑	↓	∼	∼	↑	↑	∼
Canopy-gap fraction β ↑	↓	↓	↓	↓	↓	↓
Leaf number/volume ↑	↓	↓	↓	↑	↑	∼
Dry-mat. m_{d}↑ & Sal. S ↓	↑	∼	↓	↓	∼	∼

Table 4.3: An example of effective VOD and scattering albedo for a forest stand at P- (0.4 GHz), L-(1.4 GHz), C-(6.8 GHz), X-(11 GHz), Ku-(19 GHz), and Ka-(37 GHz) bands.

	P-τ_{eff} (P-ω_{eff})	L-τ_{eff} (L-ω_{eff})	C-τ_{eff} (C-ω_{eff})	X-τ_{eff} (X-ω_{eff})	Ku-τ_{eff} (Ku-ω_{eff})	Ka-τ_{eff} (Ka-ω_{eff})
5m	0.210 (0.000)	0.334 (0.194)	0.680 (0.196)	0.969 (0.181)	1.120 (0.212)	1.100 (0.294)
10m	0.262 (0.000)	0.381 (0.166)	0.744 (0.197)	1.030 (0.184)	1.136 (0.212)	1.103 (0.294)
15m	0.308 (0.000)	0.420 (0.150)	0.784 (0.199)	1.060 (0.185)	1.141 (0.212)	
20m	0.351 (0.000)	0.454 (0.139)	0.810 (0.200)	1.075 (0.186)	1.143 (0.212)	
25m	0.390 (0.000)	0.484 (0.133)	0.826 (0.201)	1.083 (0.186)		

support improving the retrieval algorithms.

Quantitatively, Table 4.3 provides exemplary values of τ_{eff} and ω_{eff} for different frequencies f and tree heights H with a fixed canopy-gap fraction $\beta = 0.05$. The results further corroborate the conclusions drawn previously. It is seen that at C-band and higher frequencies, the H-dependence of ω_{eff} and τ_{eff} is very weak due to the reduced sensing depth. The frequency and tree height at which this effect becomes obvious depend on forest structures and dry-matter fraction, as demonstrated in Figs. 4.7-4.9.

4.4 Application in VOD retrieval

There are numerous possibilities of applying the MEMLV to enhance the multi-frequency VOD retrieval. The major aspects are outlined as follows:

(i)Parameterization of retrieval algorithms:

The sensitivity study summarized in Table 4.2 provides guidelines for adjusting the parameterization of scattering albedo in retrieval based on canopy characteristics. The values of ω_{eff} in Table 4.3 can facilitate the retrieval of variables such as VOD through the inversion of 2S-MEM. Although their direct applicability may be affected by natural forest variability, their frequency- and tree height-dependent trends are expected to hold for natural forest stands. Consequently, VOD retrieval could be optimized by using scaled ω_{eff} values while maintaining the trends of frequency- and height-dependence observed in Table 4.3. This approach offers a more dynamic and representative parameterization strategy for retrieval algorithm, potentially in combination with data from other sensors. For instance: (1) The ω_{eff} of forests can be adjusted based on tree height H measured by LiDAR or InSAR, along with canopy-gap fraction β and tree crown-shape parameter α estimated using image pattern recognition techniques. This will enable data fusions from the modelling level to improve estimations of ω_{eff} and enhance VOD retrieval for forests with various characteristics. (2) Seasonal variations in tree constituents could be factored into parameterizing ω_{eff} for VOD retrieval. For example, defoliation or reduced liquid sap water during cold periods may result in higher ω_{eff}, particularly at low microwave frequencies. Conversely, during warming periods, one might expect a decrease in ω_{eff} due to growth and associated increasing gravimetric water content of canopy constituents.

(ii) Interpretation of retrieved VOD:

Properly interpreting the retrieved VOD is often challenging due to limited knowledge about the physical interactions between canopies and electromagnetic waves at various frequencies. The simulated τ_{eff} using the MEMLV provides physics-based explanations, justifies assumptions and elucidates specific behaviors of the retrieved VOD. Furthermore, the simulated τ_{eff} helps establish the relationships between VOD and the architecture and physiology of canopies, as presented in Figures 4.7-4/9. Consequently, this facilitates the quantification of parameters such as AGB and VWC based on the retrieved VOD.

(iii)Definition of future campaigns and missions:

The crucial preparatory phase before launching microwave campaigns and missions requires the definitions of their scientific needs, advantages and goals. The MEMLV enables sensitivity analysis of microwave measurements across different frequencies and environmental conditions to support in such definitions. For instance, our simulations indicated a negligible P-ω_{eff} for coniferous forests (Table 4.3). This finding is an advantageous for future P-band missions aimed at quantifying forest state parameters, as it simplifies the retrieval process and reduces uncertainties involved in the retrieval scheme.

4.5 Conclusion and Future Work

A new compound modelling tool, the MEMLV, is developed and presented along with sensitivity analyses for coniferous forests. It is designed as a toolbox to simulate the effective radiative transfer parameters across frequencies of land covered with different types of vegetation under a variety of conditions. The primary aim of the MEMLV is to bridge the gap between highly detailed physics-based forward emission models and retrieval processes over a wide frequency range. To achieve this goal, the MEMLV incorporates three major components:

(i) the SL-DSM for calculating the single scattering albedo and VOD of a single-layer of vegetation canopies;

(ii) the TSM for representing the forest architecture in multi-layer structure and defining the characteristics of tree constituents in each layer;

(iii) the 2S-MEM for calculating the BT of forested ground and obtaining the effective radiative transfer properties that can be used in retrieval practices.

The principal outcomes of this research are the frequency spectra of effective VOD, τ_{eff}, and effective scattering albedo, ω_{eff}. The τ_{eff} generally increases as frequency increases but can saturate at high frequency. The ω_{eff} increases rapidly at low frequency and peaks at around S-band, then it decreases until X-band (10 GHz), and finally, it increases almost linearly as frequency further increases.

Another major finding concerns the sensitivity analyses of τ_{eff} and ω_{eff} with respect to crucial model parameters. They are summarized below:

(i) Canopy height H: at low frequencies, ω_{eff} slightly decreases and τ_{eff} increases when H increases. At high frequencies, τ_{eff} and ω_{eff} show minimal variation with H due to saturation.

(ii) Canopy-gap fraction β: when β increases, ω_{eff} slightly decreases while τ_{eff} significantly decreases.

(iii) Dry-matter fraction m_{d}: as m_{d} increases, ω_{eff} increases while τ_{eff} decreases at low frequency.

(iv) Defoliation: it can cause an increase in ω_{eff} across all frequencies, a slightly lower τ_{eff} at P- and L-band, and a significant lower τ_{eff} from X- to Ku-band.

Although the simulated radiative transfer properties have shown encouraging results when compared with satellite retrieval and literature values, rigorous ground-based microwave experiments are essential to fully validate MEMLV. Moving forward, it is crucial to perform close-range multi-frequency radiometer experiments on forest sites with well-documented ground-truth data. Additionally, experiments conducted under controlled environmental conditions (e.g., irrigated vs. non-irrigated) and across different growth seasons would be also very valuable for thoroughly assessing the performance of the MEMLV. In the long term, the MEMLV is anticipated to serve as a universal tool for the remote sensing community, constructively supporting and improving retrieval practices for satellites operating across frequencies from P- to Ka-band.

Chapter 5

Deriving VOD and AGB from passive microwaves (WP 2-3-4)

5.1 Introduction and purpose

The purpose of this part, is to detail the setup to derive VOD, the retrieval chains and their configurations. Data from the SMOS, AMSRE and AMSR2 sensors are used to derive the VOD at L, C and X bands. The following represents the activities developed for the tasks (work packages) 2, 3 and 4 of the project (see 3).

5.2 Input data

5.2.1 General Considerations

All the products discussed in the following are provided for the OSMOSE project as:

- **Format**: netcdf files, characterized by the .nc extension name
- **Grid system**: EASE Grid version 2 (*Brodzik et al.* (2012a), *Brodzik et al.* (2014)), that is 25km x 25km at 30deg of latitude
- **Grid dimensions**: global earth is covered by a matrix of 584x1388 pixels

5.2.2 SMOS

The Level 3 Brightness Temperature product (*Al Bitar et al.* (2017a)) is a daily global polarised brightness temperature product, arranged by incidence angle values, in full polarisations (H, V, stokes3, and stockes4). It includes all brightness temperatures acquired that day, transformed to ground polarisation reference frame (H and V polarisation), binned (5^o), and averaged into fixed angle classes. Bin 9 corresponds to SMAP incidence angle (40^o +/- 2^o), and bin 15 is a placeholder for the time being. Ascending and descending orbits are processed separately.

5.2.3 AMSRE

The Advanced Microwave Scanning Radiometer for EOS (AMSR-E) aboard NASA's Aqua satellite, provides Brightness Temperatures at Vertical and Horizontal polarisations at an incidence angle of 55 deg, for the 6

frequencies 89.0 GHz, 36.5 GHz, 23.8 GHz, 18.7 GHz, 10.65 GHz and 6.925 GHz. The OSMOSE project focuses on the C and X bands, so only the 6.925 GHz and 10.65 GHz are kept.

- Radiometric accuracy: $<$0.5 K
- Spatial Resolution: 51 x 29 km (10.65 GHz, X-band), 75 x 43 km (6.925 GHz, C-band)
- Incidence Angle: 55 deg.

The purpose of the OSMOSE project is to set up an homogenized VOD data set, on the EASE Grid version 2 (25 km). The NSIDC provides the AMSRE data on such grid, called the "MEaSUREs dataset" (*Brodzik et al.* (2016), *Brodzik and Long* (2018)), but the data are split into two products, one on the global projection, but only covering latitudes lower than 70 deg. and one on the polar grid. The two projections are then combined daily (ascending/descending orbits processed separately) to ensure a global coverage, i.e. including the high latitudes. The MEaSUREs database encompasses different passive microwave sensors (e.g. SSM/I, SSMR...), which are inter-calibated and for which the projection on the EASE Grid 2 is properly done, in case the methods developed for the OSMOSE project can be applied to the other passive microwave satellites.

5.2.4 AMSR2

The AMSR2 sensor is onboard the GCOM-W1 satellite, launched in July 2012 by the JAXA. It provides brightness temperatures at Vertical and Horizontal polarization, at an incidence angle of $\sim 55^o deg.$. The data are split between Ascending/Descending orbits, corresponding to an overpassing time of 13:30/01:30, *Maeda et al.* (2016).

- Name of dataset: LEVEL3
- original spatial resolution: 0.25 deg
- Website: https://suzaku.eorc.jaxa.jp/

5.2.5 ERA 5

ERA 5 is the re-analysis of climate data from ECMWF, which provides various atmospheric and surface geophysical information every hour, on a 0.25deg spatial resolution grid *Hersbach et al.* (2023). For the OSMOSE project, we used a subset of the catalog available from the website `https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=form`, described in the next section. The ERA5 are pre-processed for the applications to 1/ use only the closest time slot to the satellite overpassing time. It means for each satellite used in this project, and for each orbit, the hourly ERA5 data may change.

Other auxiliary data are considered in the retrieval:

- *IGBP Landcover*: The IGBP LandCover map *Broxton et al.* (2014a) is used to describe the surface land classification, and was spatially resampled from its native resolution to the EASE v2.0 25km Grid. The 17 landcovers are gathered into main classes presented in Table 5.1. The VOD retrievals are performed over the two classes NO and FO, for which sets of parameters are defined (see Table below).

- *Soil Grid texture*: The clay content in the soil comes from the Food and Agriculture Organization map *FAO* (1988). This map was reprojected onto the EASE 2.0 Grid. This information is used in the dielectric model of the soil, to compute the emission of the surface soil.

LANDCOVER CODE	Name	DT aggreg class
0	Special Discard LC Marker	None:0
1	Evergreen Needle leaf Forest	CFO:2
2	Evergreen Broadleaf Forest	CFO:2
3	Deciduous Needle leaf Forest	CFO:2
4	Deciduous Broadleaf Forest	CFO:2
5	Mixed Forests	CFO:2
6	Closed Shrublands	CNO:1
7	Open Shrublands	CNO:1
8	Woody Savanas	CFO:2
9	Savanas	CNO:1
10	Grasslands	CNO:1
11	Permanent Wetland	CNO:1
12	Croplands	CNO:1
13	urban and Built-Up	CNO:1
14	Cropland/Natural Vegetation Mosaic	CNO:1
15	Snow and Ice	CEI:7
16	Barren or Sparsely Vegetated	CNO:1
17	Ocean water (Saline)	CWS:5
18	Open Water (Pure)	CWP:4

Figure 5.1: Correspondence IBGP land cover classes and SMOS Fractions as NO: Nominal, FO: Forest, EI: Ice/permanent SNOW, WS: Saline Water, WP: pure water. "C" stands for Classes and "F" for Fraction. From CESBIO technical note SO-TN-CB-GS-057

5.3 VOD Output Data file

The last products to be detailed are the output products containing the VOD. Every band is reported as separate files, as well as the orbits which are distinguished between ascending and descending. Note that the ascending/descending orbits do not correspond for SMOS and AMSR-E/2.

The variables in the output files are summarized in the following table:

lat	latitude in degrees written as double, vector dim=1x584
lon	longitude in degrees written as double, vector dim=1x1388
Soil_Moisture	Surface Soil Moisture, in $m^3.m^{-3}$ written as double, 2D matrix dim=584x1388
Soil_Moisture_StdError	Soil moisture range due to modeling uncertainties, in $m^3.m^{-3}$ written as double, 2D matrix dim=584x1388
Optical_Thickness_Nad	or VOD, transmittivity of vegetation layer at Nadir, unitless written as double, 2D matrix dim=584x1388
Optical_Thickness_Nad_StdError	VOD range due to modeling uncertainties, unitless written as double, 2D matrix dim=584x1388
Quality_Flag	0: data OK, 1: data not recommended, 2: missing data, unitless written as byte, 2D matrix dim=584x1388
Scene_Flags	Scene Flags written as byte, 2D matrix dim=584x1388
Processing_Flags	Processing Flags, unitless written as byte, 2D matrix dim=584x1388
RMSE	RMSE between SMOS-TB and modeled TB, in Kelvin written as double, 2D matrix dim=584x1388
Soil_Temperature_Level1	ECMWF Soil Temperature at surface level 1, in Kelvin K. written as double, 2D matrix dim=584x1388
Days	Number of Days since 1/1/2000 written as integer, 2D matrix dim=584x1388
UTC_Seconds	Number of Seconds of the day written as integer, 2D matrix dim=584x1388
UTC_Microseconds	Number of Microseconds of the day written as integer, 2D matrix dim=584x1388
RFI_ratio	Index to evaluate the impact of RFI, unitless written as double, 2D matrix dim=584x1388
UTC_Minutes	Number of minutes since the start of the day written as float, 2D matrix dim=584x1388

Table 5.1: VOD file content

RFI_ratio
RFI_ratio (Table 5.1) only concerns SMOS-derived VOD. It is computed by dividing the number of RFI flags per bin (from L1c data) by the number of TB per bin. Users can filter the derived SM/VOD based on this ratio of RFI by applying a user-defined threshold. For example, applying a threshold of 0.2 means that any SM/VOD retrievals associated with a RFI_ratio greater than 0.2 should be discarded from any analysis (i.e. more than 20% of the input brightness temperatures are tagged as contaminated by RFI).

Scene_Flags
The Scene Flag is a binary flag encoded on a byte that identifies moderate or strong topography, frozen soil (if soil temperature below 273K) or if a pixel is polluted by more than 10% of Water/Urban/Permanent Ice land cover types. The table 5.2 summarizes the flag signification per bit.

Bit	8	7	6	5	4	3	2	1
Flag	No Flag	No Flag	No Flag	No Flag	Frozen Soil	Polluted Scene	Strong Topography	Moderate Topography

Table 5.2: Bit signification of the Scene_Flags

Processing_Flag
The Processing Flag is set to one when the RMSE between the modeled TB and the measured TB is above 12K.

Quality_Flag
The Quality flag is set to 1 when a Scene_Flag is raised or the Processing_Flag is raised.

5.4 Principle of the retrieval

In this pixel-based approach, the scene is considered homogeneous in terms of land classification and the properties of the pixels (roughness, vegetation, scattering albedo, etc) are area-weighted averages of the properties of the IGBP land cover classes present in the footprint.
In most cases, on appropriate pixels, a couple (SM, VOD) is fed to a forward model computing the corresponding TB at all incidence angles and for both horizontal and vertical polarizations. The values of the (SM, VOD) couple are iteratively modified until the cost function computed from the sum of the squared weighted differences between the modeled and measured TB reaches a minimum. The values of (SM, VOD) which minimize the cost function are written out to the output products. The general principle of the retrieval is described on Figure 5.2 and the following sections describe some of the processing steps more thoroughly.

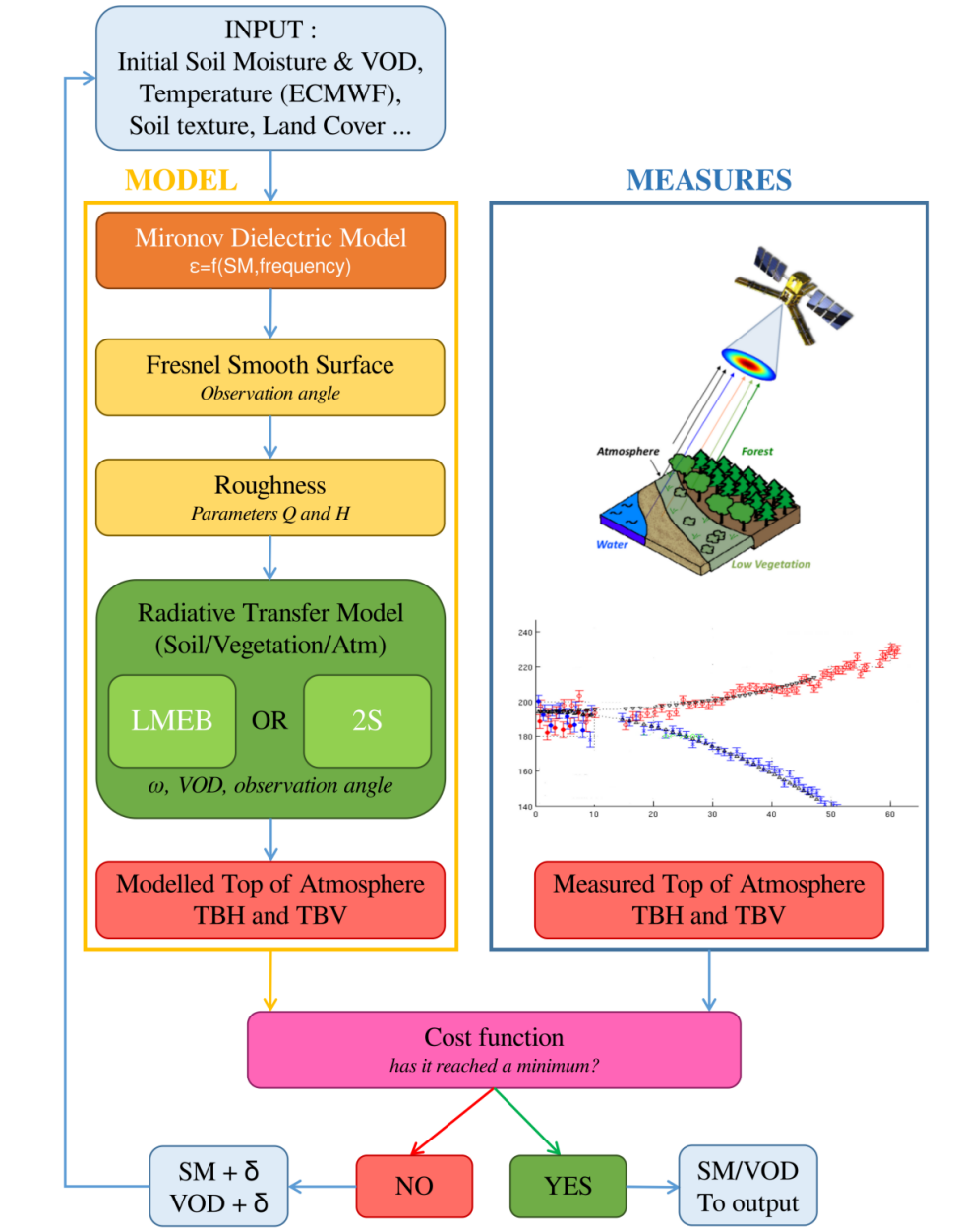


Figure 5.2: Retrieval principle of the processing platform

5.4.1 Cost function minimization

The VOD retrieval is based on the minimization of a cost function (pink box on Figure 5.2) of the following form:

$$FC(Par) = \sum_{\theta,p=\{H,V\}} \frac{(TB_{sat}(\theta,p) - TB_{model}(\theta,p,Par))^2}{\sigma^2_{TB_{sat}(\theta,p)}} + \sum_{Par=sm,tau_{nad}} \frac{(Par - Par_{prior})^2}{\sigma^2_{Par_{prior}}} \quad (5.1)$$

with $\text{TB}_{sat}(\theta, p)$ being the satellite brightness temperatures from passive microwave sensor (SMOS, AMSRE, AMSR2), at polarisation *p* (V Vertical or H Horizontal), θ the incidence angles, and $\sigma_{TB_{sat}(\theta,p)}$ the radiometric accuracy. $\text{TB}_{model}(\theta, p, Par)$ are modeled TB obtained from a radiative transfer model and derived free parameters *Par* constrained by their initial value Par_{prior} and uncertainties $\sigma_{Par_{prior}}$ (*Kerr et al.* (2012a, 2020)).
The derived parameters *Par* are the ones that minimize *FC*. In most cases, *Par* are the surface soil moisture and the VOD - Vegetation Optical Depth. For the OSMOSE project, $\text{TB}_{model}(\theta, p, Par)$ are computed using two reference radiative transfer models: the Two-Stream model and the L-MEB (L-band Microwave Emission of the Biosphere).

5.4.2 Radiative Transfer Model

5.4.2.1 L-MEB

The L-MEB is a zero-order radiative transfer model. The multi-layer approach considers the emission of soil, with a rough surface. The contribution of the vegetation is computed using the Tau-Omega model. All the elements are reported in *Wigneron et al.* (2017), and for SMOS in *Kerr et al.* (2012a). The Tau/omega model relies on a fixed omega (scattering albedo) and the free parameter Tau (τ), also called the VOD Vegetation Optical Depth.

5.4.2.2 2S

The two-stream (2S) is a first-order radiative transfer model accounting for multiple reflections, and developed by *Mätzler* (1998) and *Wiesmann and Mätzler* (1999). It is mentioned in this section as it was considered to perform the retrieval. Tests were done but could not be implemented in the processing chains.

5.4.2.3 Soil dielectric constant and emissivity

Various dielectric models exist to compute the dielectric constant of a soil. For the OSMOSE project, all cases consider the Mironov dielectric constant model *Mironov et al.* (2009), as it was designed and validated for the frequencies of interest for the OSMOSE project *Montpetit et al.* (2015). Then, the soil emssivity is computed using the fresnel law at the soil-atmosphere interface.

5.5 Operational configuration of the retrieval processing chains

This section details the setup defined in the radiative transfer model. The retrieval principle described in the previous part is applied to all frequency bands with some adjustments. The processing chain automatically identifies the sensor and the frequency based on the input TB product. The configuration files vary with the frequency, as some parameters need to be adapted. In order to tune the retrieval platform following each sensor's characteristics (e.g frequencies, multi-observation angles for SMOS or one observation angle for AMSRE and AMSR2), the values of several key parameters were tested out and optimized (see section 5.6).

The processing configuration can be summed up as follows:

Table 5.3: Configuration set up of the VOD retrieval algorithm. *RTM: Radiative Transfer Model, *: LV low vegetation, **: FO forest*

Band	**L**	**C**	**X**
RTM	L-MEB	L-MEB	L-MEB
Q	0	0	0.13
Hr	0.1 (LV*); 0.3 (FO**)	0.15	0.15
ω	0.1 (LV*); 0.06 (FO**)	0.05	0.05
derived parameters	SM-VOD	SM-VOD	SM-VOD
boundary retrieved SM	None	[0-1]	[0-1]
boundary retrieved VOD	None	[0-2]	[0-2]
Effective soil temperature	ERA5 soil level 1	ERA5 soil level 1	ERA5 soil level 1
Vegetation Temperature	ERA5 skin t.	ERA5 skin t.	ERA5 skin t.
Dielectric model	Mironov	Mironov	Mironov
SM initial VOD init(*)	0.2 global map	0.2 global map	0.2 global map

Initial values of free parameters are needed to start the first iteration of the minimization. SM is fixed at 0.2 m^3 m^{-3}. The initial VOD (* in Table 5.3) are computed as an average of VOD from a three-year run.

5.6 VOD sensitivity analysis

To produce VOD from Brightness Temperatures (TB), the OSMOSE project capitalized on an existing algorithm for the L- band and adapted it to the C and X bands. The Table below (Table 5.4) summarizes the tested parameter values.

Table 5.4: Tested parameters and associated values for tuning the retrieval platform per frequency band

Parameter	**Tested value**		
Band	**L**	**C**	**X**
RTM	LMEB or 2S	LMEB or 2S	LMEB or 2S
Hr	[0 0.15 0.3 0.4]	[0 0.15 0.3 0.4]	[0 0.15 0.3 0.4]
RFI filtering	[0.2 0.8]	N/A	N/A
ω	$\omega = f(TH)$	[0.01:0.01:0.15] $\omega = f(TH)$	[0.04 0.05 0.06 0.08]
Free parameters	VOD/SM VOD/ACard	VOD/SM VOD/ACard VOD only (SM from ERA5)	VOD/SM VOD/ACard VOD only (SM from ERA5)
Retrieval limits	Free or 0⩽SM⩽1 & 0⩽VOD⩽2	Free or 0⩽SM⩽1 & 0⩽VOD⩽2	Free or 0⩽SM⩽1 & 0⩽VOD⩽2
Veg. Temperature	Skin Temperature 2m Air Temperature	Skin Temperature 2m Air Temperature	Skin Temperature 2m Air Temperature

Here are some explanations of some tested parameters:

- *RFI filtering*: to assess the retrieval quality when applying different RFI filtering thresholds, for SMOS only. The L3TB impacted by RFI are filtered using the ratio Nb_RFI_Flags / Nviews, i.e. numbers of SMOS TB impacted by RFI over the number of TB. Two thresholds are tested, 0.2 being too restrictive removing too many TBs. A threshold of 0.8 allows us to deal with more TBs. It is justified as two more filters are applied on the VOD: removing VOD with an RMSE¿8K and outliers when computing the yearly average.
- ω: was already fine-tuned for the L-band MEB. Tests were carried out with a Tree Height (TH) dependent omega based on the results from WSL analyses (*Zhou et al.* (2024)). For C and X bands, additional tests with fixed ω values were conducted.
- *Free parameters*: As it was unsure whether SM would be a relevant parameter for C and X-bands, some tests were conducted to retrieve VOD and ACard, instead of VOD and SM. ACARD is the cardioid model *Waldteufel et al.* (2004), which is a simple model to estimate the module of the dielectric constant. It can be used to characterize particular cases (frozen/bare soils) where the soil model (i.e. Mironov dielectric model) is not adapted. In the cases of X and C bands, we tested 2-parameter retrievals with ACard instead of the SM, as it was not evident that performing a SM/VOD retrieval would lead to satisfying results in terms of VOD and also in terms of RMSE (modeled TB model versus SMOS TB). Indeed, C and X is less sensitive to the soil in particular with a vegetation layer. Additionally, the retrieval of a single parameter (with SM from ERA5) versus the retrieval of two parameters was also tested.
- *Retrieval limits*: Assess the impact on the VOD, of constraining the free parameters in the cost function.

In the following section, we present the sensitivity of the VOD at X and C bands with respect to the various parameters, as described above.

5.6.1 Influence of the Radiative Transfer Model

In the frame of this project, two Radiative Transfer Models (RTM) were tested:

- The L-band Microwave Emission of the Biosphere Model (LMEB) which uses the classical $\tau - \omega$ model *Wigneron et al.* (2017)
- The Two-Stream Model (2S) which is able to simulate higher degrees of interaction between different layers (soil, vegetation, atmosphere) *Schwank and Naderpour* (2018)

A test for the L-band was conducted over 1 month for both ascending and descending orbits and using both RTM. The ω value is converted from LMEB values to 2S values using Eq. 23 from *Schwank and Naderpour* (2018). It means that the 2S model is not used per se, but is rather fitted into the LMEB. The ω_{LMEB} values are 0.06 over forest pixels and 0.1 over low vegetation pixels. The output VOD were then averaged for each orbit type. One of the final purposes of the microwave part of the OSMOSE project is the estimation of the AGB from multi-frequency VOD. When conducting the sensitivity analysis of different parameters on the VOD, the spatial correlation of this averaged VOD with respect to the CCI AGB is a relevant indicator. In this study, the Pearson correlation coefficient (R) and Spearman correlation coefficient (Rho) are used. Figure 5.3 shows the spatial density plot between the AGB from the CCI *Santoro and Cartus* (2023) and the 1-month average VOD for ascending and descending orbits using the 2S or the LMEB for the descending (top panel) or ascending orbits (bottom panel).

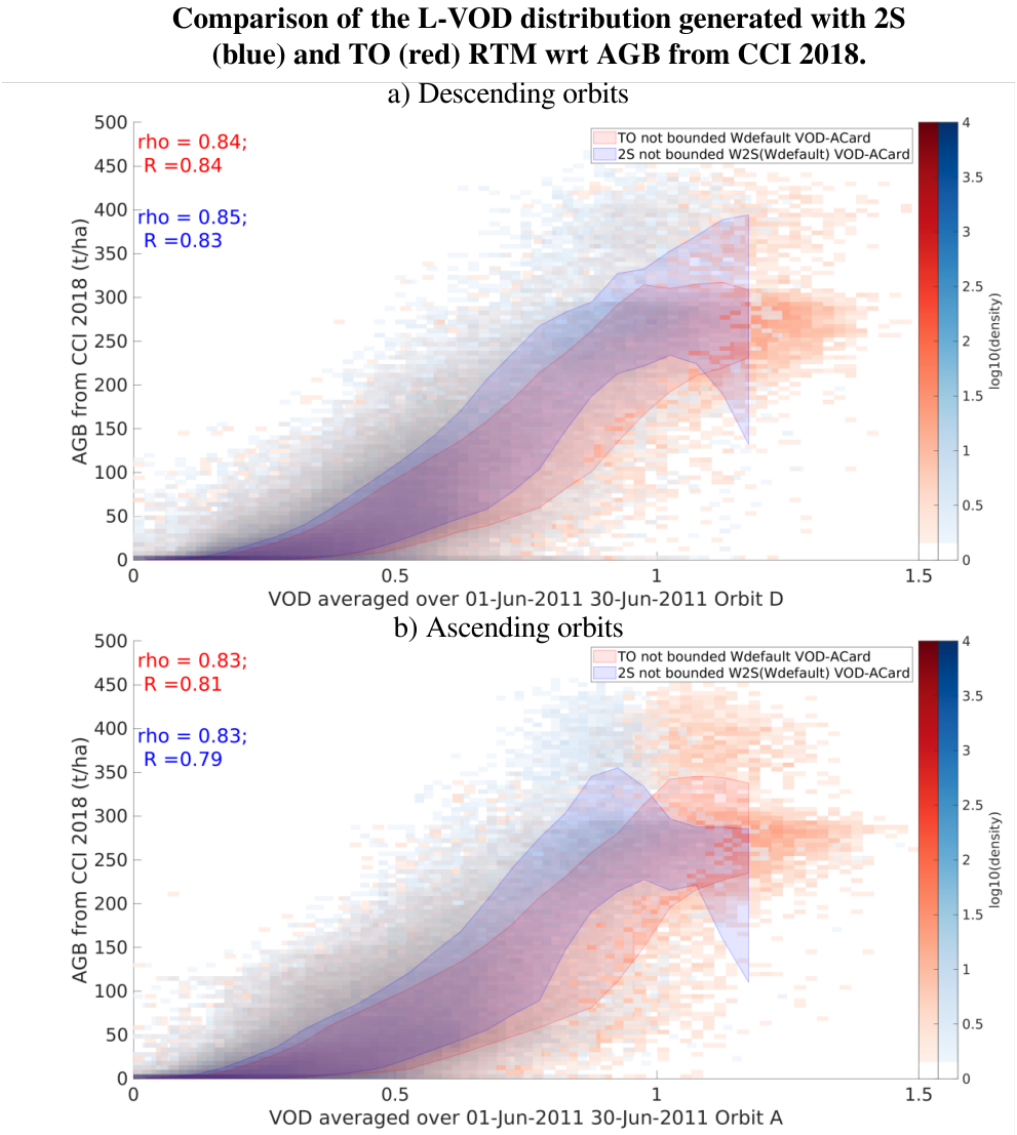


Figure 5.3: Density plot of the OSMOSE L-VOD for June 2011 against the AGB from CCI 2018. The red color refers to the use of the TO model to retrieve the VOD and the blue color refers to the use of the 2S model. In both cases, the couple (VOD/ACard) was retrieved, the retrieval was not constrained and the default values of ω were used (eq 23 of *Schwank and Naderpour* (2018) was used to convert from TO to 2S value). The shaded areas correspond to the interval in-between quantiles 16 and 84 and contains 68% of the distribution. (a) for descending orbits and (b) for ascending orbits.

Figure 5.3 shows that the RTM does not significantly impact the VOD correlation with the AGB for the L-band for both orbit types. In both cases, the 2S L-VOD saturates to a lower value than the TO L-VOD. The choice of the RTM does not greatly impact the distributions.

A similar test was conducted for C-band over July 2011. The ω value are derived from the MEMLV (outputs of OSMOSE WP1, WSL) and depends on the the TH. A map of the C-band omega for the TO RTM is displayed in Figure 5.4. The ω values from TO were converted to 2S using Eq. (23) of *Schwank and Naderpour* (2018). The results are displayed on Figure 5.5.

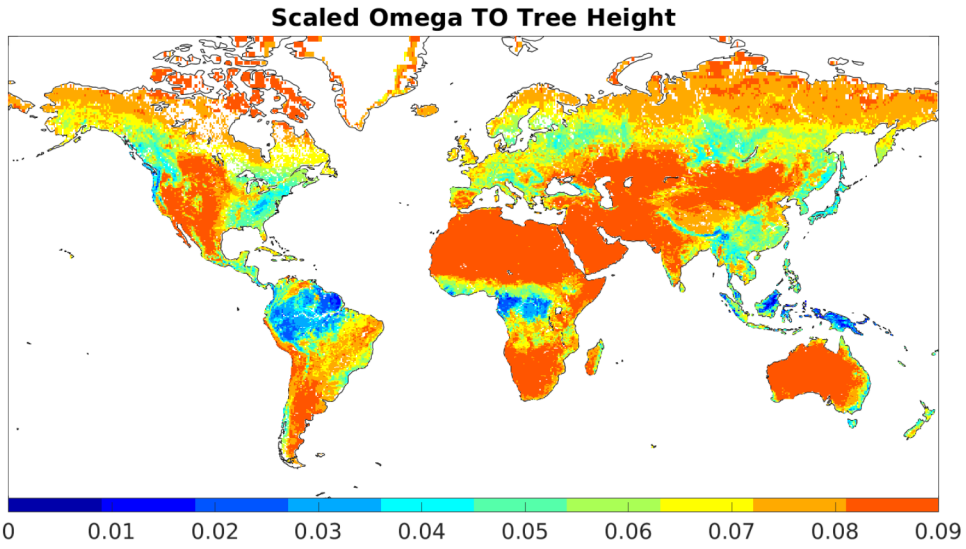


Figure 5.4: Map of ω for the C-band TO RTM based on the WSL Look-Up Tables

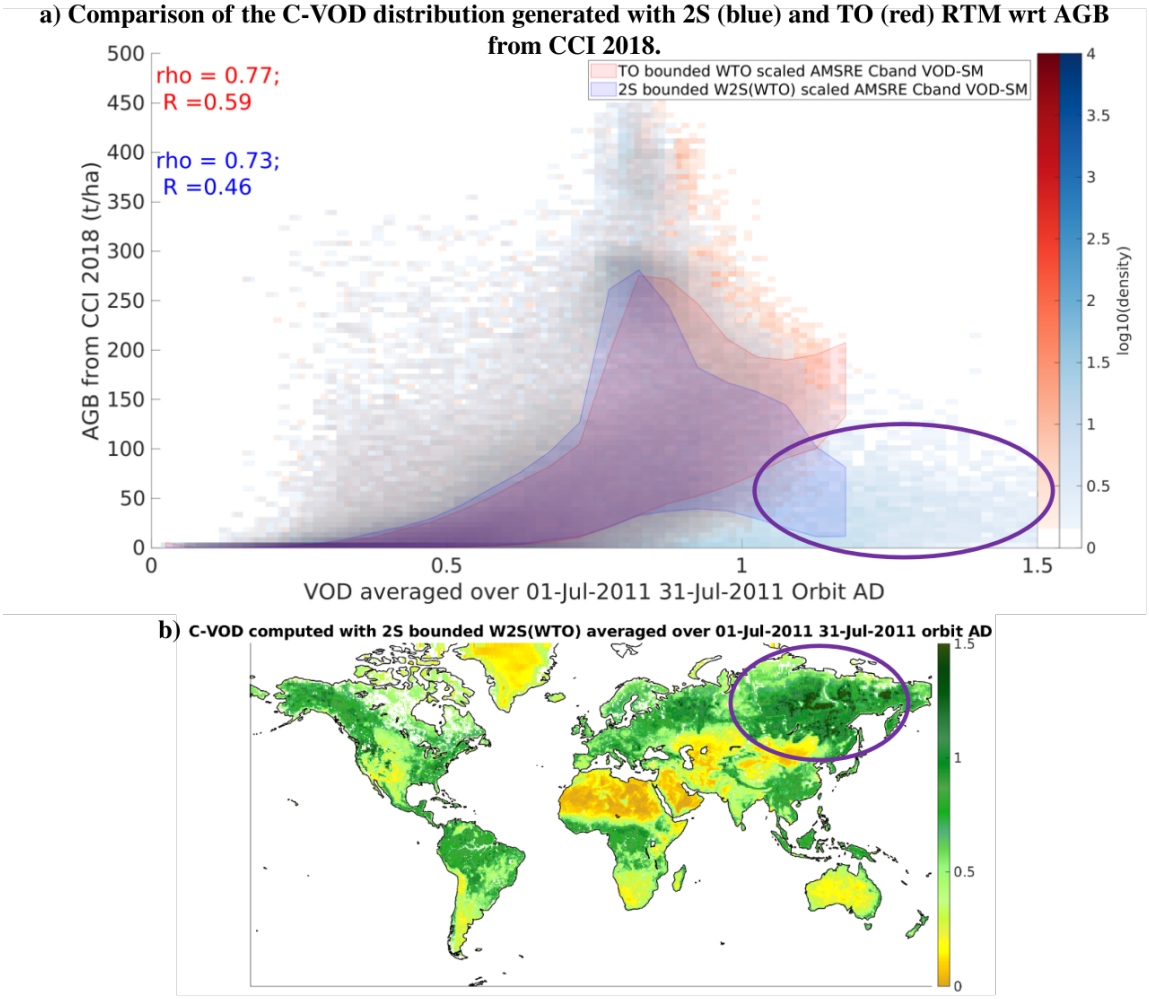


Figure 5.5: Density plot of the OSMOSE C-VOD for July 2011 against the AGB from CCI 2018. The red color refers to the use of the TO model to retrieve the VOD and the blue color refers to the use of the 2S model. In both cases, the couple (VOD/ACard) was retrieved, the retrieval was not constrained and the default values of ω were used (eq 23 of *Schwank and Naderpour* (2018) was used to convert from TO to 2S value). The shaded areas correspond to the interval in-between quantiles 16 and 84 and contain 68% of the distribution.

With the 2S configuration, a significantly large area over Eastern Russia exhibits unexpected high VOD. This area is highlighted in purple on panel (a) of Figure 5.5 and the corresponding points are also highlighted in purple on panel (b). These 2S VOD are higher than the VOD over the Tropics and do not seem realistic as they correspond to AGB values around 50-100 Mg ha^{-1}. The correlation also drops with the 2S RTM. Hence, the TO model seems to be better suited for the C-band. The same results apply to the X-band.

5.6.2 Influence of the retrieval parameters

The retrieval algorithm allows us to derive various parameters that drive the RTM, and the emission observed by the sensors. The most obvious ones are the SM or the dielectric constant for the soil layer, and the VOD for the vegetation layer, as described at the beginning of the section.

5.6.2.1 L-Band

For SMOS L-band, two couples of free parameters were tested:

- Retrieving (VOD, SM)
- Retrieving (VOD, ACard) and setting SM to the default initial value of 0.2 m^{-3}/m^{-3}

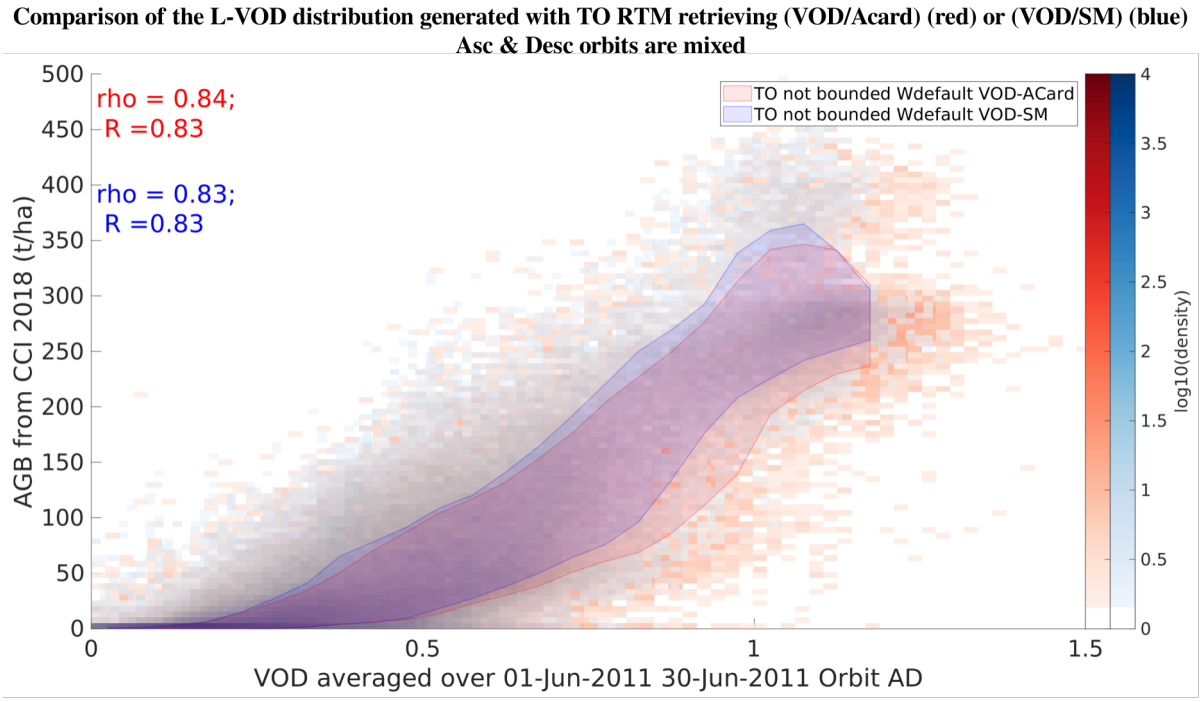


Figure 5.6: Density plot of the OSMOSE L-VOD for June 2011 against the AGB from CCI 2018. The red color refers to the retrieval of the (VOD/ACard) couple and the blue color to the (VOD-SM) couple. The TO retrieval was not constrained and the default values of ω were used. The shaded areas correspond to the interval in-between quantiles 16 and 84 and contains 68% of the distribution.

Figure 5.6 shows that retrieving (VOD/ACard) or (VOD/SM) does not significantly impact the VOD correlation with the AGB from CCI 2018. It is not surprising as the VOD are monthly averaged, which probably smooths the small differences that may occur. The derived Acard is the dielectric constant, which the SM mainly drives at L-band. SM and Acard are in that case linked and lead to similar VOD. All correlation coefficients are around 0.83. The dispersion, highlighted by the colored shaded area, is lower when using the classical (VOD/SM) couple. The (VOD/SM) couple is then considered in the OSMOSE processing chain for L-band.

5.6.2.2 C-Band

The same couples (VOD/SM or VOD/ACard) were tested for the C-band. As the observing frequency is higher than L-band, there was a possibility that the C-band observations were not sensitive to the ground underneath a vegetation layer, and that retrieving ACard would be more relevant than retrieving SM. Like for L-band, retrieving Acard rather than SM does not impact the correlation of the VOD with the AGB from CCI 2018. (VOD/SM) will hence be retrieved for the C-band. This also ensures homogeneity with L-band retrieval.

5.6.3 Influence of constraining the retrieval with upper and lower boundaries

Constraining the retrieval does not affect the L-VOD correlation to the AGB from CCI 2018. The processing chain is unconstrained for the L-band.

For the C-band, constraining the retrieval slightly improves the VOD Spearman correlation with the AGB (Pearson remains about the same). The dispersion of the C-VOD wrt AGB marginally decreases for VOD

values lower than 0.7 when setting lower and upper boundaries to the free parameters. Also, constraining the retrieval greatly improves the convergence between the demo chain (coded in matlab and hosted at CESBIO) and the operational chain (coded in Python and hosted at Magellium). In order to ensure results as close as possible to the demo trials, the retrievals are constrained between [0-2] for the VOD and [0-1] for SM for the C and X bands.

5.6.4 Influence of the ω and Hr parametrization

Another important aspect of the processing chain is the parametrization of the scattering albedo (ω and the roughness of the soil, Hr). This test was only conducted for the C and X bands as the ω and Hr parametrization for L-band had already been extensively studied and optimized, as described in the literature (*Fernandez-Moran et al.* (2017) and *Wigneron et al.* (2017)).

For this test, the OSMOSE retrieval platform was run with fixed values for ω and Hr. For the first runs, the ω values spanned [0.01-0.15] and the Hr values spanned [0 0.15 0.3 0.4]. The AMSRE C and X band VOD were derived for the month of July 2011 for ascending and descending orbits. Then, the temporal and spatial mean RMSE between the modeled (TB_{mod}) and the measured TB (TB_{meas}) was computed. This single mean RMSE value allowed us to compare the performance of the different (ω/Hr) configurations. The results of these tests for C-band over the global region are displayed in Figure 5.7. This figure shows that Hr has little influence on the RMSE compared to the value of ω. The RMSE slightly decreases from ω=0.01 to ω=0.05. The RMSE then increases with increasing values of ω. The configuration ω=0.05 and Hr=0.15 is selected for the C-band archive. The test was also conducted for the month of September 2011 and led to the same conclusion.

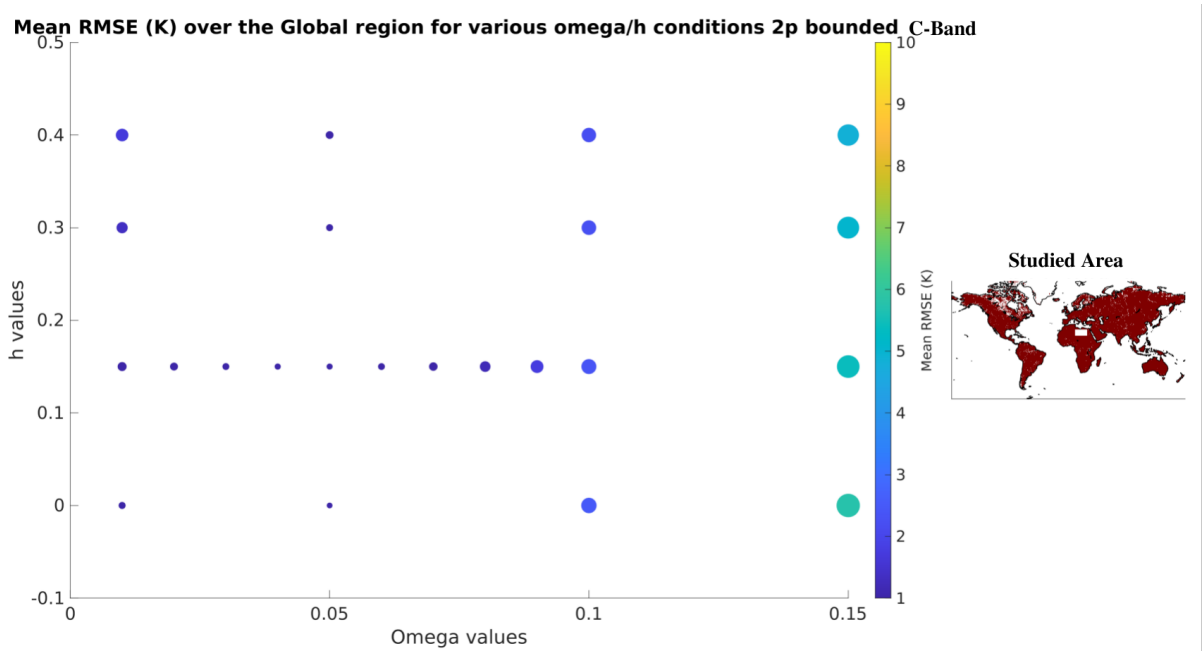


Figure 5.7: Mean RMSE between C-band $TB_{mod}TB_{meas}$ over the global region for different ω and Hr couples and a 2 parameters (VOD-SM) retrieval. The RMSE was averaged over the course of July 2011, mixing ascending and descending orbits

The same test for the X-band leads to the identical ω and Hr configuration as for the C-band.

5.6.5 Influence of retrieving one or two parameters

As AMSRE and AMSR2 only have one incidence angle, only two TB are available per observation, one for horizontal and one for vertical polarizations. Moreover, these two TB are not completely independent. Retrieving two parameters with two correlated TB may lead to erroneous results. This is why tests with a one-parameter retrieval were conducted with the same configuration described in the previous section. The results of these tests for the global region and the C-band are displayed in Figure 5.8. When comparing this figure with Figure 5.7, it is clear that a one-parameter retrieval leads to higher RMSE between the C-band TB_{mod} and TB_{meas}. The same conclusion applies to the X-band. Hence, a two-parameter retrieval is adopted for the C and X-band archives. As a side note, Figure 5.8 is concordant with Figure 5.7 and shows the little influence of the value of Hr compared to the value of ω. Again, the lowest global RMSE is achieved with ω=0.05.

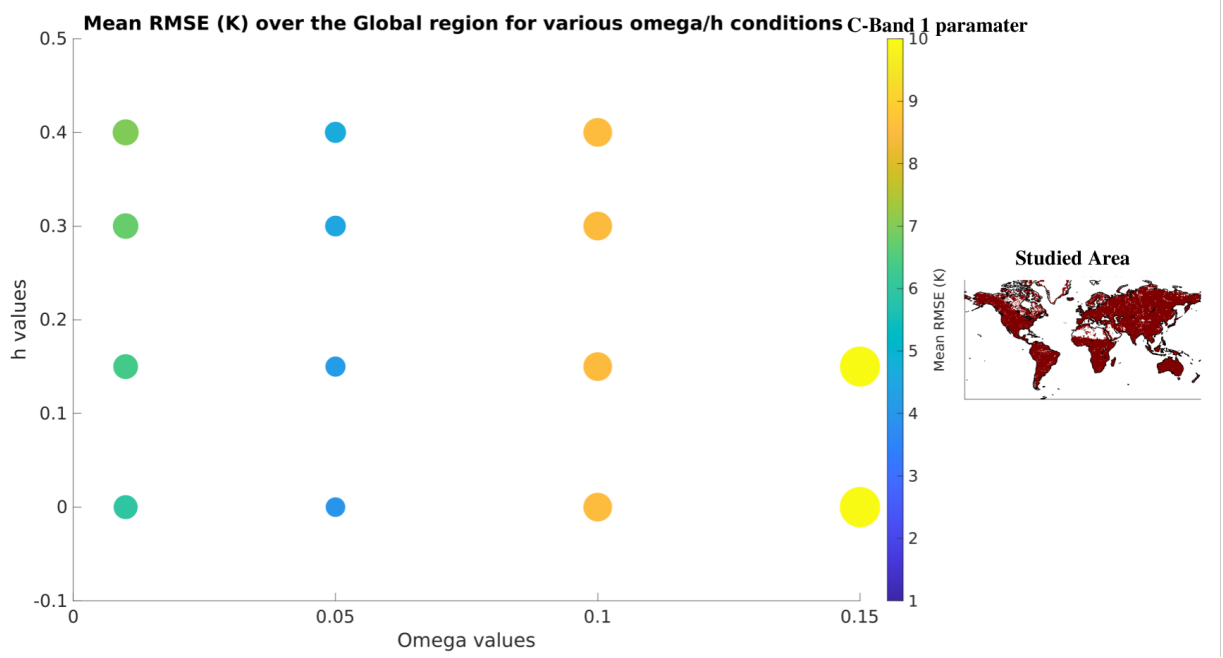


Figure 5.8: Mean RMSE between C-band $TB_{mod}TB_{meas}$ over the global region for different ω and Hr couples and a 1 parameter (VOD) retrieval. The RMSE was averaged over the course of July 2011, mixing ascending and descending orbits

5.7 Validation of VOD

5.7.1 versus datacube2

Datacube 2 is a collection of datasets from visible sensors, projected on the ESA Grid version 2 (*Brodzik et al.* (2012a, 2014)). Refer to chapter 8 for a complete description of the datacube2.

5.7.1.1 L-Band

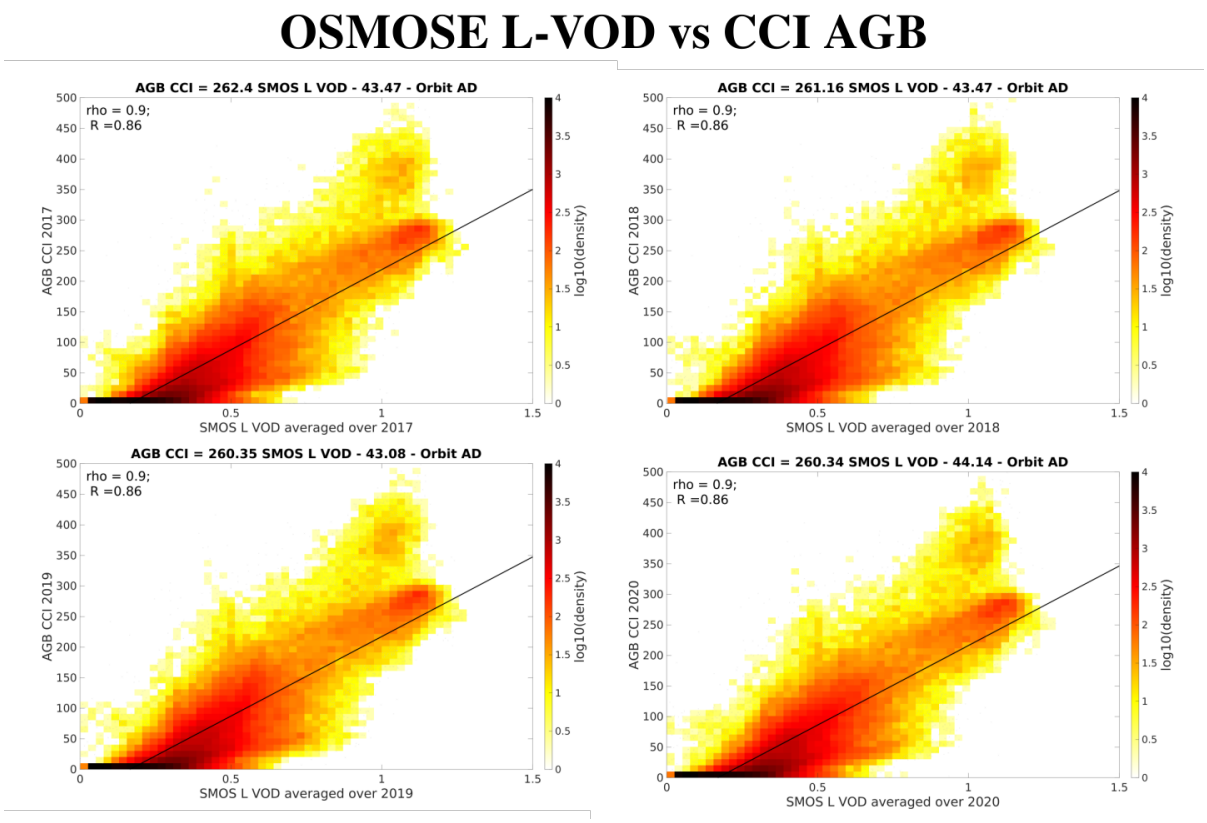


Figure 5.9: Density plot of the yearly L-VOD from OSMOSE wrt the CCI AGB for the years 2017-2020

Figure 5.9 and Table 5.5 show that the OSMOSE L-VOD is very well correlated to the CCI AGB for all years. The relation is stable from one year to the other.

Table 5.5: R between the L-VOD from OSMOSE and AGB values from CCI for the years 2017-2020.

	2017	**2018***	**2019**	**2020**
R	0.86	0.86	0.86	0.86
Rho	0.9	0.9	0.9	0.9

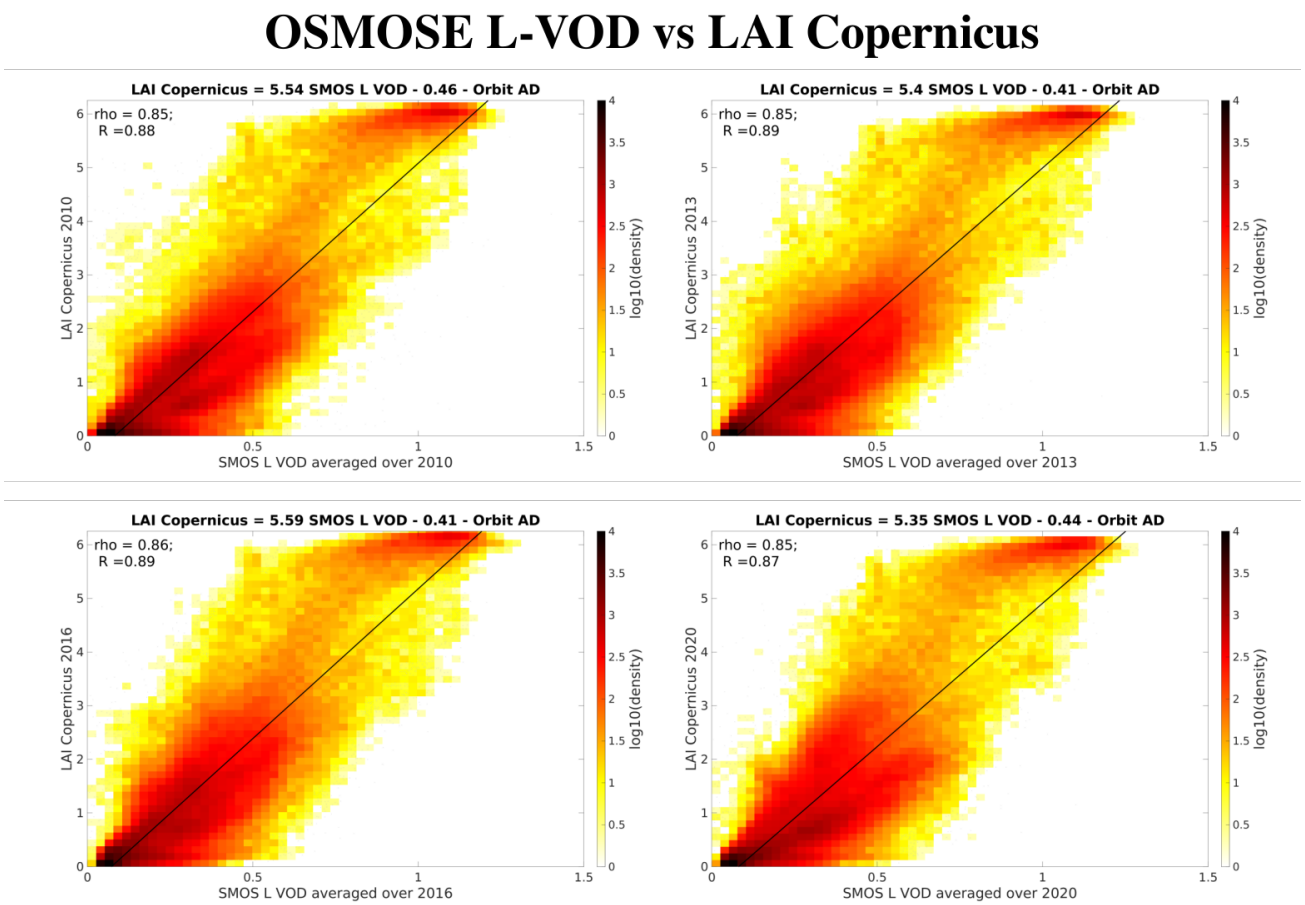


Figure 5.10: Density plot of the yearly L-VOD from OSMOSE wrt the LAI from Copernicus for the years 2010, 2013, 2016 and 2020

Figure 5.10 shows that the OSMOSE L-VOD is very well correlated to the LAI for all years. Only 4 years are displayed in Figure 5.10, but the relationship is stable from one year to the other.

Even though they are not displayed in this technical note, regressions between the L-VOD and other optical indices have been conducted:

- FAPAR (2010-2016): $0.82 \leq R \leq 0.86$ and $0.82 \leq rho \leq 0.86$
- FCOVER (2010-2020): $0.82 \leq R \leq 0.86$ and $0.81 \leq rho \leq 0.85$
- NDVI (2010-2020): $0.49 \leq R \leq 0.66$ and $0.51 \leq rho \leq 0.7$

5.7.1.2 C-Band

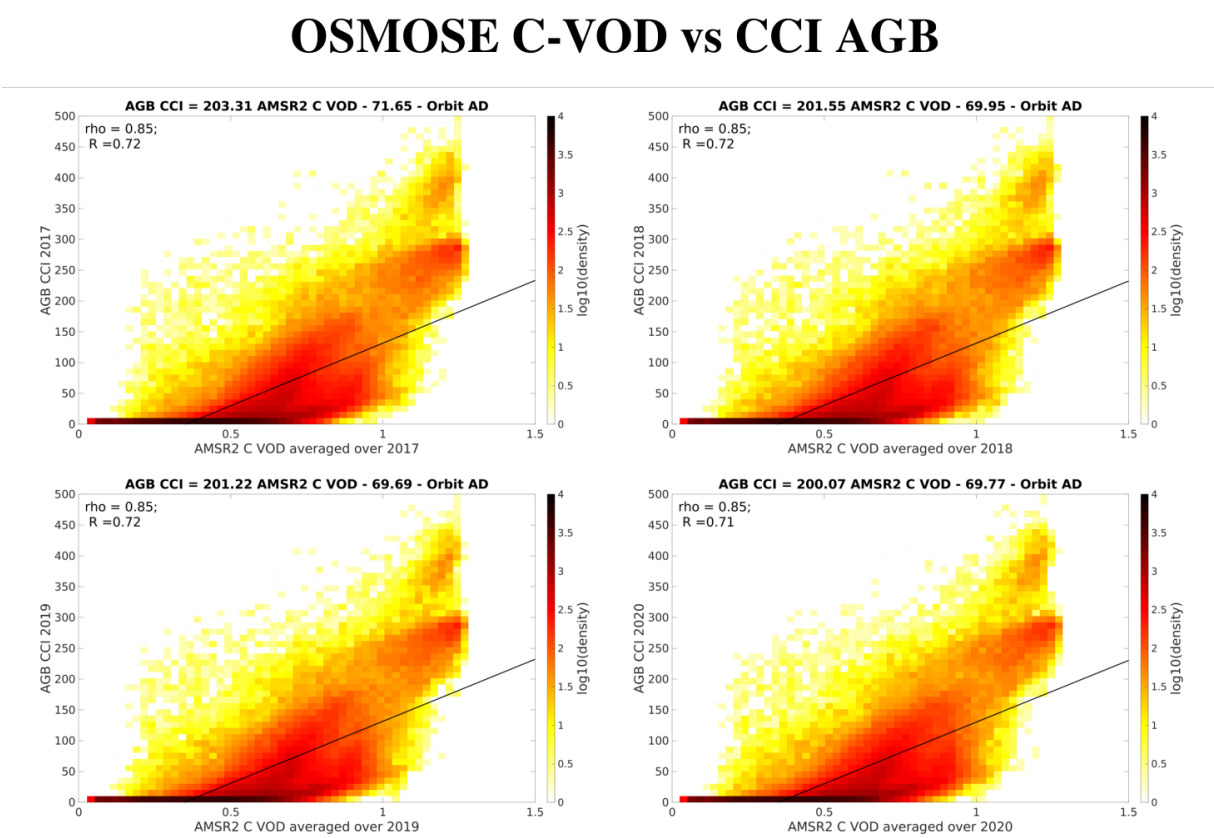


Figure 5.11: Density plot of the yearly C-VOD from OSMOSE wrt to the AGB from CCI for the years 2017-2020

Figure 5.11 shows that the OSMOSE C-VOD is well correlated to the AGB from the CCI for all years and the relation is stable from one year to the other. In agreement with the literature (*Rodríguez-Fernández et al.* (2018)), the general correlation with the AGB is not as good as with the L-VOD (see Figure 5.9).

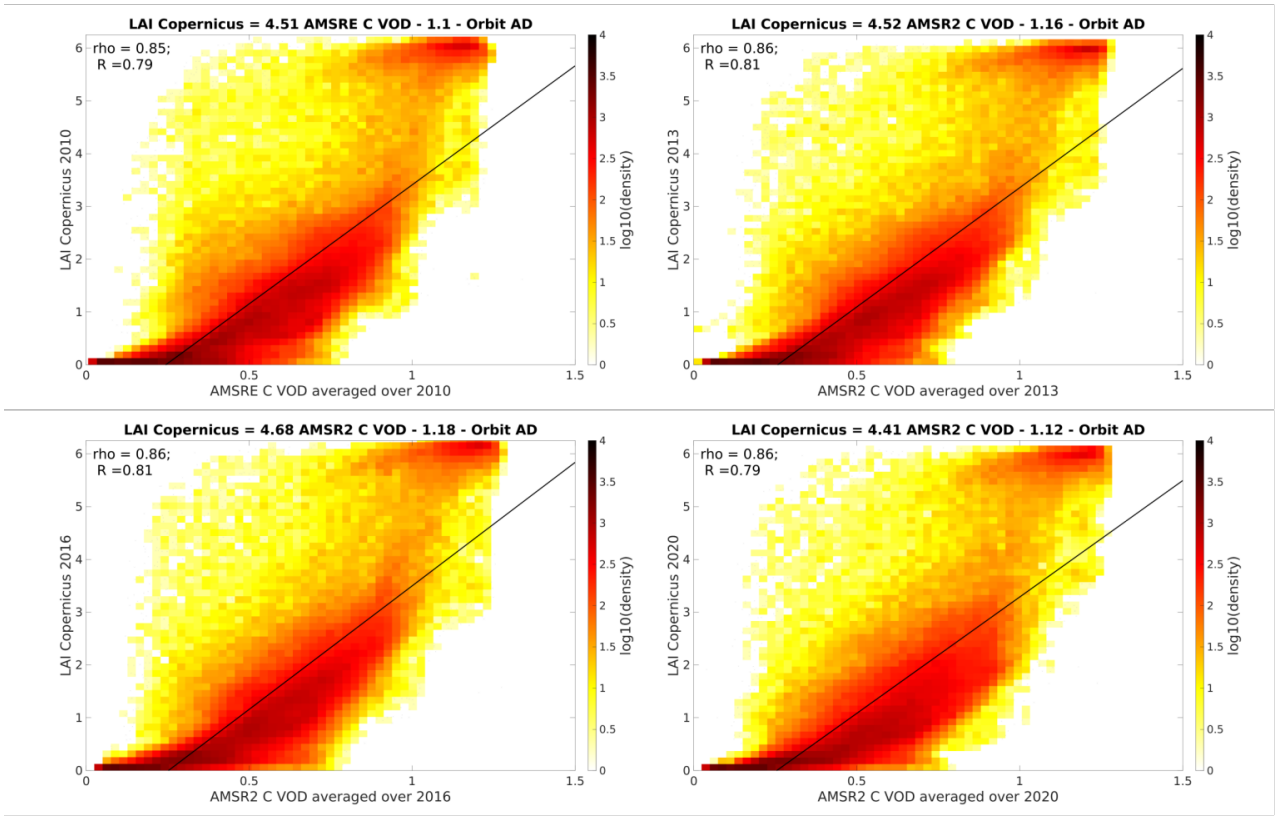


Figure 5.12: Density plot of the yearly C-VOD from OSMOSE wrt to the LAI from Copernicus for the years 2010, 2013, 2016 and 2020

Figure 5.12 shows that the OSMOSE C-VOD is very well correlated to the LAI for all years. Only 4 years are displayed in Figure 5.12 but the relationship is stable from one year to the other. The linear correlation R is not as good as the one with the LVOD but the spearman correlation coefficients are very close (see Figure 5.10).

Even though they are not displayed in this technical note, regressions between the C-VOD and other optical indices have been conducted:

- FAPAR (2002-2016): $0.84 \leq R \leq 0.88$ and $0.84 \leq rho \leq 0.87$
- FCOVER (2002-2020): $0.83 \leq R \leq 0.86$ and $0.83 \leq rho \leq 0.86$
- NDVI (2002-2020): $0.6 \leq R \leq 0.76$ and $0.62 \leq rho \leq 0.77$

5.7.1.3 X-Band

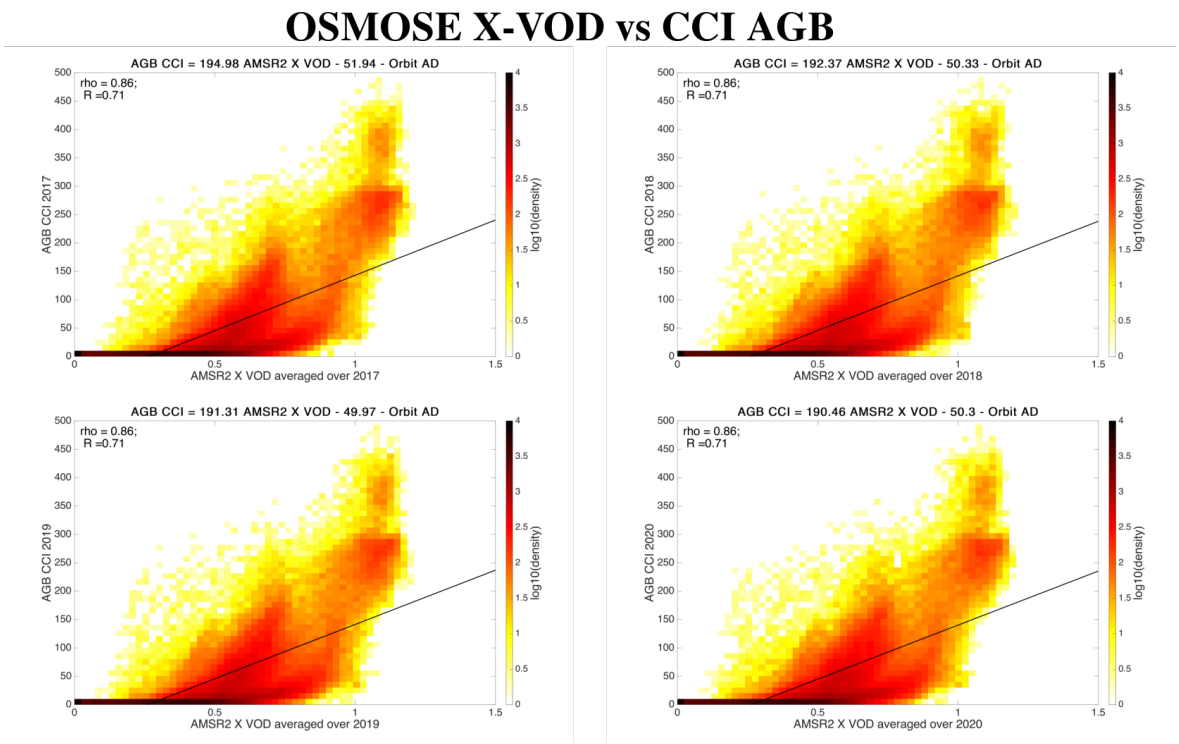


Figure 5.13: Density plot of the yearly X-VOD from OSMOSE wrt the AGB from CCI for the years 2017-2020

Figure 5.13 shows that the OSMOSE X-VOD is well correlated to the AGB from the CCI for all years and the relation is stable from one year to the other. In agreement with the literature (*Rodríguez-Fernández et al.* (2018)), the general correlation with the AGB is not as good as with the L-VOD (see Figure 5.9) but is very close to the correlation of the C-VOD with AGB (see Figure 5.11).

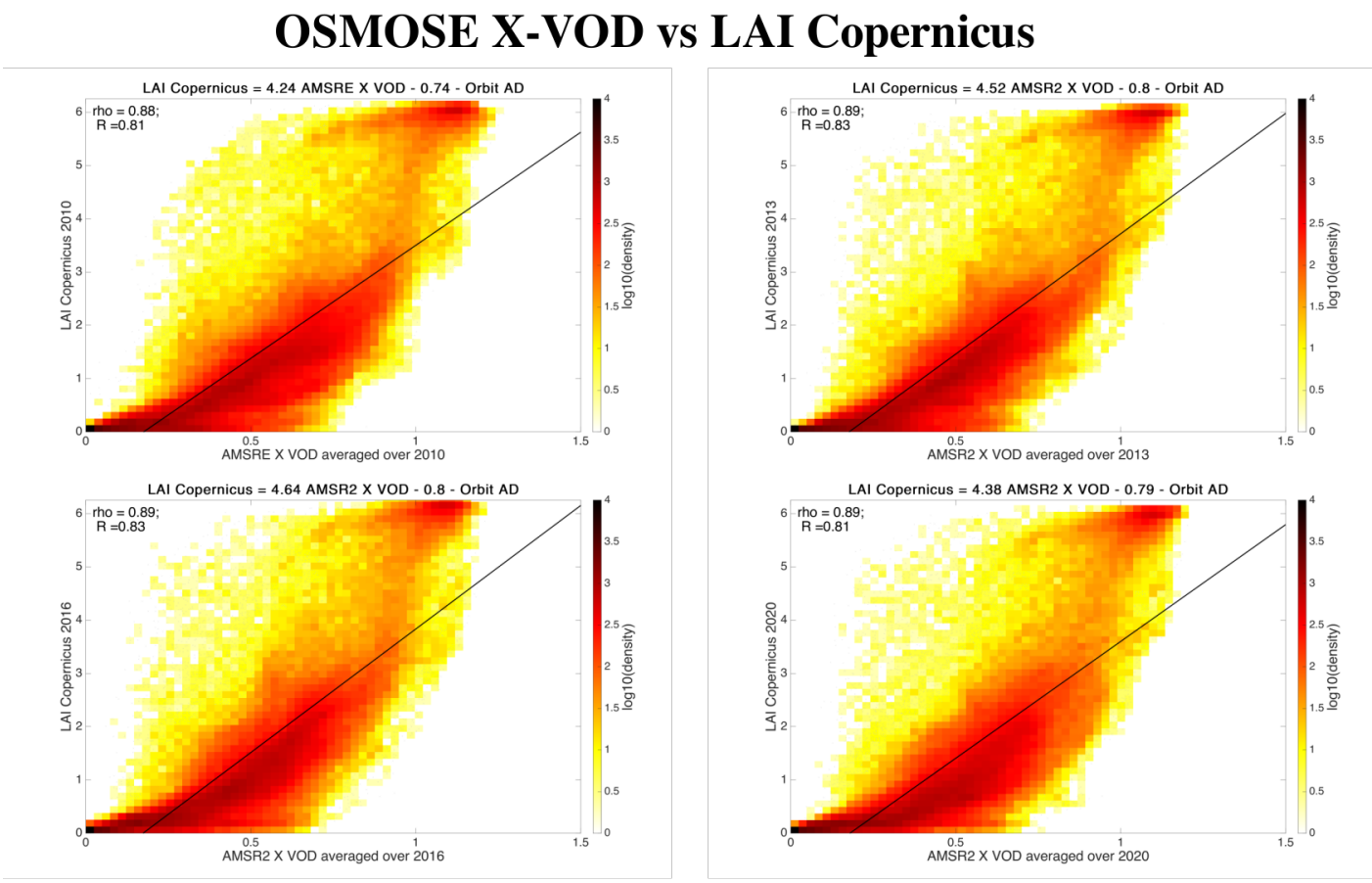


Figure 5.14: Density plot of the yearly X-VOD from OSMOSE wrt the LAI from Copernicus for the years 2010, 2013, 2016 and 2020

Figure 5.14 shows that the OSMOSE X-VOD is very well correlated to the LAI for all years. Only 4 years are displayed in Figure 5.14 but the relationship is stable from one year to the other. The linear correlation R is not as good as the one with the LVOD but the spearman correlation coefficients are very higher (see Figure 5.10). In general, the correlations (R and Rho) are better than the C-VOD ones (see Figure 5.12).

Even though they are not displayed in this technical note, regressions between the X-VOD and other optical indices have been conducted:

- FAPAR (2002-2016): $0.86 \leq R \leq 0.90$ and $0.86 \leq rho \leq 0.89$
- FCOVER (2002-2020): $0.87 \leq R \leq 0.89$ and $0.85 \leq rho \leq 0.89$
- NDVI (2002-2020): $0.66 \leq R \leq 0.79$ and $0.68 \leq rho \leq 0.8$

5.7.2 versus VODCA

The version 2 (v2) of the VODCA dataset *Zotta et al.* (2024) was just released. In this new version of the dataset, the C, X and Ku bands were merged into one daily data archive. It includes the use of AMSRE and AMSR2 descending orbits processed in the LPRM. The L-band created from SMOS and SMAP ascending orbits processed in the LPRM algorithm, is distributed as a 10-day composite data archive. The full data archive is provided on a 0.25^o grid, close to the EASE 2 grid used in the OSMOSE project. The VODCA v2 data archive was reprojected to the OSMOSE grid.

The yearly averaged OSMOSE VOD were compared to the yearly averaged VODCA dataset.

5.7.2.1 L-band

Figure 5.15 shows that the correlation between OSMOSE and VODCA for the L-band is excellent and uniform over the years. The R (Pearson coefficient) and Rho (Spearman coefficient) correlation coefficients are over 0.89 for all years between 2010 and 2020.

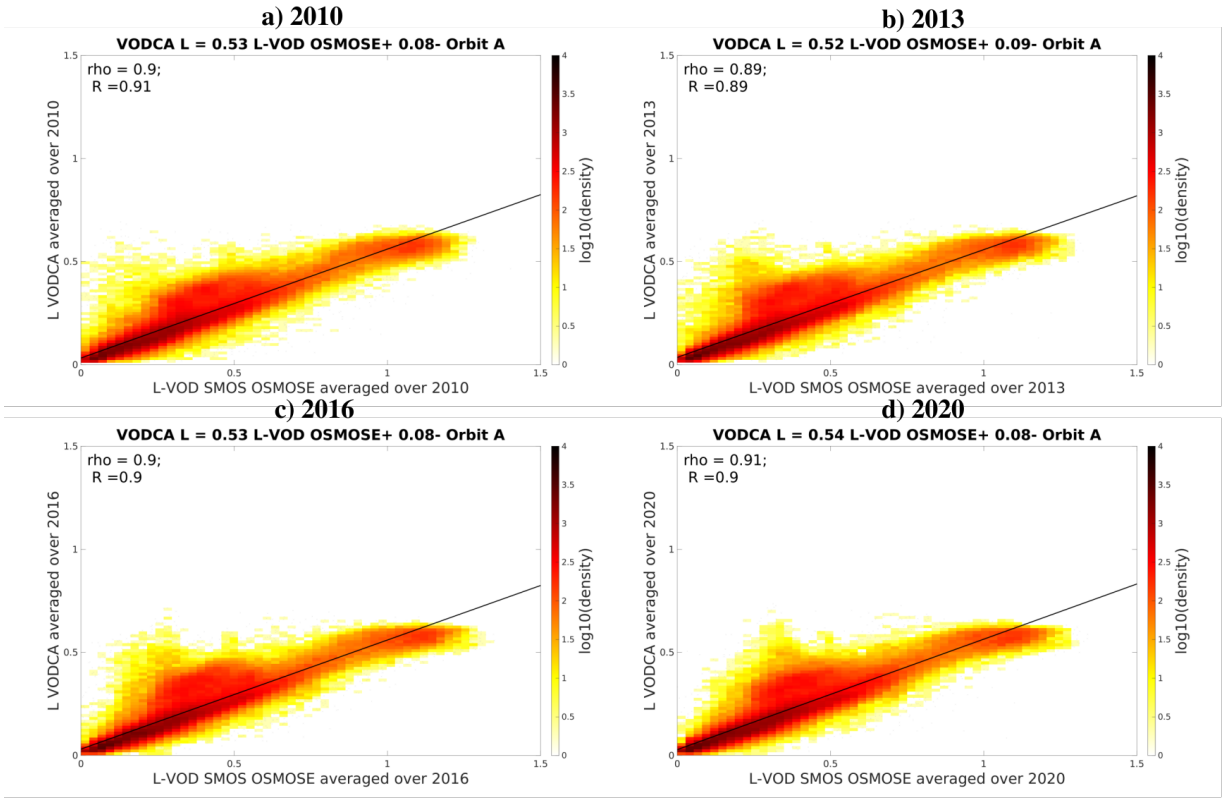


Figure 5.15: Density plot of the yearly L-VOD from OSMOSE wrt to VODCA v2 for (a) 2010, (b) 2013, (c) 2016 and (d) 2020

Generally, VODCA VOD is lower than OSMOSE VOD, except for a region where the two have more similar values (i.e. 0.2-0.5, above the linear fit black line in Figure 5.15). These points are mainly located over boreal forests, and further investigations are needed to understand better these differences. The VODCA VOD also saturates at a lower value (∼0.8) compared to the OSMOSE VOD which saturates at around 1.2 being more able to capture the vegetation dynamic.

5.7.2.2 C -band

Figure 5.16 shows that the correlation between OSMOSE and VODCA for the C-band is also excellent and uniform over the years. The ratio between OSMOSE C-band and VODCA CXKu is close to 1. The R and Rho correlation coefficients are over 0.87 for all years between 2003 and 2020.

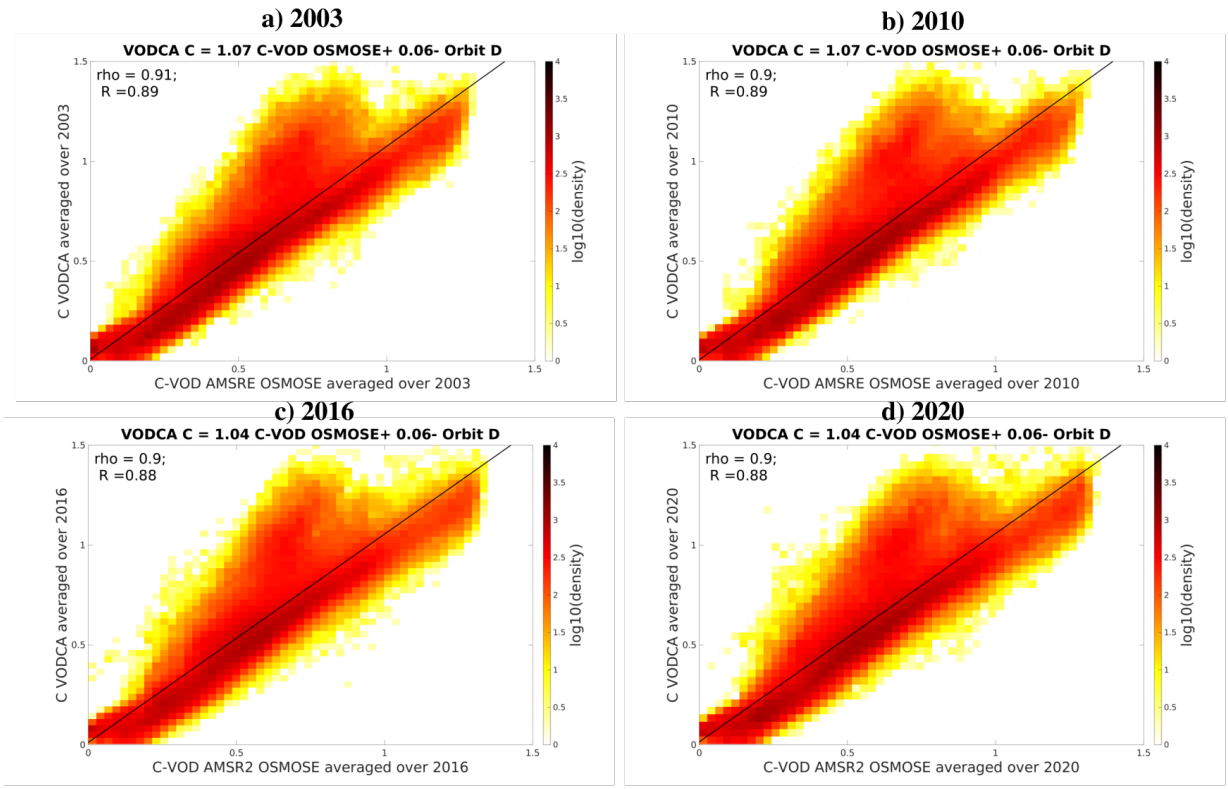


Figure 5.16: Density plot of the yearly C-VOD from OSMOSE wrt to VODCA CXKu v2 for (a) 2010, (b) 2013, (c) 2016 and (d) 2020

5.7.2.3 X -band

Figure 5.17 shows that the correlation between OSMOSE and VODCA for the X-band is also excellent and uniform over the years. The ratio between OSMOSE X-band and VODCA CXKu is close to 1. The R and Rho correlation coefficients are over 0.88 for all years between 2003 and 2020.

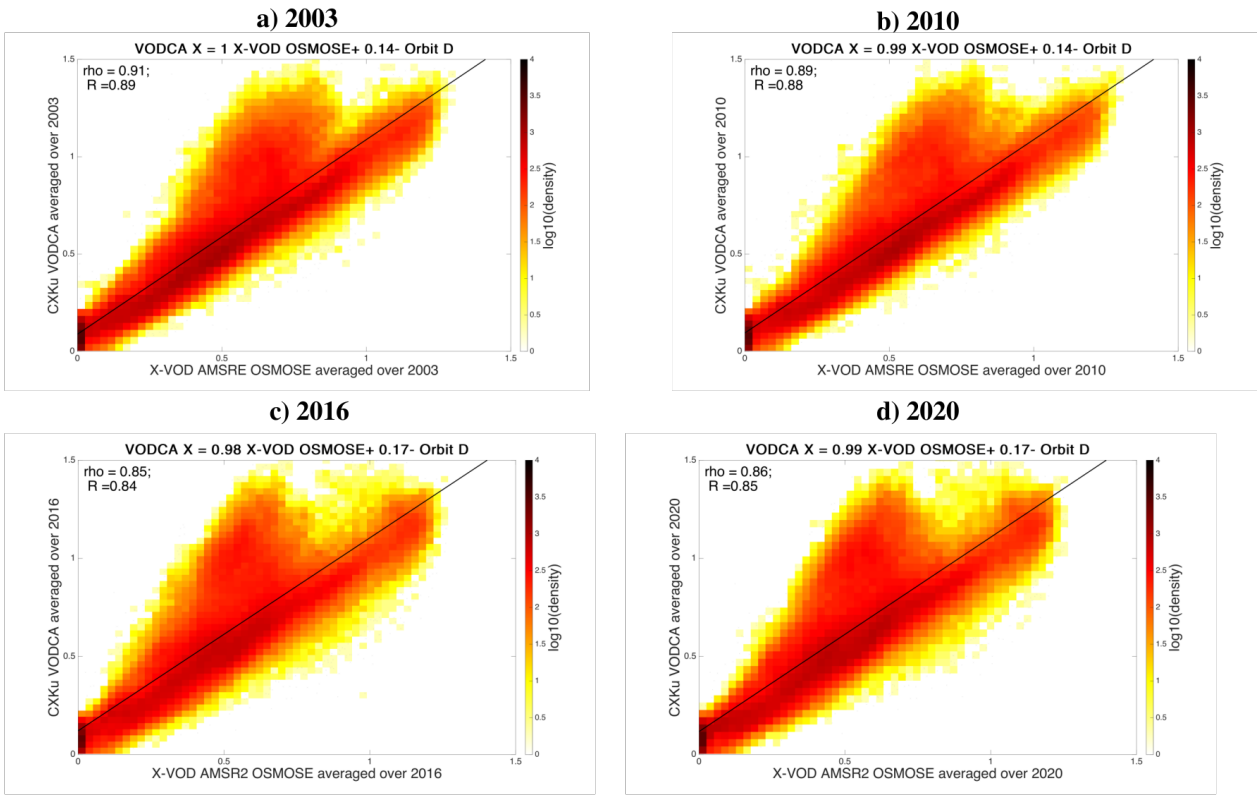


Figure 5.17: Density plot of the yearly X-VOD from OSMOSE wrt to VODCA CXKu v2 for (a) 2010, (b) 2013, (c) 2016 and (d) 2020

A similar evaluation is conducted per land cover classes. Table 5.6 gathers all the correlation coefficients per IGBP class for OSMOSE L and C VOD and VODCA L and CXKu VOD.

Table 5.6: Correlation Coefficient (R) between the CCI AGB (year 2018) and VODs from OSMOSE (L/C/X) and VODCA.

	OSMOSE L-VOD	VODCA L-VOD	OSMOSE C-VOD	OSMOSE X-VOD	VODCA CXKU-VOD
Evergreen needleleaf forests	**0.58**	0.42	0.45	0.56	0.50
Evergreen broadleaf forests	**0.71**	0.66	0.55	0.55	0.52
Deciduous needleleaf forests	0.43	**0.53**	0.3	0.44	0.34
Deciduous broadleaf forests	**0.44**	0.3	-0.18	-0.04	0.29
Mixed forests	**0.49**	0.15	0.18	0.36	0.16
Closed shrublands	0.52	**0.68**	0.08	0.42	0.66
Open shrublands	0.56	**0.59**	0.51	0.49	0.54
Woody savannas	**0.54**	0.53	0.38	0.42	0.44
Savannas	**0.74**	0.72	0.52	0.65	0.62
Grasslands	**0.54**	0.52	0.44	0.48	0.47
Permanent wetlands	0.43	**0.52**	0.32	0.39	0.51
Croplands	0.48	**0.50**	0.39	0.37	0.50
Urban	0.50	**0.54**	0.29	0.16	0.38
Cropland and vegetation mix	0.60	**0.61**	0.4	0.44	0.57
Snow Ice	**0.67**	0.20	0.34	0.41	0.04
Barren	**0.16**	0.14	0.07	0.09	0.12

5.7.3 Versus Wang X-VOD

The OSMOSE X-VOD produced with the TO model with ω=0.05 and Hr=0.15 using a constrained 2 parameter retrieval has been compared with the estimation available from *Wang et al.* (2024). In this paper, the authors conducted a regression between the 2017 X-VOD and the AGC from CCI 2017 over Africa. Figure 5.18 shows that the OSMOSE X-VOD averaged in 2017 over Africa shows a similar relationship with AGB from CCI as what *Wang et al.* (2024) found.

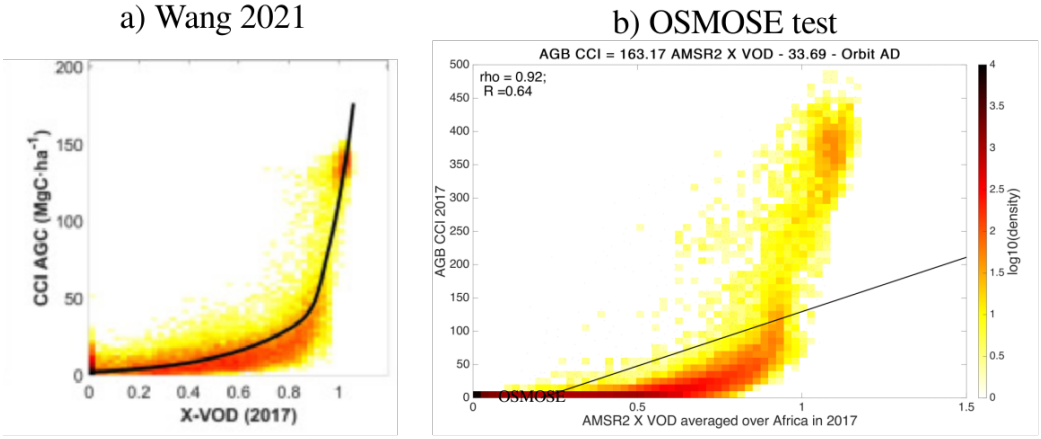


Figure 5.18: (a) Density plot of the 2017 X-VOD from Fig1 of *Wang et al.* (2024) wrt AGC CCI over Africa. (b) Density plot of the July 2011 OSMOSE X-VOD wrt to AGB CCI over Africa.

5.8 AMSRE/AMSR2

AMSR2 is the follow-up mission of AMSRE and there is a nine-month gap between both sensors. One of the objectives of the OSMOSE project being to build a 20-year time series of C and X VOD, the continuity of the AMSRE/AMSR2 data needs to be assessed and potentially corrected. Several studies have already dealt with AMSRE/2 continuity and put up different methodologies. *Wang et al.* (2021b) established a linear relationship between the AMSRE and AMSR2 TB (in H and V polarization respectively) over stable and homogeneous land covers (Evergreen Broadleaf forest and snow and ice). This relationship was then applied globally to calibrate the AMSRE TB to the AMSR2 ones. Finally, the calibrated TB were in turn processed by the X-MEB (*Wang et al.* (2021b)) to produce a consistent 20-year-long X-VOD time series. The authors of *Moesinger et al.* (2020a) and *Zotta et al.* (2024) worked directly on the VOD and employed CDF matching techniques to scale AMSR2 to AMSRE. The following sentences are directly extracted from *Zotta et al.* (2024): *VODCA v1 scaled AMSR2 to TMI if enough overlap was available or directly to AMSR-E, without temporal overlap, above and below 35° latitude N and S, respectively Moesinger et al. (2020a). This led to spatial inconsistencies in VODCA v1 [...]. Therefore, we changed the approach in VODCA v2 and used SSMI F17 Ku-band observations (scaled to AMSR-E X-band) as reference to bridge the gap between AMSR-E and AMSR.*

For the OSMOSE project, the bias and scaling are studied and performed directly on the C and X-VOD of AMSRE and AMSR2 and the datacube #2 is used to check the consistency of the correction.

5.8.1 Assessment of the AMSRE/AMSR2 bias for the C-band

Since there is no overlap between AMSRE and AMSR2, the continuity of the VOD between both sensors was studied by taking into account the last year and a half of AMSRE (1st Jan 2010 to 27th Sep 2011) and the first year and a half of AMSR2 (3rd Jul. 2012 to 31st Dec. 2014), following the method described in *Moesinger et al.* (2020a). The mean of the C-VOD was computed for both sensors over their respective period. The difference between the mean VOD is then computed. This difference is referred to as a bias even though it does not purely represent a bias as there is no overlap between the sensors.

To better understand the bias between AMSRE and AMSR2 C-VOD, it is compared to the mean difference of various optical indices from the datacube #2 (FAPAR, FCOVER, LAI, NDVI) and also to the mean difference of the OSMOSE L-VOD over the same period. The global spatial correlation between the C-VOD bias and the optical indexes/L-VOD is also computed. Results for FAPAR appear in Figure 5.19.

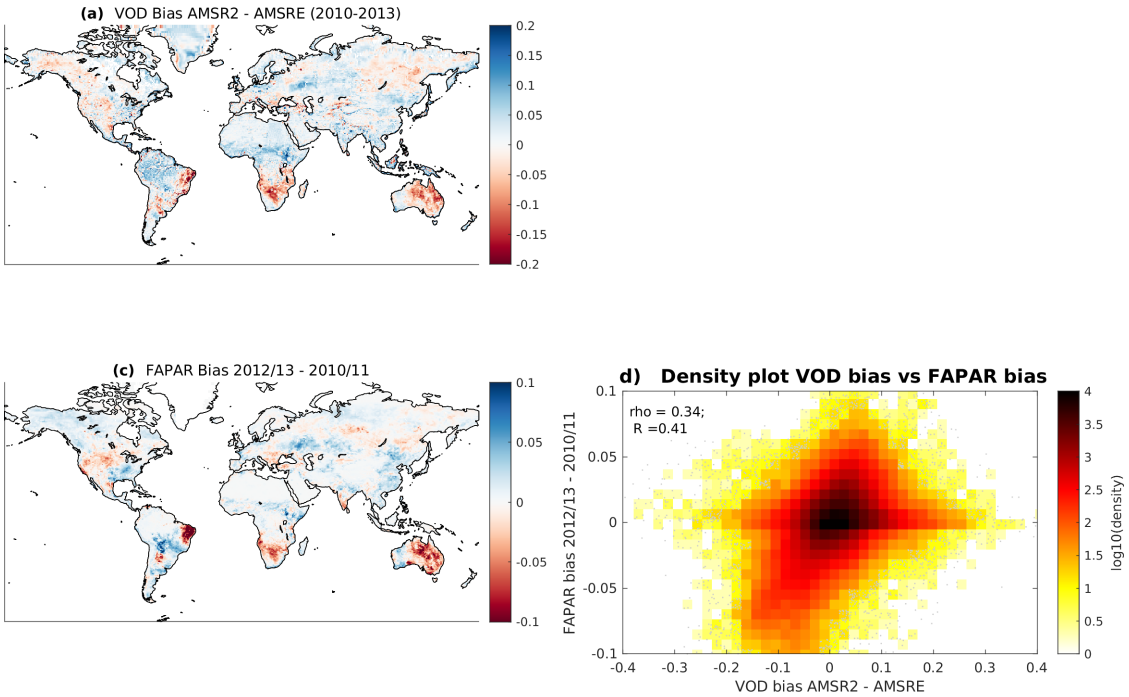


Figure 5.19: (a) Mean difference between the AMSRE C-VOD (01/01/2010-09/27/2011) and the AMSR2 C-VOD (07/03/2012-12/31/2013). (c) Mean difference of the FAPAR between 2012/2013 and 2010/2011. (d) Global spatial correlation between FAPAR and C-VOD bias.

Globally, the C-VOD bias is not correlated to any of the optical indices or L-VOD bias. R is lower than 0.41 for all cases and Rho is lower than 0.34. This indicates that globally, the C-VOD bias is not natural. Interestingly, the C-VOD bias is much more dynamic than the L-VOD bias. However, looking up Figure 5.19 and other indices show the same trend, 3 areas have similar C-VOD biases and optical/L-VOD biases: the eastern tip of Brazil, northern South Africa, and eastern Australia. For these three regions, the observed C-VOD bias is also observed in the optical indices/L-VOD bias as well, meaning that the C-VOD bias there is (at least) partly natural.

5.8.2 Correction of the AMSRE/AMSR2 bias

These comparisons highlight that, globally, the bias between AMSRE and AMSR2 needs to be corrected pixel-wise. One well-established method to correct a bias is the Cumulative Distribution Function (CDF) matching as described and applied in section III.B. of *Madelon et al.* (2022). The following sentences are directly extracted from *Madelon et al.* (2022): *The CDF is a specific way to describe the distribution of a discrete or continuous variable X. The purpose of this function is to give the probability that X will take a value less than or equal to a certain threshold. [...] CDF matching consists of transforming the CDF of one variable X (source data) to mimic that of another variable Y (reference data) by using a function f.*

For this study, AMSRE is rescaled to AMSR2 as AMSR2 is the most recent sensor. As a consequence, the source data are the last two years of AMSRE (28th Sept. 2009 to 27th Sept. 2011) and the reference data are the first two years of AMSR2 (3rd Jul. 2012 to 2nd Jul 2014). Different CDF matching functions can be used. The most simple one is a linear function using the mean and the standard deviation of the source and reference data. To use the linear CDF matching, the data time series need to be Gaussian. Several statistical tests (Chi-square goodness-of-fit test, the Lilliefors test *Lilliefors* (1967) and the Jarque-Bera test

Jarque and Bera (1987)) were applied on the AMSRE and AMSR2 C-VOD time series. These stastistical tests showed that the 2-year long time series could not be considered as Gaussian. Hence, the Polynomial fitting CDF matching from *Brocca et al.* (2011) was used to rescale AMSRE to AMSR2. *Madelon et al.* (2022) summarizes the principle: It *uses a function approximation technique to find the relationships between the source and the reference data CDF. First, the source and reference CDFs are computed with the number of percentiles equal to the number of SM observations in their respective time series. Second, the source CDF is linearly interpolated on the percentiles of the reference CDF, and the difference between the two is computed. Then, the difference is plotted against the interpolated source CDF, and a polynomial fit is performed. Finally, the polynomial function is used to compute the correction to apply to the source data. As mentioned in Aires et al. (2021), this method has the advantage to provide a continuous transformation, but it is more complex to control. Indeed, it can be subjected to instabilities and extreme events are not sure to be well preserved*.

A polynomial of degree three was used to rescale the C-VOD time series of each pixel. If there are fewer than 50 points in either the source or reference time series, the pixel is dropped (4 points in total).

5.8.3 Validation of the CDF matching

Time series of VOD with/without CDF matching correction are checked against SMOS L-VOD and optical indices from OSMOSE data cube 2. An example is displayed below:

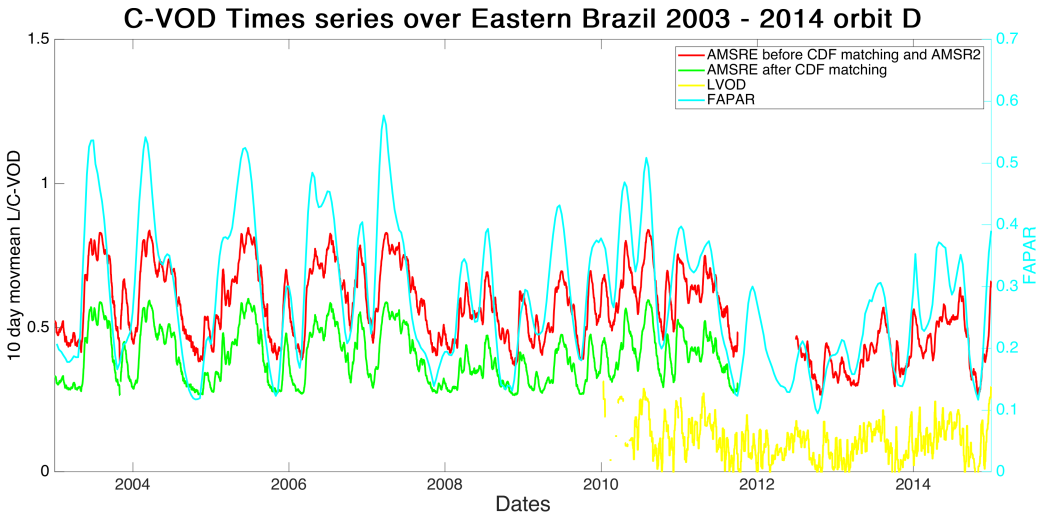


Figure 5.20: For the point [-11.41° N, -39.77°E] (eastern tip of Brazil): 2003-2014 time series of the AMSRE/AMSR2 C-VOD before CDF matching (red), 2003-2011 time series of AMSRE C-VOD after CDF matching with AMSR2 (green), 2010-2014 time series of the SMOS L-VOD (yellow), 2003-2014 time series of FAPAR from Copernicus (cyan). The L/C-VOD time series are 10 day moving averages.

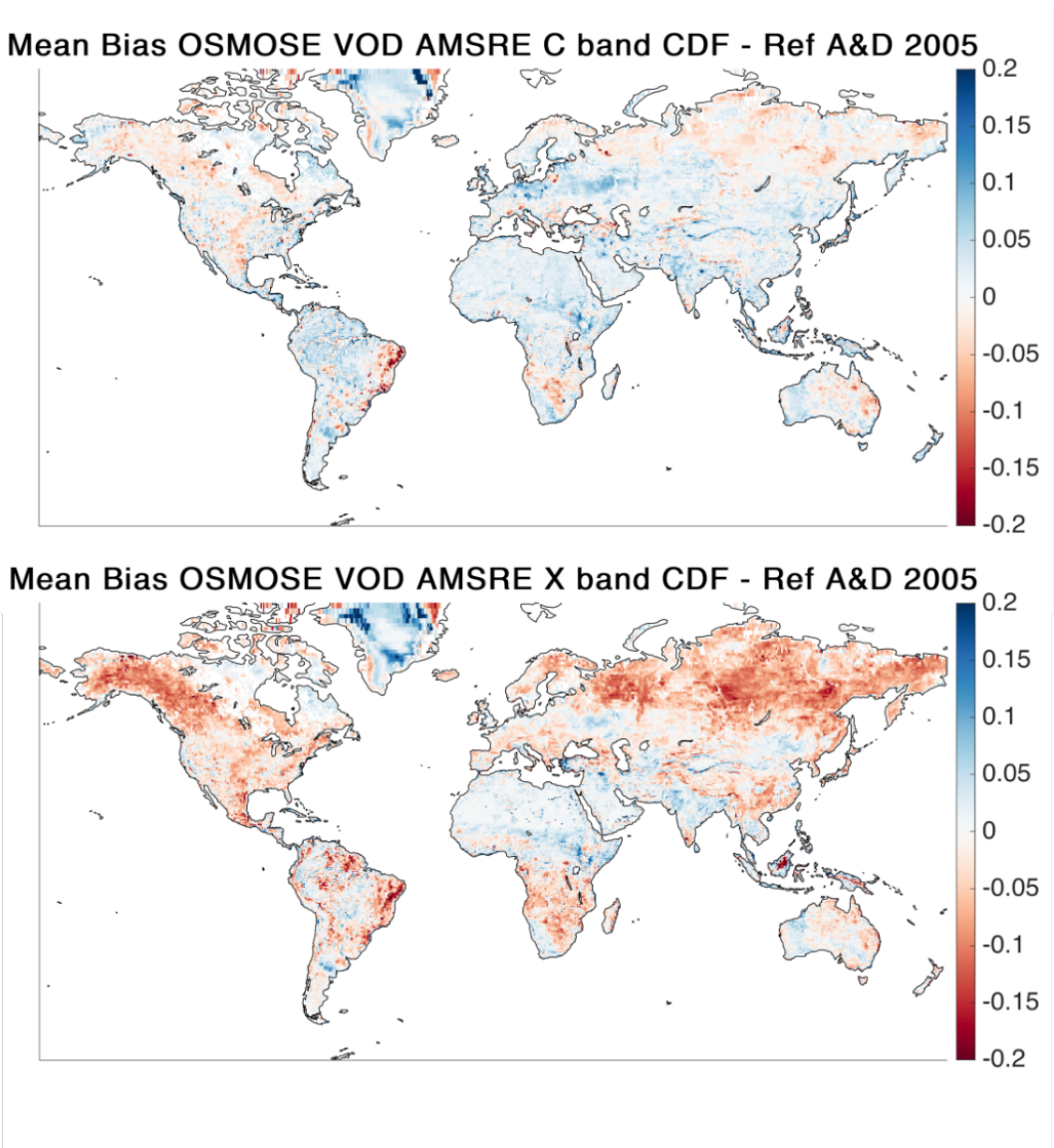


Figure 5.21: Yearly VOD Bias for AMSRE after minus before CDF matching for C (top) and X (bottom) band

Correction of the VOD from the AMSR sensors is mandatory. Figure 5.21 shows the significant differences when applying the CDF matching correction. For our analysis, we chose to correct the AMSRE VOD, as there is no absolute method to clearly state which sensor (i.e. AMSRE or AMSR2) provides the reference TB. The correction could also have been done at the TB level as in *Wang et al.* (2021a). It constitutes the most appropriate method. However, we chose to keep the "raw" VOD, so that we could perform an analysis with optical data from the datacube2. It seems that on a few nodes, the correction is not justified. One has to keep in mind that it may be the case for other grid nodes. Both VODs (i.e. not corrected and corrected with CDF matching) are then produced and delivered for the OSMOSE project, as one can further investigate the need to apply a correction, whether on AMSRE or AMSR2 sensors.

Chapter 6

Analysis of AGB (WP5)

6.1 Introduction

The purpose of this section is to detail the setup of the AGB estimation. Data from the SMOS, AMSRE and AMSR2 sensors are used to derive the AGB. The following represents the activities developed for the work package 5 of the project.

6.2 Data

6.2.1 L-Band Level 3 Brightness Temperature and VOD/SM data-sets

Level 3 TBs (L3TB) are provided in fourteen angle bins of 5º width (excluding bin number nine). Hereafter bin No.1 denotes incidence angles from 0 to 5º, bin No.2 from 5º to 10º and so on until bin No.14; exceptionally, the complementary bin number 9 is centered at 40º with a 2º width. L3TB TBs are provided in horizontal and vertical polarizations, hereafter H and V polarizations, respectively.

The Level 3 L-VOD/SM product (L3SM version 339), hereafter VOD (and SM), and TBs (version 330) are produced through the implementation of the L-MEB (L-Band Microwave Emission of the Biosphere) radiative transfert model *Wigneron et al.* (2007a) on the SMOS Level 3 retrieval algorithm *Al Bitar et al.* (2017a).

Both the L3SM and L3TB data sets are provided by the CATDS (Centre Aval de Traitement de Données SMOS) from 2010 until 2023 on a daily time step, and are projected on the EASE (Equal-Area Scalable Earth) grid version 2. For this study, only the observations corresponding to the years of the AGB reference maps (2010, 2017 and 2018) were used. TBs from the ascending orbits were used. To remove the dependence to seasonal variations of VWC on the signal, annual averages of each variable were calculated for the aforementioned years.

A daily filtering was performed on the TB and VOD/SM products before calculating the annual averages. The probability of appearance of Radio Frequency Interference (RFI) was used as a quality criterion. In the case of TBs, any observation whose probability of appearance of RFIs (number of RFIs detected/number of daily observations) by angle of incidence was greater than 0.7 was discarded. For VOD products, in addition to the RFI probability filter ($>$0.2), a χ^2 coefficient (which measures the goodness of fit from TB estimates and TB observations during the minimization process of the VOD/SM) lower than 2 was selected.

Additionally, and taking into account that the SMOS retrievals over coastal pixels tend to be strongly affected by the fraction of water in the observed scene (especially at high incidence angles), all observations one pixel away from the coasts (defined by the USGS land sea mask used by the SMOS algorithm) were removed.

6.2.2 C- and X-band VOD

Producing AGB with a multi-frequency approach requires a harmonized multi-frequency VOD dataset. The frequencies considered in the OSMOSE project are the L-band from SMOS and the C and X-bands from the Advanced Microwave Scanning Radiometer (AMSRE) and its follow-up mission AMSR2. The multi-frequency VOD dataset results from the processing of L, C and X-bands TB through a RTM (the so called $\tau - \omega$ model) which parametrization varies from one frequency to the other. For now, L and C-VOD have been produced in the frame of the OSMOSE project. The L-VOD dataset spans 2010-2020 and the C and X-VOD product span 2002-2020. The products and the retrieval chain are described in the Chapter 5.

6.2.3 Above Ground Biomass reference maps

Three AGB maps from ESA's CCI biomass project with global coverage (*Santoro et al.* (2021)) were used as vegetation biomass references for the years 2010, 2017 and 2018 (datasets version 3.0), from now on referred to as CCI2010, CCI2017 and CCI2018, respectively. Reference maps are provided with a spatial resolution of 100 m, and representing biomass density in Mg/ha. The CCI2010 is produced from ALOS-1 (PALSAR-1) and Envisat ASAR data-sets, the CCI2017-18 are produced from ALOS-2 (PALSAR-2) and Sentinel-1 observations of SAR backscatter. According to the validation protocol applied by the CCI biomass project (comparing the CCI maps with true AGB observations from forest field inventories), the CCI2017 and CCI2018 were in better agreement with AGB references than the one from 2010. Though the CCI2010 showed a lower agreement with field estimations, it still offers an approximation of the biomass for that reference year, and therefore it was used in this study. In spite of their possible uncertainties, the CCI AGB maps will be referred to as 'references', assuming that they are a good approximation to the actual state of vegetation biomass globally.

In addition to the maps from the CCI project, we also tested as reference the AGB map from Avitabile et al. (2016). The latter developed a 1-km resolution AGB map, specifically for the pantropical region, by merging two earlier AGB maps created by *Saatchi et al.* (2011) and *Baccini et al.* (2012), targeting the year 2010. This combination utilized a novel collection of field data and adjusted high-resolution AGB maps. These elements were standardized and combined, yielding nearly 15,000 reference AGB data points at a 1-km resolution. To ensure accuracy, a data merging method that included bias correction and weighted linear averaging was applied across specific zones.

For this study, all AGB reference maps have been re-gridded to the common grid of SMOS L3 products (EASE grid v2) by averaging the original 100 m pixels within the $\sim$25 km EASE grid cells.

6.2.4 Data for AGB evaluation

We used for AGB validation lidar-based AGB data, in situ AGB data and FAO data at the national scale. We also compared our maps with AGB maps from *Xu et al.* (2021).

6.2.4.1 Lidar-based AGB data

Four Lidar-based AGB dataset were used in this study for evaluating our AGB maps. The characteristics of these AGB maps are summarized in Table 6.1. They have been processed in the frame of the ESA CCI Biomass project by Nicolas Labrière from the Évolution et Diversité Biologique Laboratory, University Toulouse 3 Paul Sabatier, Toulouse, France. The four dataset were acquired in Australia, USA, Brasil and Indonesia, and include 499 lidar-based AGB maps at 100m pixel size, representing in total an area of 24 340 km^2, acquired from 2011 to 2019. This timeframe is particularly suited to our analysis, in which we provided maps from 2011 to 2022.

The work of Nicolas Labrière was similar for the Australian TERN (Terrestrial Ecosystem Research Network), the National Ecological Observatory Network (NEON), the Sustainable Landscape Brazil (SLB) and the Carbon Monitoring System (CMS) consisted in the processing of tree inventory and airborne LiDAR data. After data cleaning, tree inventory data included approximately:

- TERN: 28 800 living trees from 9 inventories at 5 different sites scattered across Australia. They were converted into plot- and subplot-level AGB estimates and associated uncertainty, for a total of 188 units (34, 1 and 153 of 1ha, 0.5ha and 0.25ha, respectively),
- NEON: 58 800 living trees from 126 inventories at 37 different sites scattered across the conterminous United States, Alaska, Hawaii and Puerto Rico. They were converted into plot-level AGB estimates, for a total of 3 386 plots of 0.04ha (20×20m square plots)
- SLB: 32,500 living trees from 31 inventories at 22 different sites scattered across the Brazilian Amazon. They were converted into plot-level AGB estimates and associated uncertainty, for a total of 625 plots ranging from 0.16ha (40×40m square plots) to 1ha (20×500m transects),

Regarding CMS data, tree-by-tree data was not accessible but plot-level information was, for a total of 104 plots ranging from 0.1ha (20×50m rectangular plots) to 1ha (40×250m transects) covering both drylands (n=82) and wetlands (n=22).

Then, for each of the four dataset, LiDAR-derived canopy height models (CHMs) were produced for all the lidar point cloud data. Mean top-of-canopy height (TCH) was calculated within the footprint of each tree inventory unit from the corresponding CHM, and a power law model was built between plot AGB and TCH using all the plots units. This model was applied to all LiDAR footprints, and AGB maps and associated uncertainty were produced by Nicolas Labrière.

Table 6.1: Lidar-based AGB data characteristics

Source	Country	Area in km²	Nb. lidar footprints	Nb. sites	Year
TERN	Australia	370	13	13	2012-2015
NEON	USA	21 650	125	-	2013-2019
SLB	Brasil	820	286	69	2011-2018
CMS	Indonesia	1 500	75	-	2014
	Total	24 340	499	-	2011-2019

6.2.4.2 In situ AGB data

We used for evaluation the AGB plots dataset from *Ploton et al.* (2020a), accessible online to all for free. This dataset consists in the AGB estimations and associated uncertainty for nearly 60,000 1-km pixels aggregated from nearly 100,000 ha of in situ forest management inventories for the 2000-early 2010s period. The data specifically targets five central African countries in the African humid tropical forests, providing detailed insights into this ecologically important region. The uncertainty on AGB estimations derived from 0.5-ha inventory plots (8.6-15.0%) is only moderately higher than the error obtained from 1-ha scientific sampling plots (8.3%). This uncertainty is acceptable, considering that larger plots (here 1-ha versus 0.5-ha) exhibit smaller uncertainties, and that management inventories present higher uncertainties that scientific sampling plots.

6.2.4.3 FAO AGB data

The FAO Forest Ressource Assessment (FRA) provides a worldwide analysis of forest resources, including data on forest extent, condition, and changes over time, with a specific section dedicated to forest biomass *FAO* (1988). FAO thus offers a comprehensive overview of the world's forest resources, with key insights into forest AGB. Estimates of AGB and carbon stocks are provided for forests defined as vegetation patches with a minimum tree cover of 10% and minimum area of 0.5 ha. It is important to note that the FAO estimates are based on reports provided by each country, and some countries have limited capacities to estimate their biomass stocks accurately. However, the comparison between our estimates and the FAO estimates provide an idea of the amount of biomass estimated nationally.

6.2.4.4 AGB maps from Xu et al. (2021)

Xu et al. (2021) calculated the above- and below-ground biomass of global terrestrial ecosystems on an annual basis, covering the years 2000 to 2019. Reference data, including approximately 100,000 plots coupled with airborne LIDAR data covering more than 1 Mha of tropical forests globally and satellite lidar survey detailing the height structure of global vegetation across over 8 million sample footprints, were used to feed a machine learning model. Reference data were converted into estimates of biomass using established allometric models and root-to-shoot ratio, then used as training data for the machine learning algorithm, together with microwave and optical satellite imagery collected from 2000 to 2019. Note that these maps were used for comparison purpose only.

6.3 Methods

6.3.1 AGB estimation from multi-angular L-band TBs

The proposed methodology for estimating AGB maps from ANN is based on a purely statistical relationship between inputs, in this case multiangular TBs, and targets, the CCI reference maps. Yet, each polarization of the TBs is grouped into 1 of 14 bins (according to their angle of incidence - See Section 6.2.1). As a result, the number of possible predictors can be as large as 28. In order to determine the optimal set of predictors to retrieve biomass from ANNs, an evaluation of the estimated AGBs was performed. Taking advantage of the multi-angle capabilities of the SMOS mission, the quality of the produced biomass maps was evaluated in two ways. First, the influence of the TBs' angle of incidence on the quality of the AGB inversions was assessed. Then, the influence of different predictors was analyzed, being TBs' H and V the main ones. Other variables such as the Polarization Ratio (*PR*, Eq. 6.1) and soil temperatures (TS) were used as complementary predictors. Soil temperature was also used to estimate emissivities (E) from the TBs, calculated as E = TB/TS; Henceforth, E in horizontal and vertical polarizations will be referred to as E_h and E_v, respectively. The polarization ratio is defined as:

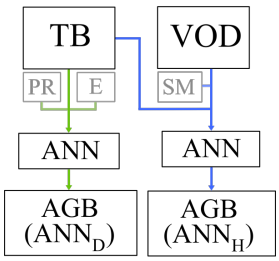
$$PR = \frac{TB_V - TB_H}{TB_V + TB_H} \tag{6.1}$$

Scattering events within the vegetation strongly depolarize the ground emission. Therefore, low PR values are associated with dense vegetation and higher values with sparsely vegetated soils

In addition, a hybrid approach for the estimation of AGB was evaluated. In this approach, the outputs of a physically based radiative transfer model (VOD or SM) were used along with the TBs as complementary predictors in the ANN input vector. To summarize, AGB estimates will be presented from two ANN approaches: the first referred to as 1) the data-driven approach (ANN_D), which is based solely on TBs (and their derived variables), and the second referred to as 2) the hybrid approach, which uses both TBs and VOD (or SM) as inputs for the networks (ANN_H), see Fig 6.1. An additional case using parametric functions based solely on the VOD will be used and discussed later in the document (VOD function).

The learning sets used by the networks are divided in three subsets. The first one, the training subset, is used to train the network and adjust its parameters (weights and biases) in order to reduce the difference between estimates and targets. The validation subset is used to monitor the performances of the ANN at each step of the training process and to detect overtraining if the performances continue to increase on the training subset but start decreasing on the validation subset. The last subset, the test one, is used to evaluate the ability of the ANN to generalize the patterns learned during the training stage to a set of new data. The training, validation and test subsets were built from a random selection of the global observations contained in the learning set, and represent respectively 50 (∼58000 pixels/year), 20 (∼23000 pixels/year) and 30% (∼35000 pixels/year) of the original set.

Figure 6.1: Sketch of pure data-driven approach (AGB(ANN_D)) using TBs or derived variables (like E or PR) and the hybrid approach (AGB(ANN_H)) using TBs and variables from radiative transfert models (VOD or SM).



The optimization of the ANN parameters is done using the Levenberg-Marquardt algorithm, the performance of the network is calculated using the Mean Square Error (MSE). A simple architecture is used: a single hidden layer with 10 neurons with a logistic-sigmoidal activation function, together with an output layer with a linear activation function. Initially, the training is carried out on input-target pairs of the same year. However, thanks to the three available years, a temporal cross-validation is applied. In other words, once a model is trained on data from a specific year, it is then applied to data from another year in order to compare the resulting inversion to the reference map of that period, e.g. an ANN trained with the input-target pair: TBs 2017-CCI2017, is subsequently applied to TBs of 2018, to compare their inversion to the CCI2018 map, and so on for all other possible combinations. An alternative training scheme, based on a multi-year learning-set (a concatenation of inputs and references from the three available years) was explored to assess the flexibility of such an ANN for a temporal cross-validation; the same ANN architecture mentioned before is used, the construction of the learning subsets (training, validation and test) is carried out from a random selection of the three-year concatenated data set.

6.3.2 AGB estimation from parametric function for L-band

Currently, the common approach to estimate AGB from passive microwaves is based on determining a relationship between VOD and AGB using a functional fit. Previous studies have explored this methodology and have proposed different parametric functions to estimate AGB from VOD (*Liu et al.* (2015); *Rodríguez-Fernández et al.* (2018)). Here, the one proposed by *Rodríguez-Fernández et al.* (2018) was used.

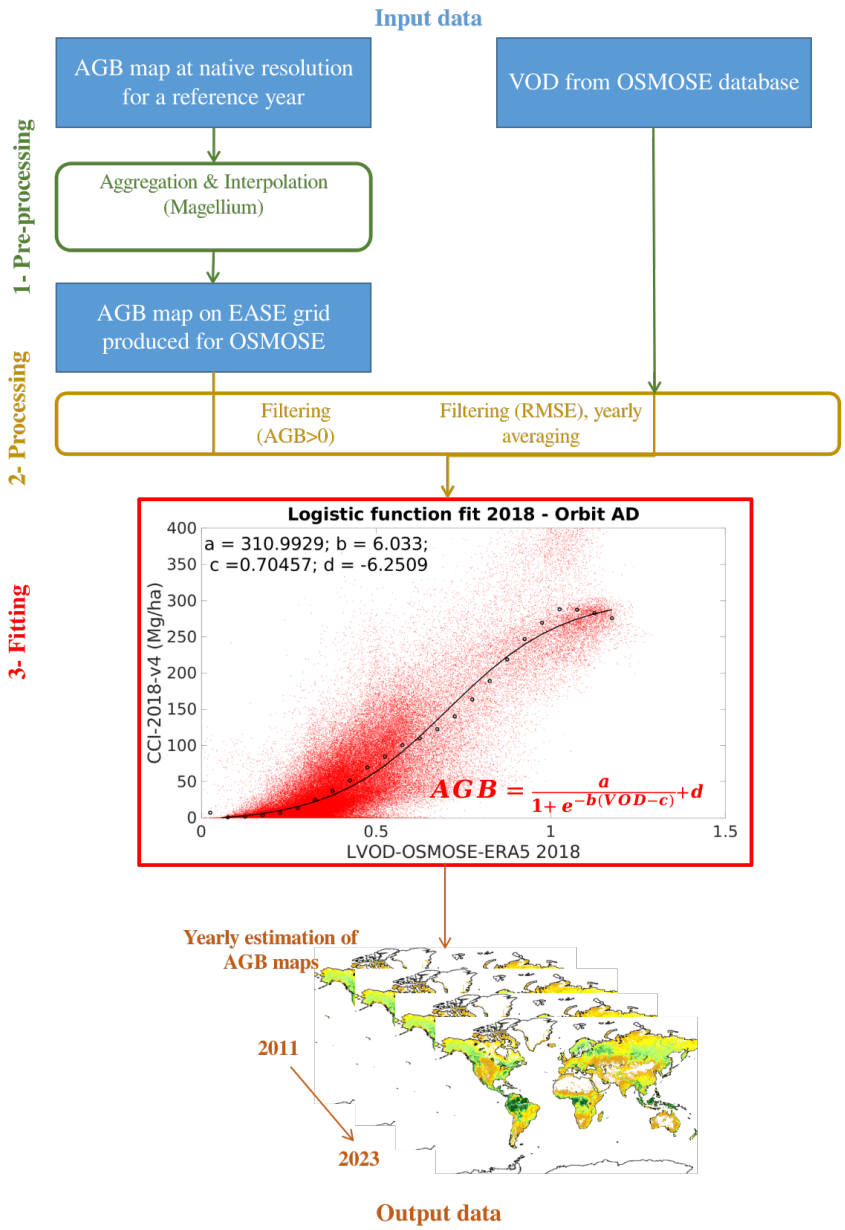


Figure 6.2: Overview of the methodology for estimating yearly AGB maps from SMOS L-VOD and an AGB reference map.

The workflow is divided into three steps numbered from 1 to 3 in Fig. 6.2. The step 1 (pre-processing) refers to the aggregation and interpolation of the AGB calibration map. This was done by Magellium and it is described in Chapter 8.

In step 2 (2- Processing) the SMOS L-VOD daily measurements are masked, filtered, and temporally averaged. Only pixels over continental surfaces are considered, as provided in the 1km USGS (US Geological Survey) land-sea mask aggregated into the EASE 2 grid. Low-quality VOD are removed based on the $RMSE$ between the modeled and measured TB. The SMOS footprints with a RMSE larger than $8K$ are filtered out. The VOD temporal series of each footprint are also checked for potential outliers: the values outside an interval of two sigmas around the yearly average are discarded. The cleaned VOD time series, are then temporally averaged on a yearly basis to iron out the effects on L-VOD of the diurnal and seasonal variations of the vegetation water content. At this stage, footprints with an AGB reference value equal to 0 are discarded, as they are not useful to our AGB estimation.

In step 3 (3- Fitting in Fig. 6.2), the AGB reference map is then compared pixel-wise against the annual SMOS L-VOD map (red points in the 3-Fitting part of Fig. 6.2) for the same year (for example SMOS L-VOD in 2018 against ESA Biomass CCI AGB map for the year 2018). The annual L-VOD is binned into 0.05-width bin. In each bin, the mean AGB from the reference map is computed (black points in the $3-Fitting$ part of Fig. 6.2) and the set of parameters of the logistic function that best fits the mean AGB/L-VOD distribution is estimated. This logistic function is defined in Eq. 6.2:

$$AGB = \frac{a}{1 + e^{-b(VOD-c)}} + d \tag{6.2}$$

where a, b, c and d are the free parameters to obtain the logistic function that best fits the mean AGB distribution in the L-VOD bins (black curve in the $4-Fitting$ part of Fig. 6.2). In equation 6.2, AGB is in $Mgha^{-1}$ and the $L-VOD$ is dimensionless. Hence a and d are in Mg ha^{-1} and b and c have no dimension.

6.3.3 AGB estimation from multi-frequency VOD

Multi-frequency approaches to estimate AGB from passive microwaves VOD was studied using a multi-linear regressions, neural networks and random forest.

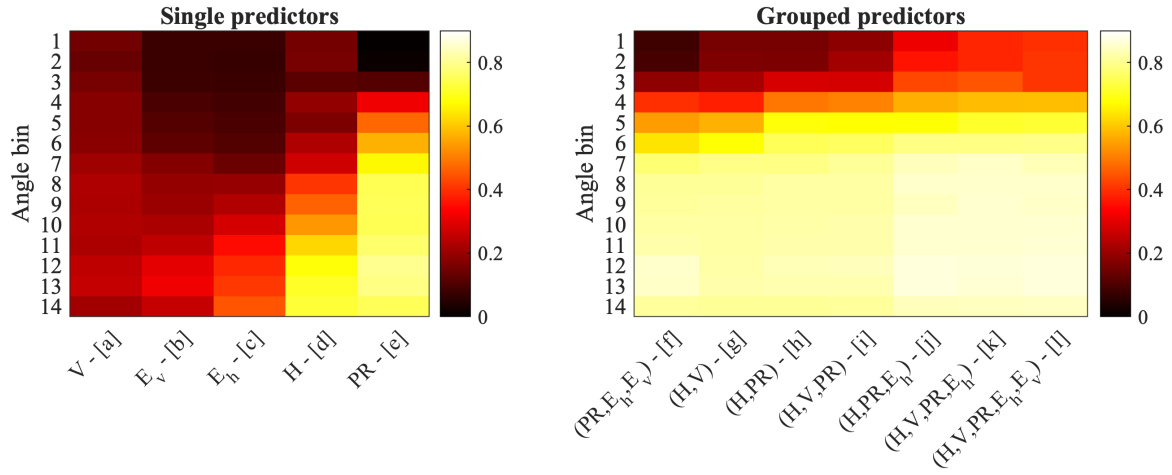
6.4 Results

6.4.1 Estimation of AGB from L-band brightness temperatures

6.4.1.1 Quality evaluation and predictors selection

Variations in the R^2 scores between AGB references and estimates, with respect to the angle of incidence and different predictors, are shown in Fig. 6.3. Several combinations of ANN predictors were tested, using single or grouped variables. Higher R^2 values were found in retrievals from higher incidence angles (bin >8, incidence angles >35º). In single predictors, the H-polarization and PR presented the highest R^2 values (around 0.72 and 0.80, respectively). The V and emissivities alone presented non-significant values, around 0.1 for the most part and 0.45 as the highest value (for E_h between 60 to 65º [Bin No. 14]). The highest R^2 were obtained when multiple predictors were used at the same time. The most performant combinations were the cases *j*, *k* and *l* in Fig. 6.3, all with similar values close to 0.85; the use of E_v and V did not significantly increased the coefficient' scores. The case *f*, which did not use TBs but only variables derived from TB and TS, and case *i*, which used only TBs and PR, showed similar scores for angles on incidence >35º (R^2 ~0.82). In inversions with multiple predictors, the less performant approach was that of case *g*, with slightly lower R^2 scores, at around 0.81.

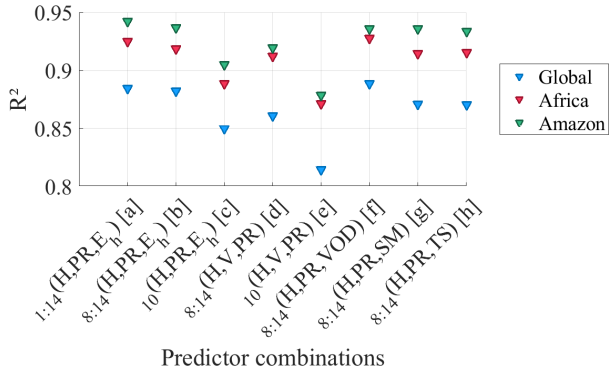
Figure 6.3: R^2 values for AGB estimates depending on angle of incidence (Y axis) and predictors (X axis). Bins from 1 to 14 covering angles of incidence between 0 to 65º, in intervals of 5º -see Section 6.2.1-. For E the subindex represents the polarization: E_h corresponds to the horizontal polarization, and E_v to the vertical one. The inversions presented were produced from 2017 data-sets using the CCI2017 map as a reference.



Thus, from two of the most performant groups of predictors: case *j*, with the highest values of R^2, and case *i* with relatively lower results but using exclusively TBs, the quality of AGB estimates (in terms of R^2) was evaluated for different ranges of angles of incidence on a global and continental scale (see Fig. 6.4). AGB estimates obtained from ANN_D(H, PR, E_h) were evaluated for: all angles (14 angle bins and 3 variables, therefore 42 input predictors), angles between 35 and 65º [bins 8 to 14], and between 40 and 45º [bin number 10] (inputs *a*, *b* and *c* respectively in Fig. 6.4). Inversion from ANN_D(H, V, PR) were evaluated for: angles between 35 and 65º [bins 8 to 14], and between 40 and 45º [bin number 10] (inputs *d* and *e* in Fig. 6.4). Three additional sets of predictors were studied, two for the hybrid approach (using VOD and SM as complementary predictors to H and PR) and one using TS due to its strong influence in TB observations (inputs *f*, *g* and *h* in Fig. 6.4).

Estimates from a single angle of incidence had the lowest R^2 values among all combinations. However, using one angle with the predictors (H,PR,E_h) presented similar scores to those produced from angles 35 and 65º [bin 8 to 14] with (H,V,PR), around 0.86 in both cases. On the global scale, the most performant values come from the combination *a*, *b* and *f* in Fig. 6.4, with an R^2 of ∼0.89. Finally, the hybrid case using SM (and TS as complementary predictor) showed similar coefficients (close to 0.87).

Figure 6.4: Influence of different angles of incidence on the R^2 of the best predictors. The label '8:14' represents the TBs between 35 and 65º (see Section 6.2.1). The '1:14' tag is equivalent to 0 to 65º. The inversions presented were produced using data-sets from 2017 and the CCI2017 map as a reference.



A similar analysis to that of Fig. 6.3 and Fig. 6.4 was done for 2010 and 2018 (at a global scale and for Africa and South America), similar results were obtained. Consequently, henceforth the following groups of

predictors will be used: *i*) the optimal combination of predictors (H,PR,E_h), as it gives the best compromise between angles of incidence and higher R^2 scores of the estimates, *ii*) the combination (H,V,PR), that despite having relatively low values for the coefficient of determination, exhibits the results that can be expected from multi-angular TBs alone, and *iii*) both hybrid cases, as they presented considerable R^2 values without the use of emissivities; Despite having similar results to the latter, the combination that includes TS as complementary predictor will not be used, since it is considered that the effect of the surface temperature on the AGB estimates was better captured from the emissivities (derived from TB and TS).

Model inversions from the selected cases were evaluated in both their ability to explain the variability of the reference AGB maps (R^2) and their dispersion with respect to them (RMSE). The first three matrices (from top to bottom) of Fig. 6.5, present the R^2 values obtained for the AGB inversions produced for the three reference years. In Fig. 6.5 the title of each matrix represents the year of the data-set that was used for the training of models, therefore, the second matrix titled '2017' contains the R^2 of all the models trained using input variables from 2017 as predictors and the CCI2017 as reference. The columns, on the other hand, represent the year of the data-set [reference map] to which the trained model was applied to [compared to], e.g. the value of the upper right corner of the '2017 matrix' (so 0.87), represents the R^2 between: 1) the CCI2018 reference map and 2) the AGB estimate from the ANN_D(H,PR,E_h) trained with 2017 data and applied to data-sets from 2018.

On average, when the AGB estimates produced by ANN_D(H,V,PR) were evaluated with respect to their respective reference map, a R^2 of around 0.85 units was obtained; this approach presented slightly lower R^2 values than the other cases. When looking at the remaining ANNs and the temporal extrapolation of models, two patterns were observed. First, both ANN_H(H,PR,VOD) and ANN_D(H,PR,E_h) presented similar performances (R^2 $\sim$0.87) for each year. However, the use of E_h or VOD as a complementary predictor yielded relatively (but not significantly) higher values than the use of SM; the inclusion of horizontal polarized emissivities as an additional predictor of ANNs slightly increased the R^2 value (in $\sim$0.03 units) with respect to the ANN_D(H,V,PR). Second, the temporal cross-validation of models showed a decrease in R^2 values. Models trained in 2010 and applied to 2018 reduced their R^2 by about 0.2 units, when applied to 2017 data-sets the difference was not significant. A similar decrease was observed when applying models trained in 2018 to data from 2010 ($\sim$0.2). The difference in R^2 between 2017 and 2018 models was not significant.

Similar patterns were observed in the RMSE values (not shown here), i.e. a) an improvement from the inclusion of E_h (and similar results from the VOD hybrid case) and b) a deterioration when applying models to data-sets from years different than those used during training. AGB estimates from ANNs trained using (H,V,PR) from 2018, 2017 and 2010 showed RMSE in the order of 38.7 [54%], 38 [54%] and 38.6 Mg/ha [55%] respectively. The ANN_D(H,PR,E_h) case reduced the previous values by 3.7 [4%], 3.8 [4%] and 2.4 Mg/ha [5%] respectively (similar decreases were observed with the ANN_H(H,PR,VOD)).

Finally, ANNs were trained with learning-sets composed of the three reference years (multi-year learning-set mentioned at the end of Section 6.3.1); the temporal cross-validation of models trained under these conditions is shown in the '3 years' matrix in Fig. 6.5 (bottom matrix). The evaluation is done using only test subset (Sect. 6.3.1). The ANNs trained under this configuration, as in the previously mentioned cases, showed slightly higher results from both the ANN_D(H,PR,E_h) and the ANN_H(H,PR,VOD) approach (increase of $\sim$0.03 units) than from the ANN_D(H,V,PR). Additionally, a significant improvement in the cross-validation of models trained with multiple years was observed. The aforementioned interannual differences in R^2 (0.2 for models trained in 2010 [2018] and applied to 2018 [2010] data-sets), were not significant under the 3-years training configuration ($<$0.01 units). When comparing the R^2 (and RMSE) from multi-year trained models with the results from single-year models (trained and applied on data from the same year), a slight (but not significant) decrease in R^2 was observed in some cases. However, to carry out the evaluation of AGB estimates, the models trained with a multi-year learning set (concatenation of three years) were chosen, since they are the ones that showed greater stability for inter-annual analyses.

6.4.1.2 Evaluation of AGB estimates from ANNs

Once the quality metrics for the different cases were evaluated, and the optimal training configuration defined, the spatial distribution of residuals was explored (that is, the differences between references and estimates). This analysis was performed exclusively on three cases of interest, more precisely: ANN_D(H,PR,E_h),

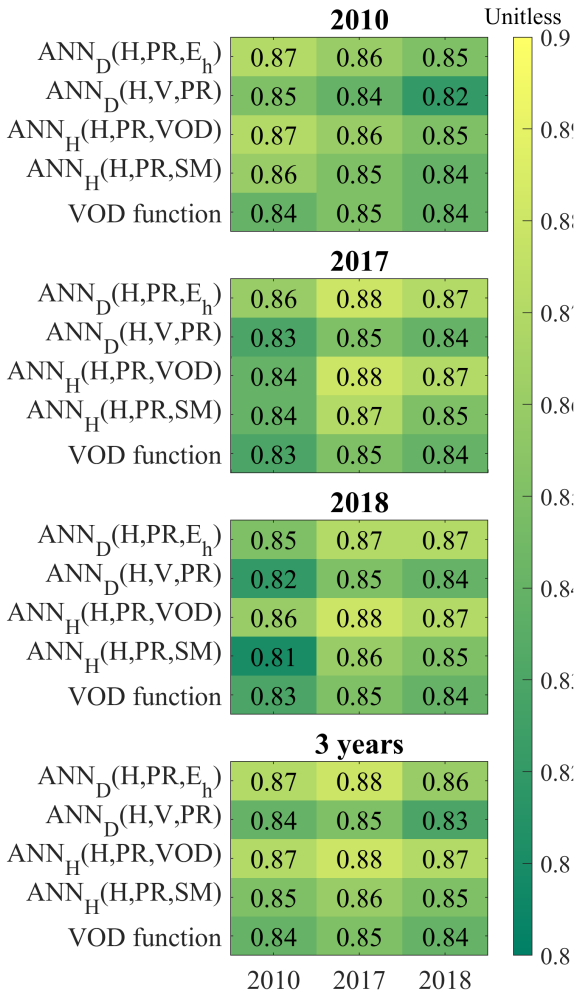
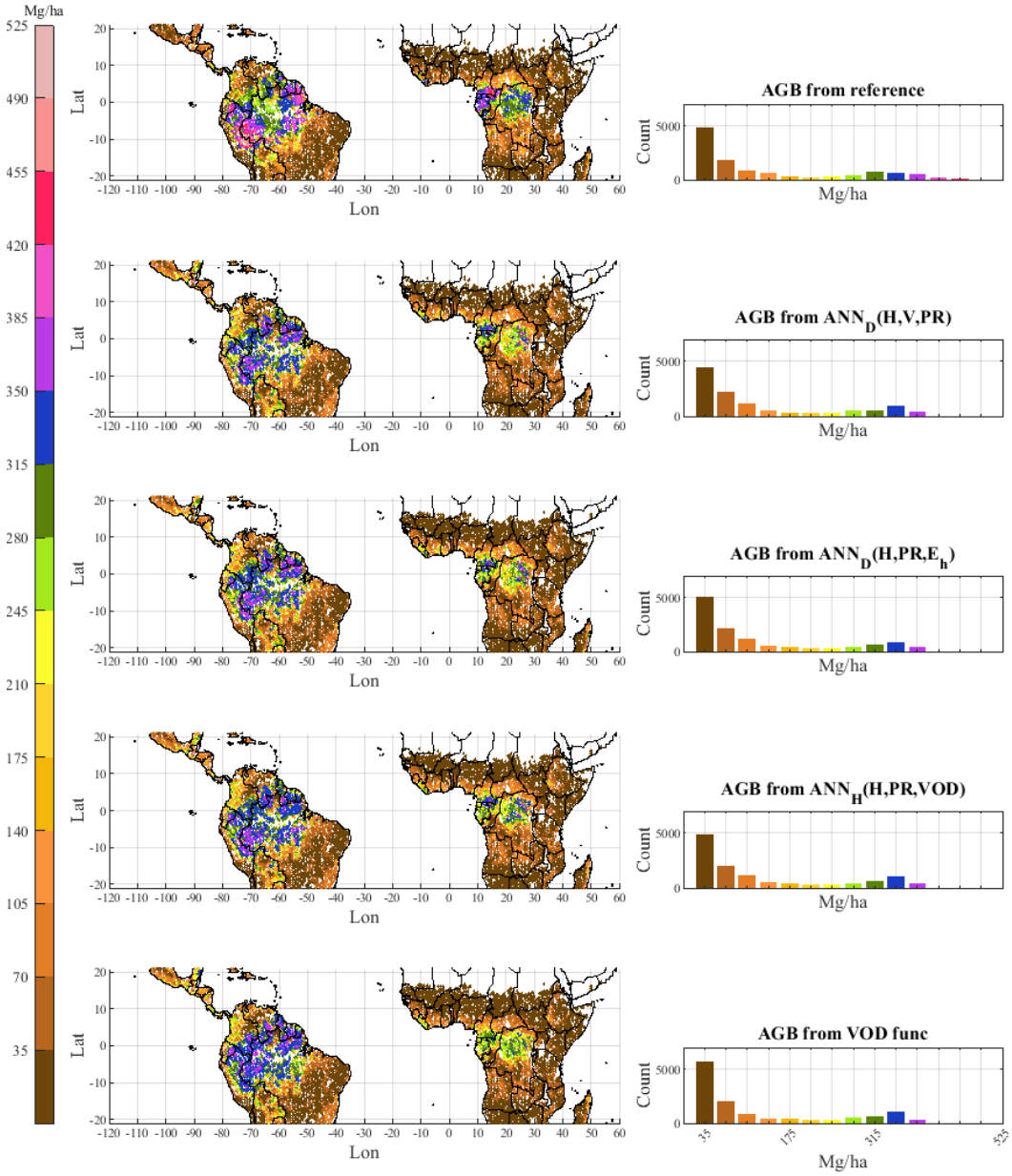


Figure 6.5: Temporal cross-validation of models (in terms of R^2). The title of each matrix represents the year of the data-set used for training the models. On the Y axis of each matrix: The models used to estimate AGB. On the X axis : The years of the datasets to which the trained models were applied to.

ANN_D(H,V,PR) and ANN_H(H,PR,VOD); the results by the ANN_H(H,PR,SM) presented lower quality metrics to those produced by using VOD as an additional predictor, only the results of the latter are shown.

Figure 6.6: Comparison of the spatial structure of AGB maps. Inversions are produced by models trained with a multi-year data-set and applied to 2018 observation, the reference is the CCI2018 map. AGB is regrouped in bins of 35 Mg/ha of width.



The evaluation of the spatial patterns of AGB estimates produced from the selected ANN_D and ANN_H over tropical Africa and America are shown in Fig. 6.6. In order to represent the spatial distribution of vegetation biomass, the AGB was regrouped into bins of 35 Mg/ha of width. In the left column, and in descending order, are the AGB maps from: CCI2018, the data-driven cases ANN_D(H,V,PR) and ANN_D(H,PR,E_h), and the hybrid case ANN_H(H,PR,VOD) (the final panel will be discussed in Section 6.4.1.3). In the right column the count per AGB bin. All inversions are from models trained with the multi-year data-set and applied to 2018 observations.

Just like shown by the other quality metrics (bottom box in Fig. 6.5), both the ANN_D(H,PR, E_h) and the ANN_H(H,PR,VOD) presented similar results. However, none of the three cases was capable of reproducing

the highest AGB values, e.g. the redish spots in the Amazon rain-forest (around [Lat: -10, Lon: -70], or [Lat: 0, Lon: -55]), where models reproduced values around 385 Mg/ha whilst the reference show AGB values around ∼500 Mg/ha. Similarly, even if the main structural features of reference maps were captured by ANNs, e.g. lower values in northern and southern tropical Africa, and higher values in the more equatorial region of both America (Amazon rainforest) and Africa (Congo forest), subtle spatial variations like the transition from densely to sparsely vegetated areas in the Amazon rainforest (region between Lat: [-10, 0] , Lon: [-70, 60]) were not reproduced accurately. Nevertheless, the cluster of high AGB values from the reference map, located near the Atlantic Ocean on the tropical America, were more closely reproduced by ANN_D(H,PR,E_h) and ANN_D(H,V,PR), than by the hybrid approach. Regardless, ANNs tended to correctly replicate the main features of the distribution of the reference map (histograms on the right side of Fig. 6.6).

6.4.1.3 Comparison of estimates from ANN and VOD fitted functions

In view of the demonstrated potential of the methodology proposed here to estimate AGB, a comparison between AGB estimates produced from ANNs with those produced from the fitted VOD function was finally explored. The same evaluation criteria applied to the ANNs were applied to the VOD-based estimates. To ensure a fair comparison, the training of models and the comparison of estimates were performed on the same data-sets for both methods.

The fifth row of each matrix in Fig. 6.5 (R^2 results), gives a comparison of the performance of the VOD function with respect to ANNs. The R^2 scores between references and AGB estimates from the VOD functions presented similar results independently of year used for the training. R^2 values from ANN_D(H,PR, E_h) are generally ∼0.03 units higher than those from the VOD parametric function; a similar difference was observed when compared to ANN_H(H,PR,VOD). The RMSE values associated with the inversion of the VOD function were around 2.8 Mg/ha larger than those on the ANN_H(H,PR,VOD), the difference with respect to the ANN_D(H,PR,E_h) case was on the order of ∼2 Mg/ha.

The lower panels of Fig. 6.6 show the spatial patterns and histogram of AGB estimates produced from VOD. Estimates of the parametric function tended to simplify the spatial variations of biomass. In regions with higher AGB densities, the estimates of the parametric function deviated from the references more strongly than those from ANNs, resulting in inversions larger than the reference AGB in the Amazon rain forest while lower than the AGB reference in tropical Africa. Furthermore, while it is true that both methods (the ANNs and the parametric functions) underestimated the highest AGB values (Fig. 6.6), it is also true that the ANNs were closer to the latter than the VOD function. In general, the AGB reference distribution was reproduced more accurately by the ANNs (right-side Figures in Fig. 6.6). For example, the Amazon rain-forest was mostly represented by values between 315 and 385 Mg/ha under the VOD-based models, while according to the estimates of the reference, it can encompass densities between 245 and >500 Mg/ha.

One might think that a more complex approach such as using an ANN with VOD as only input instead of a parametric function could give better performances, however an ANN with VOD as input did not change significantly the results obtained with the parametric function.

6.4.2 L-AGB OSMOSE

Following the methodology described in Sect. 6.3.2, a yearly AGB time series has been estimated from the project L-VOD dataset. An example of the L-AGB in 2018 is displayed in Fig. 6.7. This L-AGB has been generated using the year 2018 as a reference. This means that the CCI calibration map reprojected to EASE 2 for the year 2018 and the 2018 L-VOD average, mixing ascending and descending orbits, have been used to calibrate the parameters of the logistic function (Eq. 6.2). The maximum AGB estimates are in the equatorial belt and reach 300 Mg ha^{-1}. The boreal forests, in the northern high latitudes, present AGB estimates around 100-150 Mg ha^{-1}. Temperate and arid regions show an estimated AGB lower than 50 Mg ha^{-1}. The spatial distribution is coherent with CCI 2018 (see Fig. 6.8) and with the input L-VOD.

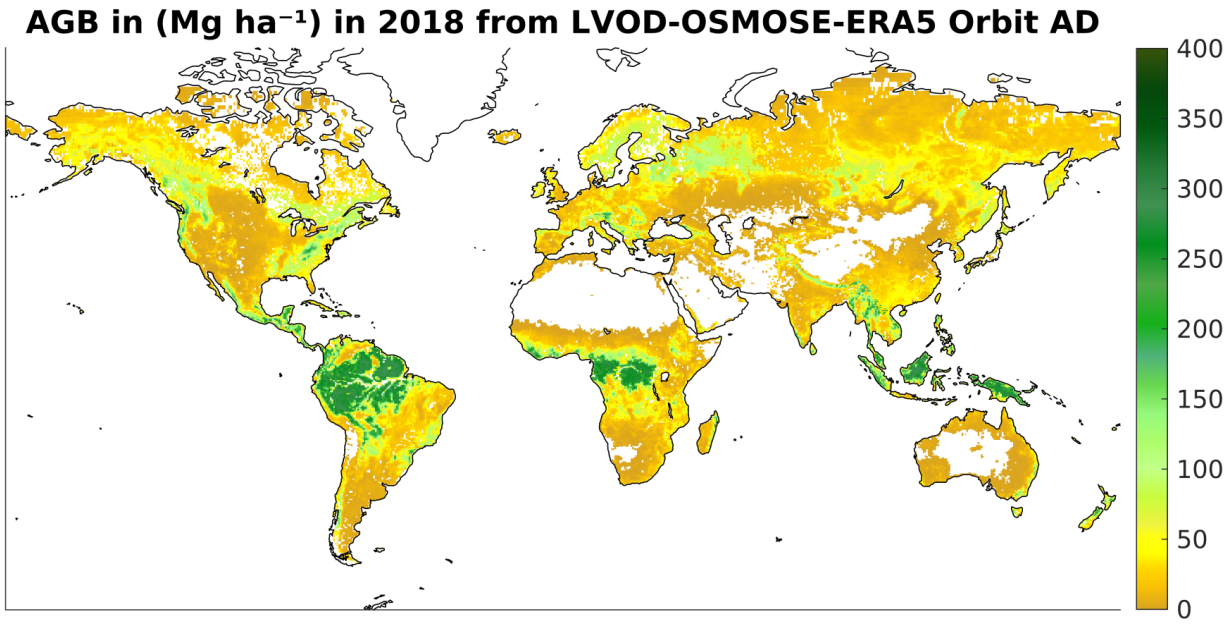


Figure 6.7: AGB estimated from the OSMOSE L-Band VOD, mixing ascending and descending orbits

Panels (a) to (d) in Fig. 6.8 show the spatial distribution of the differences across the four years (2017-2020) of the CCI reference maps. The maps are very similar from one year to the other. The region with the highest differences (more than 75 pixels with differences greater than 200 Mg ha^{-1}) is the equatorial part of Africa. The statistics (see Table 6.2) are also steady over the years. The AGB estimated from OSMOSE L-VOD is very well correlated to the CCI maps from 2017 to 2020 (R above 0.88). On the global scale, the OSMOSE derived L-AGB shows similar distribution, presents more variability, and slightly lower estimations compared to the CCI. Table 6.2 shows a global bias of ≈ -2.5 Mg ha^{-1} and Fig. 6.7 highlights that this negative bias is particularly high over the Congo basin and the boreal forests of eastern Europe and Russia.

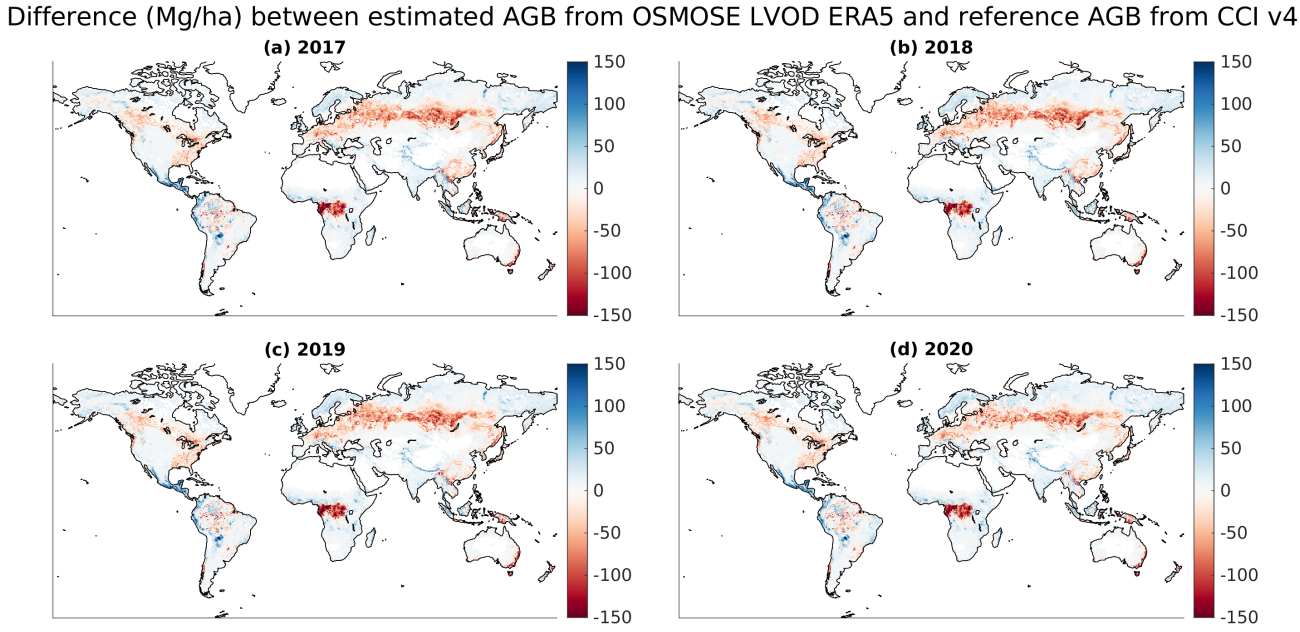


Figure 6.8: (a)-(d) Difference in Mg ha^{-1} between the yearly AGB estimated from the 2018 global logistic fit minus CCI AGB values for the years 2017-2020.

Table 6.2: R, bias (Mg ha^{-1}) and RMSE (Mg ha^{-1}) between the estimated AGB from OSMOSE LVOD and AGB values from CCI for the years 2017-2020. AGB was estimated using the global logistic fit computed from the VOD and the CCI AGB for the year 2018. *: year of reference.

	2017	**2018***	**2019**	**2020**
R	0.88	0.89	0.89	0.88
bias	-2.46	-2.42	-2.64	-2.26
RMSE	36.30	35.89	35.41	35.93

Overall, on a yearly basis, the AGB generated from the OSMOSE L-VOD compares well with the optical indices from the data cube 2.

6.4.3 AGB evaluation

6.4.3.1 AGB evaluation method

The major problem inherent in comparing SMOS AGB maps and in situ data lies in the drastic difference in spatial resolution between these dataset. In fact, we are comparing SMOS AGB maps at 25 m pixel size with in situ data of pixel size ranging from 100 m (lidar-based AGB maps) to 1km (data from *Ploton et al.* (2020a)). The surface area of in situ data located in a single SMOS pixel therefore represents a very small fraction of the SMOS pixel surface area itself. In this particular case, in situ data are not necessarily representative of the SMOS pixel. For example, a SMOS pixel embedded with 50% of dense forest (mean AGB of 300 Mg/ha) and open forest (mean AGB of 100 Mg/ha) has a mean biomass of 200 Mg/ha, to be compared with in situ data located either in dense forest or in open forest. To account for this bias, we need to account for the representativity of the in situ plot on the ground.

To do so, we corrected in situ data AGB values using the 2010 tree cover density map at 30m pixel size from *Hansen et al.* (2013), as detailed in Figure 6.9. We first updated the 2010 tree cover map to 2020 by setting to 0 all the pixels considered as forest loss from 2010 to 2020 in the forest loss maps from *Hansen et al.* (2013). From the 30 m pixel size 2020 tree cover map, we then derived 100 m and 1 km tree cover maps (average interpolation) to fit with reference in situ data, and 25 km pixel size tree cover map by averaging all the 30m pixels located in the 25 km SMOS AGB pixels. We finally computed the adjusted in situ AGB pixels as follows:

$$AGB_{ref,adj_i} = AGB_{ref_i}.\frac{E(tc_{smos})}{tc_{ref_i}} \tag{6.3}$$

with AGB_{ref_i} the AGB reference (in situ) value at pixel i, $E(tc_{smos})$ the 25 km tree cover pixel and tc_{ref_i} the 100m or 1km tree cover value at pixel i.

We finally performed the validation process by comparing AGB pixel values at 25km with adjusted reference in situ data values AGB_{ref,adj_i}.

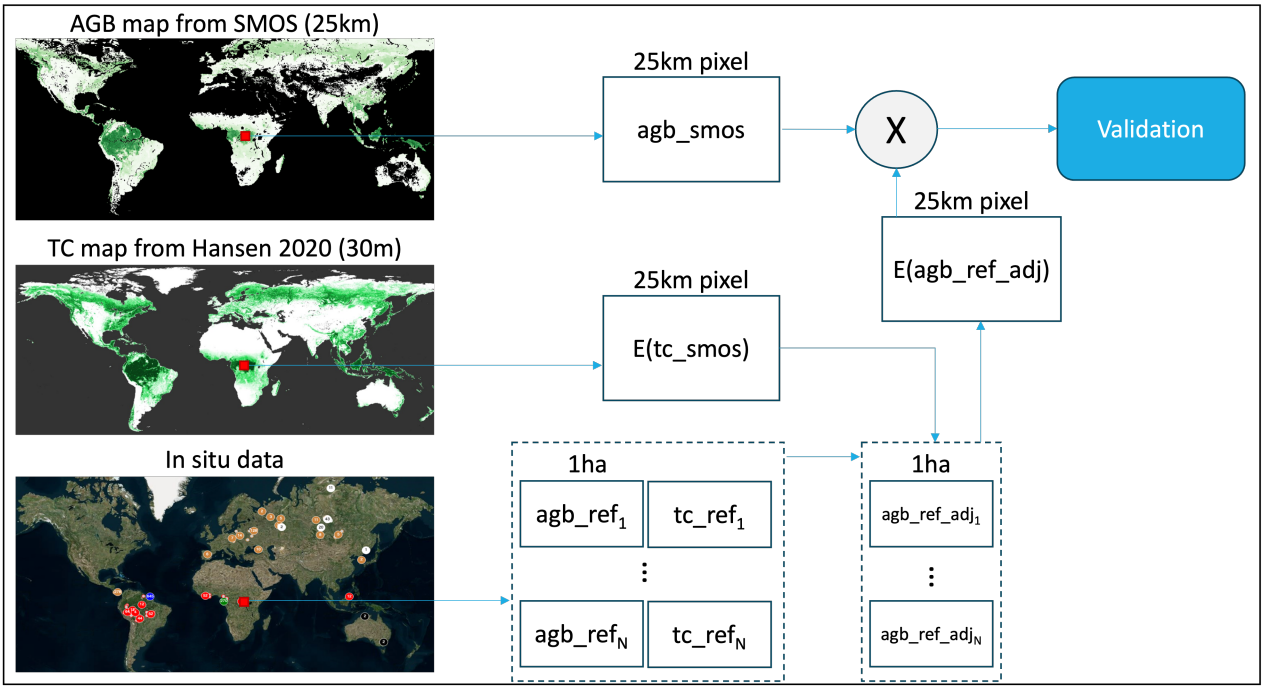


Figure 6.9: Overview of the methodology for the evaluation of yearly AGB maps from SMOS L-VOD and AGB in situ data described in section 6.2.4.

6.4.3.2 AGB evaluation results

Lidar-based AGB data

The validation results associated with all lidar-based AGB maps described in table 6.1, compared with 2020 SMOS AGB from VOD calibrated with CCI AGB map are presented in figure 6.10. Results associated to CMS (Indonesia), NEON (USA), SLB (Brazil) and TERN (Australia) data are shown in green, red, blue and yellow respectively. The accuracy assessment resulted in an R value of 0.59, a bias of -58.5 Mg/ha and a RMSE of 104.8 Mg/ha. Considering AGB values from lidar maps greater than 100 Mg/ha, 150 Mg/ha and 200 Mg/ha lowered the validation results. The R value decreased from 0.56 to nearly 0, the bias from -58.5 Mg/ha to -122 Mg/ha and the RMSE increased from 104.8 Mg/ha to 158.1 Mg/ha when considering AGB values from lidar maps greater than 200 Mg/ha.

The TERN network demonstrated the best performance in term of correlation, achieving an R value of 0.86. However, the bias and RMSE were found to be large, that is to say -145.5 Mg/ha and 205.5 Mg/ha respectively, showing a consequent under-estimation of AGB estimated from SMOS. The CMS network exhibited a R value of 0.74, a small bias of -1.9 Mg/ha and a RMSE of 67.2 Mg/ha, the NEON network a R value of 0.58, a bias of -68.4 Mg/ha and a RMSE of 93.8 Mg/ha, and the SLB network a R value of 0.51, a bias of -65.8 Mg/ha and a RMSE of 111.6 Mg/ha. As an example, the detailed results related to the CMS network, for all years, is shown in figure 6.11.

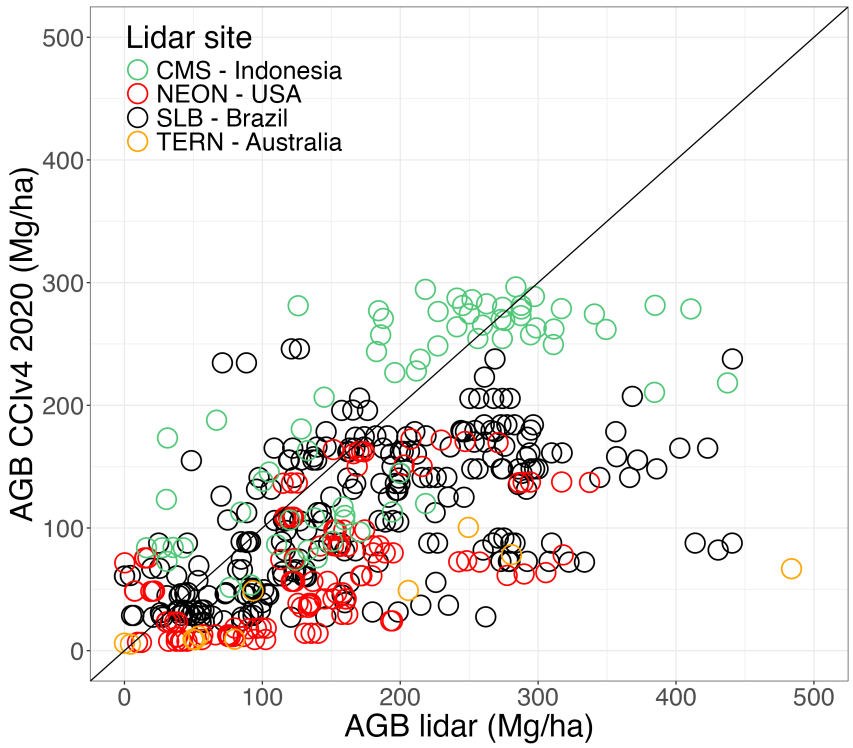
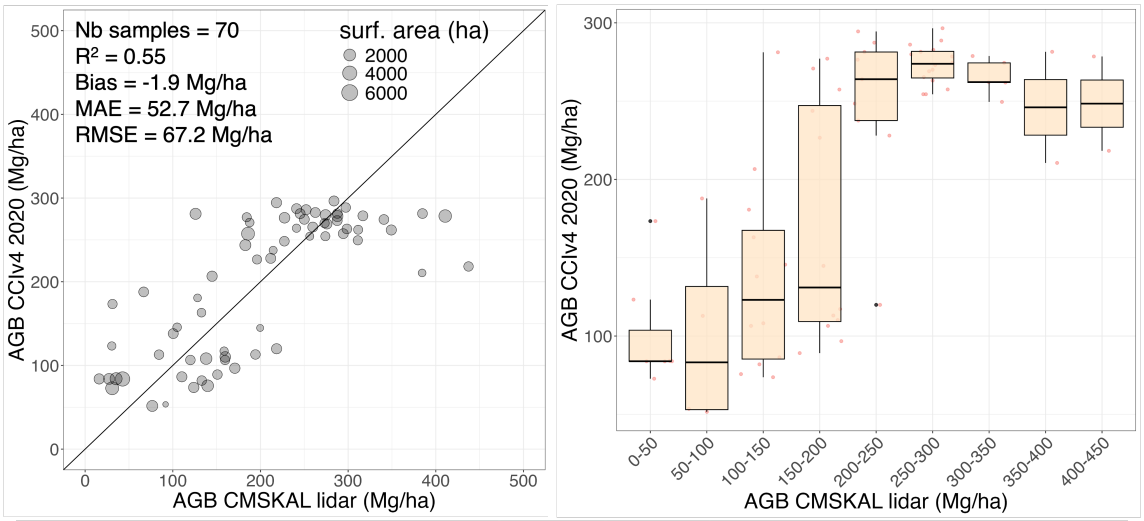


Figure 6.10: Validation using all lidar-based AGB vs 2020 SMOS AGB calibrated with CCI maps.

Table 6.3: R, bias, MAE and RMSE between the estimated AGB from OSMOSE LVOD calibrated with CCI maps, and lidar-based AGB maps >0 Mg/ha, >100 Mg/ha, >150 Mg/ha and >200 Mg/ha.

	All	**AGB>100 Mg/ha**	**AGB>150 Mg/ha**	**AGB>200 Mg/ha**
nb. sample	441	317	232	147
R	0.56	0.35	0.22	-0.04
bias	-58.5	-79	-97.9	-122
MAE	76.4	91.2	105.4	127.4
RMSE	104.8	119.8	134.8	158.1

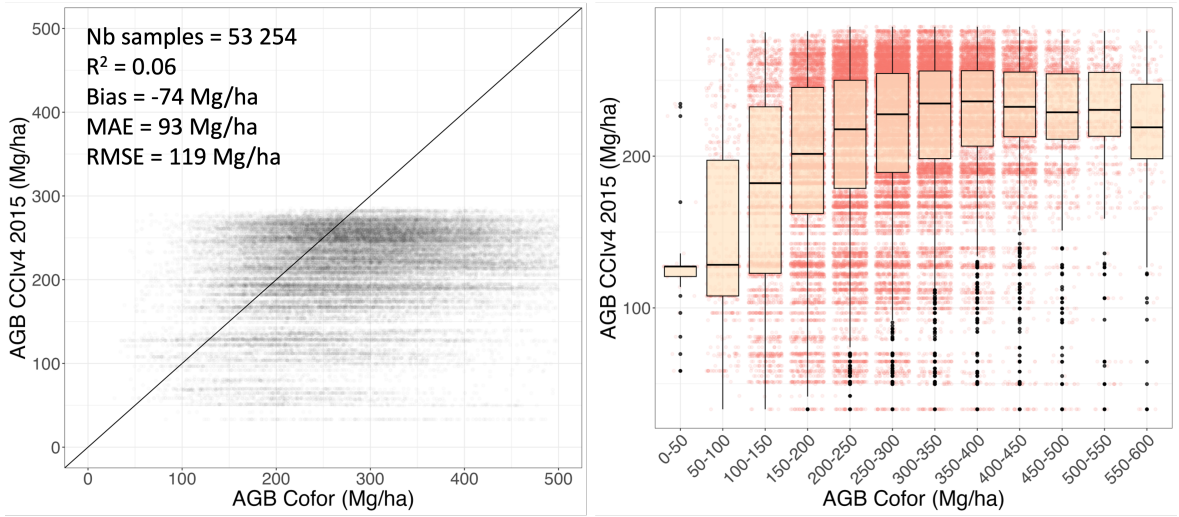


CMS - Indonesia	R	R²	bias (%)	MAE (%)	RMSE (%)	bias (Mg/ha)	MAE (Mg/ha)	RMSE (Mg/ha)
CCIv4_2011	0.70	0.49	38.10	59.33	125.21	-0.86	55.51	71.94
CCIv4_2012	0.72	0.52	43.03	62.75	133.55	6.24	56.35	70.11
CCIv4_2013	0.75	0.56	42.60	59.65	126.61	10.54	53.24	67.54
CCIv4_2014	0.74	0.55	48.72	64.23	142.72	12.79	54.06	68.83
CCIv4_2015	0.72	0.51	40.06	59.92	123.89	8.18	56.03	70.32
CCIv4_2016	0.75	0.57	26.90	49.28	92.19	5.11	54.10	67.37
CCIv4_2017	0.73	0.54	34.27	55.09	111.36	5.59	54.38	68.89
CCIv4_2018	0.72	0.51	42.73	61.26	131.14	9.39	55.23	70.49
CCIv4_2019	0.72	0.51	44.12	61.75	135.64	8.66	54.23	70.44
CCIv4_2020	0.74	0.55	26.00	50.31	98.18	-1.91	52.69	67.23
CCIv4_2021	0.75	0.56	35.15	53.97	111.61	5.55	51.90	66.51
CCIv4_2022	0.74	0.54	33.76	53.55	109.29	4.49	52.60	67.92

Figure 6.11: Validation using lidar-based AGB from the CMS network in Indonesia vs SMOS AGB calibrated with CCI maps.

In situ AGB data from Ploton et al. (2020a)

Validation results using in situ AGB data and SMOS AGB calibrated with CCI are presented in Figure 6.12. Compared to lidar-based AGB data, results showed very low correlation with r^2 from 0.03 to 0.08, a strong under-estimation of AGB estimates (bias ranging from -66.8 to -90.1 Mg/ha) and a RMSE from 112.8 up to 133.6 Mg/ha. Note that validation results when SMOS VOD is calibrated with Avitabile's map does not show better correlation ($r^2 < 0.8$), although the bias is found a bit smaller (from -25.6 to -48.5 Mg/ha). Finally, SMOS maps directly based on Tbs showed slightly better results ($r^2 = 0.9$ and bias from -22.6 to -1.6 Mg/ha). We don't think that SMOS AGB estimates are poorer over Africa. Instead, AGB estimates from commercial inventories in *Ploton et al.* (2020a), at 1km pixel size but originally with a 0.5ha resolution, are probably not representative enough. This hypothesis highlights the need of lidar-based AGB data calibrated with in situ tree inventories for an accurate statistical assessment of SMOS AGB maps.

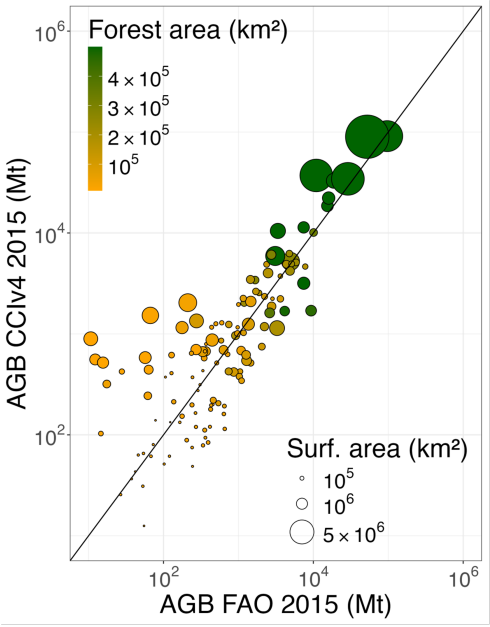


	r	r²	bias (%)	MAE (%)	RMSE (%)	bias (Mg/ha)	MAE (Mg/ha)	RMSE (Mg/ha)
CCIv4_2011	0.26	0.07	-25.31	34.77	41.43	-90.11	104.15	130.54
CCIv4_2012	0.27	0.07	-23.93	34.16	40.93	-86.73	101.90	127.62
CCIv4_2013	0.28	0.08	-21.51	32.27	38.87	-81.00	96.89	122.21
CCIv4_2014	0.29	0.08	-17.87	30.78	38.02	-71.53	90.97	115.59
CCIv4_2015	0.24	0.06	-18.48	31.52	38.96	-73.58	93.36	119.28
CCIv4_2016	0.27	0.08	-16.71	30.78	38.50	-68.66	90.06	114.91
CCIv4_2017	0.18	0.03	-17.50	32.79	41.00	-72.42	95.96	122.10
CCIv4_2018	0.17	0.03	-24.04	35.78	43.27	-88.71	106.21	133.60
CCIv4_2019	0.26	0.07	-21.75	33.49	40.57	-82.12	99.22	123.97
CCIv4_2020	0.25	0.06	-20.21	32.38	39.67	-78.33	96.08	121.33
CCIv4_2021	0.26	0.07	-18.86	31.75	39.16	-74.22	93.60	118.83
CCIv4_2022	0.28	0.08	-15.92	30.28	37.85	-66.80	88.58	112.84

Figure 6.12: Validation using in situ AGB data from *Ploton et al.* (2020a) vs SMOS AGB calibrated with CCI maps.

FAO AGB data

Comparison of SMOS AGB calibrated with CCI data with FAO national statistics is presented in Figure 6.13. Correlation coefficient r^2 is comprised between 0.85 to 0.90. Estimates are slightly over-estimated for some countries, in particular where biomass stocks are very low in North Africa (Lybia, Algeria, Tunisia, Egypt, Niger, Mali) and Central Asia (Kazakstan, Turkmenistan, Afghanistan). Our estimates suggest a somewhat higher overall carbon stock than what the FAO reports, which can be attributed to the fact that we provided AGB estimates for entire countries. In contrast, the FAO focuses solely on carbon stocks within forests, defined as areas with at least 10% tree cover and spanning no less than 0.5 ha. This discrepancy highlights the potential carbon stock overlooked when limiting the scope to forests alone. Additionally, it is worth mentioning that FAO's Figures are derived from national reporting, and the capability to precisely gauge carbon stocks varies significantly across countries. In general, countries with limited capabilities for AGB estimation have no means of estimating their carbon stocks, and declare inaccurate values. Countries showing the highest values of AGB estimates from SMOS VOD and FAO are Brazil (almost the same estimates), Russia, Democratic Republic of the Congo, Indonesia, Colombia and China. SMOS AGB estimates calibrated using Avitabile's map are more correlated to FAO data (r^2 is comprised between 0.83 to 0.94) but are a bit more biased. Finally, SMOS AGB estimates using Tbs are the most correlated to FAO data ($r^2 = 0.95$) but also the most biased.

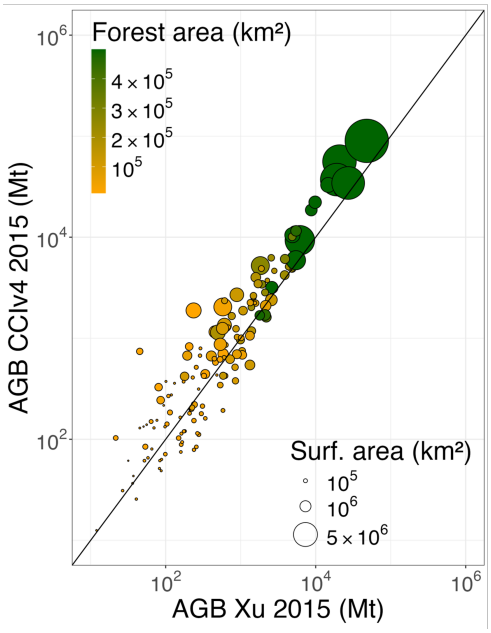


	r	r²	bias (Mt)	MAE (Mt)	RMSE (Mt)
CCIv4_2011	0.94	0.89	559.88	1334.43	4353.96
CCIv4_2012	0.94	0.89	650.70	1431.98	4424.59
CCIv4_2013	0.93	0.87	745.52	1504.16	4881.01
CCIv4_2014	0.95	0.90	669.18	1378.06	4086.75
CCIv4_2015	0.94	0.88	730.35	1448.12	4459.39
CCIv4_2016	0.94	0.88	723.21	1461.69	4589.89
CCIv4_2017	0.93	0.87	704.90	1473.26	4510.15
CCIv4_2018	0.93	0.86	687.38	1506.03	4550.65
CCIv4_2019	0.92	0.85	720.61	1550.24	4822.89
CCIv4_2020	0.93	0.86	716.51	1555.26	4739.81
CCIv4_2021	0.94	0.88	614.92	1423.59	4185.20
CCIv4_2022	0.93	0.86	700.17	1479.13	4557.66

Figure 6.13: Validation using national FAO data vs SMOS AGB calibrated with CCI maps.

AGB maps from Xu et al. (2021)

We compared SMOS AGB estimates from VOD calibrated with CCI maps with AGB maps from *Xu et al.* (2021) at the national scale, and showed comparison results in Figure 6.14. We found that both maps were very correlated with r^2 from 0.86 to 0.96 depending on the year. Estimates from SMOS are a bit over-estimated compared with the maps from *Xu et al.* (2021). We found similar results when SMOS AGB were calibrated using Avitabile's map (r^2 from 0.91 to 0.94 with a higher positive bias) or estimated using Tbs (r^2 from 0.95 to 0.96 with a higher positive bias as well). When comparing directly the map from *Xu et al.* (2021) with FAO, comparison results were the best with r^2 from 0.91 to 0.92 and a negative bias. According to these results, it is difficult to assess which one of AGB maps from SMOS or from *Xu et al.* (2021) is the best one, but both provide excellent comparison results with FAO data.



	r	r²	bias (Mt)	MAE (Mt)	RMSE (Mt)
CCIv4_2011	0.98	0.96	1514.23	1634.55	5923.51
CCIv4_2012	0.98	0.96	1620.26	1756.22	6066.86
CCIv4_2013	0.98	0.96	1699.93	1785.02	6348.13
CCIv4_2014	0.98	0.96	1598.92	1672.08	5824.85
CCIv4_2015	0.98	0.96	1661.39	1754.50	6040.28
CCIv4_2016	0.98	0.96	1652.34	1746.80	6054.94
CCIv4_2017	0.98	0.96	1622.19	1721.88	5868.07
CCIv4_2018	0.97	0.95	1594.67	1691.50	5833.51
CCIv4_2019	0.97	0.95	1621.13	1704.36	5879.05
CCIv4_2020	0.98	0.95	1638.98	1711.43	5856.69
CCIv4_2021	0.98	0.95	1558.23	1635.56	5495.15
CCIv4_2022	0.98	0.95	1634.64	1705.28	5794.96

Figure 6.14: Comparison of SMOS AGB calibrated with CCI maps with AGB maps from *Xu et al.* (2021) at the national scale.

6.4.4 Estimation of AGB from multi-frequency VODs

The three methods mentioned in Sect. 9.3 (multi-linear regression, a feed-forward neural network and a random forest algorithm) were applied to multi-frequency VOD data with and without NDVI.

The Pearson correlation between estimated AGB and the reference increases from multi-linear regression (0.87) to neural networks (0.9) and random forest (0.92). A similar trend is found for Spearman ρ. Numerically, very similar results in terms of correlation were obtained when also adding NDVI for the multi-linear regression and the neural networks.

AGB estimations for different years were done using the multi-frequency approach establishing the relationship in 2018 using the CCI2018 AGB map. Maps were also computed with the different methodologies and the different predictors sets with and without NDVI. A summary of the results is shown in Fig. 6.15-6.16. The best results are found for the random forest approach including NDVI as predictor. Table 6.4 shows different quality metrics for other years than 2018, confirming these results.

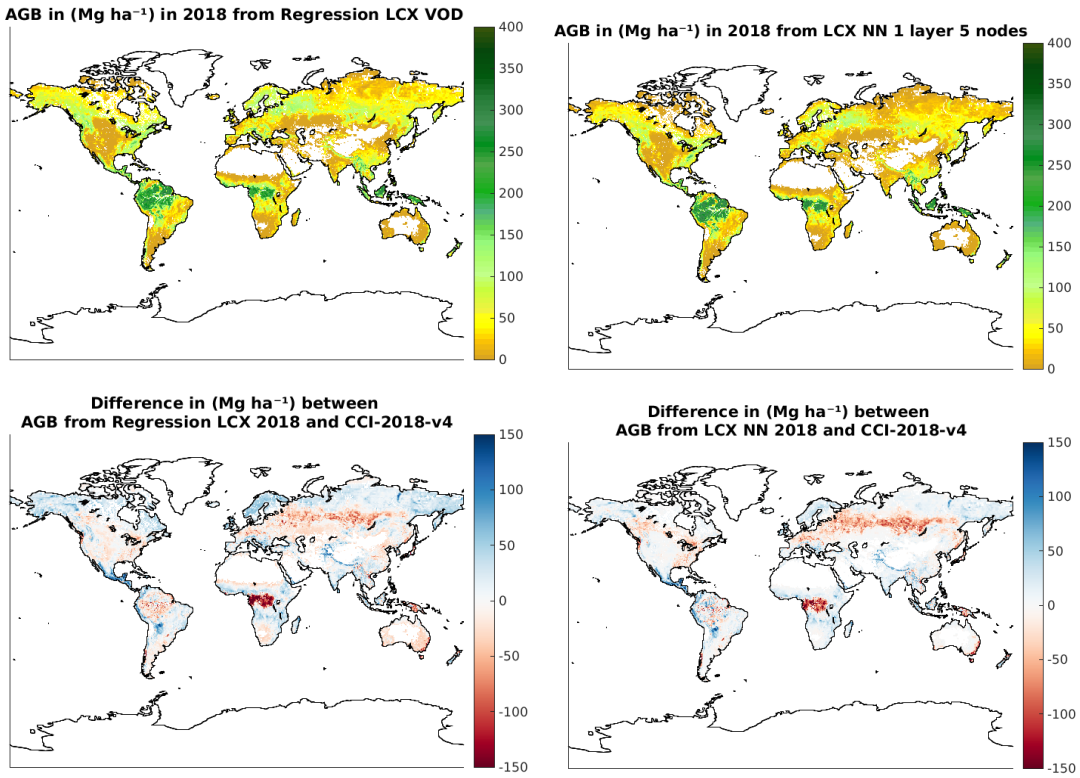


Figure 6.15: Maps of the AGB estimation for year 2018 and differences with respect the CCI map using L-, C- and X-VOD a multi-linear regression (left), a neural network approach (right).

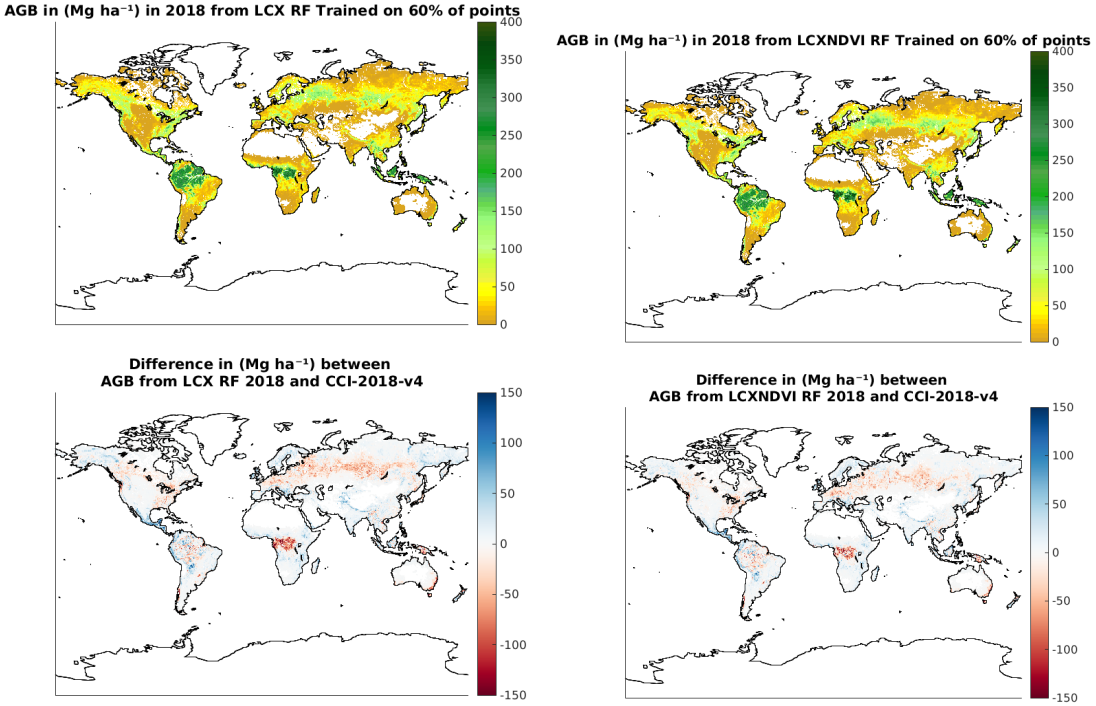


Figure 6.16: Maps of the AGB estimation for year 2018 using a random forest approach with L-, C-, X-VOD as predictors (left) and also NDVI (right).

Table 6.4: Comparison of the performances of different machine learning approaches for the estimation of AGB using multi-frequency data. Multi-linear regression (MultiReg), Neural Network (NN) and Random Forest (RF).

Method	2017	2018	2019	2020
MultiReg	0.88	0.88	0.88	0.88
NN	0.91	0.91	0.91	0.90
RF	0.92	0.95	0.92	0.90

Pearson correlation coefficient of estimated AGB LCXNDVI with CCI

Method	2017	2018	2019	2020
MultiReg	36.35	37.17	36.81	36.69
NN	32.26	32.59	32.57	34.04
RF	30.22	24.5	30.55	33.19

RMSE (Mg ha^{-1}) of estimated AGB LCXNDVI with CCI

Method	2017	2018	2019	2020
MultiReg	2.85	3.43	3.30	3.92
NN	0.03	0.31	0.43	-1.08
RF	-0.37	-0.04	-0.36	-2.1

Bias (Mg ha^{-1}) of estimated AGB LCXNDVI with CCI

6.5 Conclusion

This chapter discusses different ways to estimate AGB from L-band data:

- Using SMOS L-VOD and fitting a logistic function
- Using SMOS brightness temperatures and training a neural network
- Using SMOS L-VOD complemented by X- and C-band VOD and NDVI by multi-linear regression, neural networks and random forest

The first two approaches were evaluated extensively against a large number of AGB data from in situ plots, LIDAR estimations and national inventories.

Most of the published works using AGB estimations from L-band data use L-VOD and a reference AGB map to fit a parametric function. The first alternative methodology for the estimation of AGB from L-band passive microwave observations, which directly exploits the statistical relationships between biomass and multi-angle TB observations using Artificial Neural Networks. The results show that the estimation of AGB from the proposed method presented certain advantages with respect to the method using parametric functions used by most of the works published to date on the subject. The estimated AGB maps, presented a slightly higher correlation than the current methods, with the reference AGB maps (R^2 0.87 instead of 0.84), a stronger agreement with the structural patterns of the reference maps over tropical regions, residual maps with lower absolute values and smaller uncertainties when compared to the state of the art AGB maps. The results suggest that accurate biomass estimates can be produced directly from SMOS L-Band multi-angular TBs. Being the optimal set of predictors for the estimation of AGB directly from TBs: the TBs in H polarization, the PR and the E_h (for angles of incidence $>35^{\circ}$).

In addition, it has been shown that using a single year to establish the relationship of the input variables and AGB introduces significant uncertainties when this relationship is used to study the evolution of AGB/Carbon Stocks. An alternative training scheme was proposed based on multi-year learning sets, which accounted for the differences mentioned above and produced more stable estimates for different years. Despite all of the

above, a more detailed study should be carried out on the possible synergies that the observations of different instruments (frequencies) could have with other canopy components, and on the potential applications that this new methodology may have in the future, like the estimation of carbon stocks or AGB temporal evolution.

The evaluation against a large number of in situ estimations of AGB show a good agreement (R $\sim$ 0.6) with the in situ estimations extended using LIDAR tree height data. The estimations using SMOS brightness temperatures give slightly better results than those starting from LVOD. In contrast, the correlation with respect to in situ plots estimations by *Ploton et al.* (2020b) is very low. This is likely due to the fact that those estimations provided at a 1 km spatial scale come actually from 0.5 ha plots and they are probably not representative of the spatial scale of SMOS observations. It is interesting to remark that the comparison of AGB estimations from SMOS at the national level show a very good agreement with FAO inventories ($R \sim 0.9$), probing that when compared at large scale, AGB estimations from SMOS are very accurate.

Finally, different methods were studied to estimate AGB from multi-frequency VOD with or without NDVI. Random forest showed the best performances of the three methods studied. They also showed hints of providing better AGB estimates than L-VOD alone. More research on this point would be needed to explore the full potential of this promising approach.

Chapter 7

Optical indices (WP 6-7)

7.1 Introduction and purpose

The purpose of this section is to describe the development of the algorithm used to estimate the vegetation water content from optical sensors, as well as the corresponding validation results. The amount of vegetation water content can be assessed through several variables, such as leaf water content (LWC, unitless), which is defined as the total mass of water related to the total fresh or dry biomass of the leaves, and the Equivalent Water Thickness (EWT, g.cm-2) that is defined as the mass of water contained in a unit of leaf area *Ceccato et al.* (2001). At the canopy level, the amount of water in the vegetation can be related to any vegetative living organ (stem, trunk, leaves, reproductive organs etc.) or solely green leaves. Vegetation water content can thus be assessed through several variables such as live fuel moisture content (LFMC, %) *Yebra et al.* (2013) or canopy water content (CWC, gcm-2) also named vegetation water content (VWC, gm-2) *Hornbuckle* (2023). CWC correspond to the product of the canopy Plant Area Index (PAI, e.g. one-sided vegetation element area per unit horizontal ground area, unitless) and EWT, and are related to all the living vegetation elements *Jackson et al.* (1982).
We used MODIS, SENTINEL-2 and LANDSAT-7 Data to derive the CWC products. The validation lies in two steps, first validate the CWC product at the SENTINEL-2 scale against ground measurements that are easier to obtain, and then inter-compare the CWC MODIS product with aggregated S2 CWC. The following represents the activities developed for the tasks (work packages) 6 and 7 of the project (see Chapter 3).

7.2 Review of methods to estimate Vegetation traits from reflectance observations (WP6)

WP6 consisted in reviewing methods for estimating different vegetation traits, including CWC. Since the 80's, considerable improvement in the quality of vegetation traits derived from satellite or airborne systems has been achieved due to the advances of measurement capability of satellite instruments and to our understanding of the radiation regime within vegetation canopies *Liang* (2005). However, remote sensing observations sample a limited fraction of the radiation field reflected or emitted by the surface, and thus do not provide directly estimates of the targeted traits. It is therefore necessary to transform the radiance values recorded by the sensor into vegetation trait estimates using dedicated methods. The methods should ideally be accurate, precise and computationally efficient. Most importantly, they should require minimal calibration since they are supposed to be applied over diverse locations, seasons and conditions *Walthall et al.* (2004).

7.2.1 Biophysical parameter retrieval approaches from optical remote sensing

Many methods have been proposed to retrieve vegetation traits from remote sensing observations *Myneni et al.* (1988), *Goel* (1989), Pinty and Verstraete 1991, XX Kimes et al. 2000; *Baret et al.* (2007), Houborg

and Boegh 2008; *Verger et al.* (2011), *Laurent et al.* (2013). They include empirical methods with calibration over experimental data sets. These simple methods are limited by the size and diversity of the calibration data set as well as by the uncertainties attached to the ground measurements. More complex ones based on the use of radiative transfer models have been proposed where no in situ calibration data set is required. Radiative transfer models describe the physical processes involved in the photon transport within vegetation canopies. They simulate the radiation field reflected by the surface for a given observational configuration once the vegetation and the background are known. Retrieving canopy characteristics from the radiation field as sampled by the sensor aboard satellite needs to 'invert' the radiative transfer model, i.e. to estimate some input variables from the measurement of the outputs of the model. This can be done explicitly with a radiative transfer model, or based on datasets where the satellite measurements are matching the canopy traits targeted as measured or estimated from an independent method. Note that the use of ground-based measurement is generally not possible for coarse resolution sensors because of the very demanding spatial sampling required. The retrieval methods can be split into two main approaches depending whether the emphasis is put on the inputs (the canopy biophysical variables driven approach, e.g. vegetation indices or machine learning approach) or the outputs (radiometric data driven approach, e.g. iterative optimization methods that minimize a cost function, and Look-Up-Tables) of the radiative transfer model Figure 7.1.

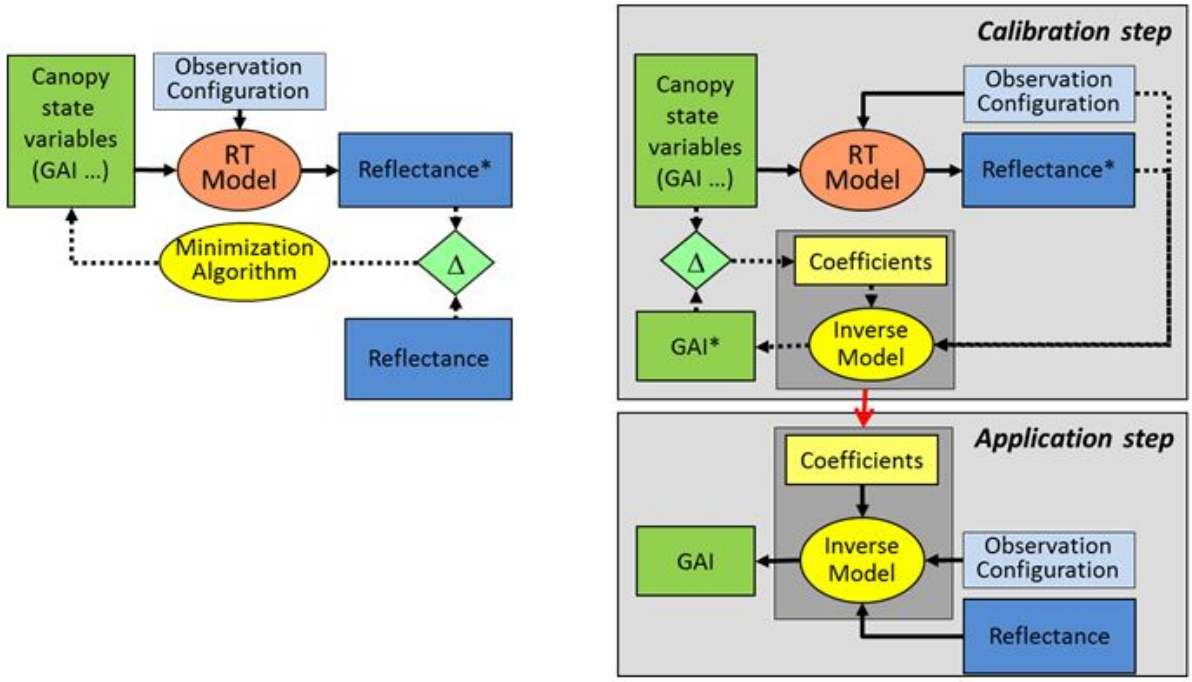


Figure 7.1: The two main approaches used to estimate vegetation traits (exemplified by LAI) from remote sensing data. On the left side, the approach focuses on radiometric data showing the solution search process leading to the estimated LAI value, LAI*. On the right side, the approach focuses on the biophysical variables showing the calibration of the inverse model (top) and the application using the inverse model with its calibrated coefficients (bottom). Δ represents the cost function to be minimized over the biophysical variables (right) or over the radiometric data (left).

In a recent study, *Wang et al.* (2022) propose to use a Bayesian scheme to estimate some vegetation traits including GAI from observations of a range of satellites, allowing to assess a better prior information regarding the distributions and co-distributions of the radiative transfer model input variables. Domain adaptation techniques based on using GAN (Generative Adversarial networks) could also be investigated to make radiative transfer model simulation better reflect the actual measurements. Finally, hybrid methods *Dong et al.* (2024) and ensemble products are also interesting solutions to achieve better accuracy on the products *Xiao et al.* (2014).

7.2.2 Pro and cons of the different approaches

7.2.2.1 Computation requirements

Machine learning approaches, once calibrated, are obviously very little demanding in terms of computation. The inverse model is generally relatively simple and could be run very quickly. However, the calibration (or learning or training) process could require large computer resources, particularly for complex parametric model with a significant number of coefficients to be tuned and when the training dataset is large. The implementation of a LUT technique in algorithmic operational chains is very efficient because the radiative transfer model is run off-line. Conversely, iterative minimization methods require large computer resources because of its iterative nature. Improvements are possible using a meta-model. Further, automatic segmentation or discretization of the reflectance space (Pinty et al., 2011) will also reduce the number of inversions to be completed over a whole set of images.

7.2.2.2 Flexibility of the observational configuration

Iterative optimization methods allow retrieving canopy characteristics from several observational configurations. It is even possible to invert radiative transfer models concurrently over several pixels. This opens great potentials for exploiting additional temporal or spatial constraints *Atzberger* (2004); *Lauvernet et al.* (2008). LUT could theoretically cope with variable configurations at the expense of the dimensionality and thus the size of the tables, making them more difficult to manipulate. Conversely, machine learning methods require a fixed number of inputs. The characteristics of the configuration need thus to be used as inputs of the inverse parametric model, where the illumination and view directions are explicitly used . However, this increases the dimensionality of the system, making the calibration step more demanding and more difficult. One alternative is to calibrate several parametric models for each individual configuration and then select the proper calibrated inverse model.

7.2.2.3 Integration of prior information

The radiometric data driven approaches integrate the prior information directly in the cost function using a regularization term *Combal et al.* (2002). However, in the case of LUTs, it is also possible to restrict the simulations to the range of situations to be encountered as is done within the MODIS LAI and FAPAR algorithm that depends on the biome type considered *Shabanov et al.* (2005). For the machine learning approaches, the prior information is introduced through the distributions and co-distributions of the inputs of the radiative transfer model: when vegetation traits have to be estimated under situations where the type of canopies and their stage of development are known, it is more efficient to calibrate a specific inverse model for each individual situation. Note that *Qu et al.* (2008) proposed to use Bayesian networks where model simulations could be exploited along with a description of the distribution of the variables that may depend on growth stages or canopy types.

7.2.2.4 Associated uncertainties

The radiometric data driven approaches allow getting some estimates of the uncertainties associated to the solution by propagating the uncertainties associated to the measurements and to the model using the partial derivatives of the cost function with regards to the measurements (Lauvernet 2005). When using LUTs, uncertainties could be estimated by Monte-Carlo methods or approximated by the standard deviation of the ensemble of solutions defined by the uncertainties in the measurements (Knyazikhin et al. 1998). For machine learning methods, the error on the measurements may be assessed in different ways as proposed by Aires et al. (2004), which may also include the errors associated to the retrieval process itself. A more simple alternative solution is also proposed by *Baret et al.* (2013) based on the training dataset. Although the estimation of uncertainties of the retrievals is possible, it is generally limited by the poor knowledge on the input uncertainties associated to the reflectance measurements and radiative transfer models used. *Knyazikhin et al.* (1998) used a 20% relative uncertainty applied to MODIS top of canopy reflectance for LAI and FAPAR retrieval. *Baret*

et al. (2007) propose to use an additive uncertainty around 0.05 and a multiplicative uncertainty around 3%. This example shows that the uncertainties attached to each band is poorly known. Further, the structure of the uncertainties may also play an important role and is unfortunately very difficult to describe.

7.2.2.5 Robustness of the retrieval and quality assessment

A quality index needs to be associated to the retrieved values to inform about the status of the inversion process. For iterative optimization techniques, it could be criterions relative to the stop of the iterations (Gilbert 2002). As a matter of facts, the algorithm may sometimes encounter numerical problems occurring generally with very small values of J. No numerical problems are expected for LUT and machine learning approaches, and the quality index should mainly indicate whether the input reflectances were inside the definition domain and if the output solution is in the expected range of variation *Baret et al.* (2013). The performances of the approach will both depend on the minimization algorithm itself and on the level of illposedness of the inverse problem as a function of measurement configuration and model and measurement uncertainties.

7.2.3 Using VIs or reflectance for biophysical parameter retrieval?

The literature on the subject started about 30 years ago focusing on a restricted number of VIs available. At that time, VIs were the main approach used to characterize the vegetation. With the multiplication of satellite systems and the progress of our understanding of the physical principles governing the radiative transfer in vegetation canopies, other approaches were proposed as described in the previous section. VIs are combinations of reflectance in several bands, generally limited to 2 or three bands. They therefore exploit only the spectral dimension of the observation. They are using the differential absorption properties of the vegetation and background materials that govern the spectral variation of canopy reflectance. However, if canopy reflectance depends on the vegetation traits of interest, several other factors should be considered like soil background, atmosphere with main effects for the shorter wavelengths, sun/sensor geometry with the bidirectional effects, topography effects, scaling effects when the pixel is composed of patches containing contrasted objects, canopy architecture when the targeted trait is the biochemical composition or the biochemical composition when the targeted trait is related to canopy architecture.

VIs are therefore designed empirically with the objective to maximize the sensitivity to the given vegetation trait targeted, while minimizing the impact of confounding factors such as the soil background, atmospheric effects, topography effects, or the effect of other vegetation characteristics.

VIs have been used since the 1970's. At that time, even if the physical principles of the radiative transfer in vegetation canopies was well understood, there was no mature models and the required computer resources to use them to calibrate retrieval algorithms. VIs correspond to a parsimonious way to combine few spectral bands into a single quantity to be related to a vegetation variable of interest. Simple statistical models can be therefore calibrated over a small training dataset. For coarse resolution satellites for which the collection of ground data is difficult, VIs offered a convenient way to build relatively consistent long time series of vegetation observations.

The advancement of the radiative transfer models, the availability of high-quality and high spatial and temporal resolution satellites, the development of machine learning techniques, including deep-learning allows nowadays to develop more sophisticated approaches that will fully and optimally exploit the spectral information available to get more robust and accurate estimate of the few main vegetation traits. The reader is referred to WP6 report, section 4, *Baret and Weiss* (2024) for a full description of the development of VIs, how they are related and their sensitivity to vegetation traits, and how to mitigate with confounding effects to better assess biophysical variables from optical measurements.

7.3 Designing the algorithm for CWC retrieval from optical measurements

This study has the objective of proposing global and consistent CWC products at both decametric (Landsat-8, Sentinel-2) and 500m (MODIS) scale. For this purpose, we relied on former developments for operational algorithms *Bacour et al.* (2006) ; Baret et al. 2007b), including the one developed for the S2 SNAP toolbox *Weiss et al.* (2023) and developed a unified approach based on machine learning trained with RTM simulations, with particular attention to the following aspects: (1) Optimizing the representativeness of the training database by combining the knowledge extracted from two open databases (e.g. the newest version of TRY for plant traits and OSSL for soil background); (2) Proposing an appropriate validation scheme by firstly assessing CWC at decametric resolution that are much more compatible with the ground measurement footprint using ground data covering diverse biome types over the world, and then comprehensively evaluating the MODIS CWC using Landsat8 and Sentinel-2 CWC at the global scale; (3) Testing the effectiveness of optical indices and different band combinations for optimizing CWC retrievals more widely, and comparing the performances of CWC retrievals with those of VOD retrieved from the passive microwave sensors. The reader may also refer to the paper by Ma et al, submitted to Remote Sensing of Environment, August 2024.

7.3.1 Material

Here, we present the satellite optical data that are used as inputs to the CWC retrieval algorithm, then the databases employed to build a realistic training dataset using the PROSAIL radiative transfer model, and finally the site network that we used to evaluate our algorithm, including ground validation sites and those used for the inter-comparison exercise between different satellite products. In this way, we followed the recommendations of the CEOS (Committee on Earth Observation Satellite) LPV (Land Product Validation) group for vegetation biophysical variables.

7.3.1.1 Optical satellite data

In this study, two high resolution satellites including Sentinel-2 and Landsat-8 (20m-30m), as well as the moderate resolution satellite MODIS (1km) are used. For each sensor, we used surface reflectance L2A data at the time of acquisition date for Landsat-8 (https://www.usgs.gov/) and Sentinel-2A and 2B (https://scihub.copernicus.eu/). For MODIS, we considered the MCD43A4 Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted Reflectance (NBAR, V6.1) product at 500m spatial resolution. This product results from a kernel driven model normalization in the nadir direction over a 16-day period centered at the 9th day. For Sentinel-2A and 2B, the 10m green (B3) and red (B4) bands were resampled at 20m through spatial averaging, to be consistent with the spatial resolution of the other considered bands. we used a single NIR band for Sentinel-2, e.g., B8a which has a closer spectral range to Landsat-8 than B8 to keep better consistency between the sensors. It can be also noted that, in the SWIR domain (e.g. CWC sensitivity bands), MODIS has an additional band (i.e., SWIR3, 1230nm 1250nm) compared to Sentinel-2 and Landsat8. Additionally, the geometry information including sun zenith angle, view zenith angle and relative azimuth angle (i.e., solar azimuth minus view azimuth) were extracted for the three satellites. The three sensor surface reflectance products are derived with different algorithms due to both the inherent characteristics of the sensors (e.g., presence of a thermal band for the detection of clouds for Landsat-8 and MODIS) and choice made for the processing (e.g., atmospheric correction scheme, cloud screening, 16 days synthesis for MODIS). The largest differences are related to the cloud screening since different cloud definitions are used (e.g. thin cirrus, *Skakun et al.* (2022). Additionally, the compositing period and the coarse spatial resolution of MODIS may affect the surface reflectance product when small clouds occur within the 500m pixel *Roy et al.* (2008). To avoid such situations and inconsistencies between products, we filtered the data by considering that NDVI values should be consistent between the three sensors at the same acquisition date for a same location. For the same reasons as for the cloud influence, we also used the MODIS snow mask.

7.3.1.2 TRY and OSSL databases

As the biggest global effort to comprise a massive repository of global plant traits, TRY is capable of representing the actual leaf traits variations at an unprecedented spatial, temporal and climatic coverage *KATTGE et al.* (2011), *Kattge et al.* (2020). Since its establishment in 2007, TRY has grown continuously, thus improving representativeness for the terrestrial leaf traits, and we used the data available up to April, 2023. We extracted three key related leaf traits (leaf chlorophyll content, Cab, leaf dry matter content, Cdm, and LWC) that were used to determine the distribution of the PROSPECT model inputs after removing potential outliers by keeping only the data between the first and 99th percentile to 99% for each trait. With regard to the soil spectral data, the CWC retrieval algorithm relies on former developments, which used a soil spectral database collected by INRAE Avignon over different sites distributed all over France *Weidong et al.* (2002). However, some weaknesses were observed by validation studies of the SNAP algorithm for low vegetation areas *Vinué et al.* (2018), *Estévez et al.* (2020), which indicated that this database is not sufficiently representative of the diversity of the soil at the Earth surface. We therefore combined it with OSSL, another open access and massive database for soil spectral properties at the global scale *Hengl et al.* (2021). We resampled the OSSL soil spectra using a cubic convolution interpolation approach into 1nm spectral resolution, before combining it with the INRAE database. The resulting dataset of 35835 soil spectral samples was then exploited to improve the representativeness of the soil spectral background in our training database.

7.3.1.3 Ground CWC data

We evaluated the decametric CWC products using ground destructive measurements from five ground sites representing 3 biome types (2 grasslands, 2 forests and 1 cropland), distributed in diverse continents (Table 7.1, 7.1) by matching the Sentinel-2 and Landsat-8 acquisition dates with the date of measurements (+/- 2 days) during which we assume CWC exhibits very slight variations at the satellite overpass time. For the sites of Dahra, SMAPVEX16 and EucFACE, both Sentinel-2 and Landsat-8 could be used, while only Landsat-8 was present for VALSE2 and Qinghai with earlier observation periods. Regarding the spatial matching, the Dahra site can be considered as a homogeneous grassland, and the 28 CWC measurements of $1m^2$ plot were averaged *Tagesson et al.* (2015); for EucFACE, the ground CWC was measured in rings of 25m diameter *Quan et al.* (2017) *Griebel et al.* (2023); while for Qinghai sites, the ground CWC was acquired at each 30m*30m plot. The size of sample range for these two sites are close to one pixel of Sentinel-2 (20m) or Landsat-8 (30m). In terms of temporal matching, only the decametric optical satellite data were extracted at the dates of the ground measurements (+/- 2 days). Here it should be noted that for SMAPVEX16, ground CWC of diverse crop types including soybean, wheat, canola, oat, and corn etc. were collected. As also pointed out in existing studies by *Djamai et al.* (2019) and *Garcia-Haro et al.* (2020), the ground data in this area were calculated by the difference of wet and dry weights of leaves (for soybean, canola and corn) or of combinations of leaves and stems (for wheat and oat). Moreover, the data for canola during the crop flowering stage (DOY from 180 to 203) were not considered for the CWC assessment, as the flowers may be absent in PROSAIL simulations adopted in this study. Unfortunately, we could not access additional sites or data via destructive approaches to complement our validation exercise at this spatial scale, as we tried to make our best efforts to guarantee the representativeness and broadness of the CWC assessment in this study.

Table 7.1: Ground CWC sites

Ground Sites	Continent/Country	Time range used	Matching satellite	Landcover	Reference
Dahra	Africa/Senegal	2013 July-2019 October	Landsat-8/Sentinel2	Grassland	Tagesson et al. 2015
SMAPVEX16	America/Canada	2016 June-2016 July	Landsat-8/Sentinel2	Cropland	McNairn et al. 2017
VALSE2	Europe/France	2013 June-2013 September	Landsat-8	Forest	Weiss and Baret 2013
EucFACE	Oceania/Australia	2019 September-2021 June	Landsat-8/Sentinel2	Forest	Griebel et al. 2023
Qinghai	Asia/China	2014 July-2015 August	Landsat-8	Grassland	Quan et al. 2017

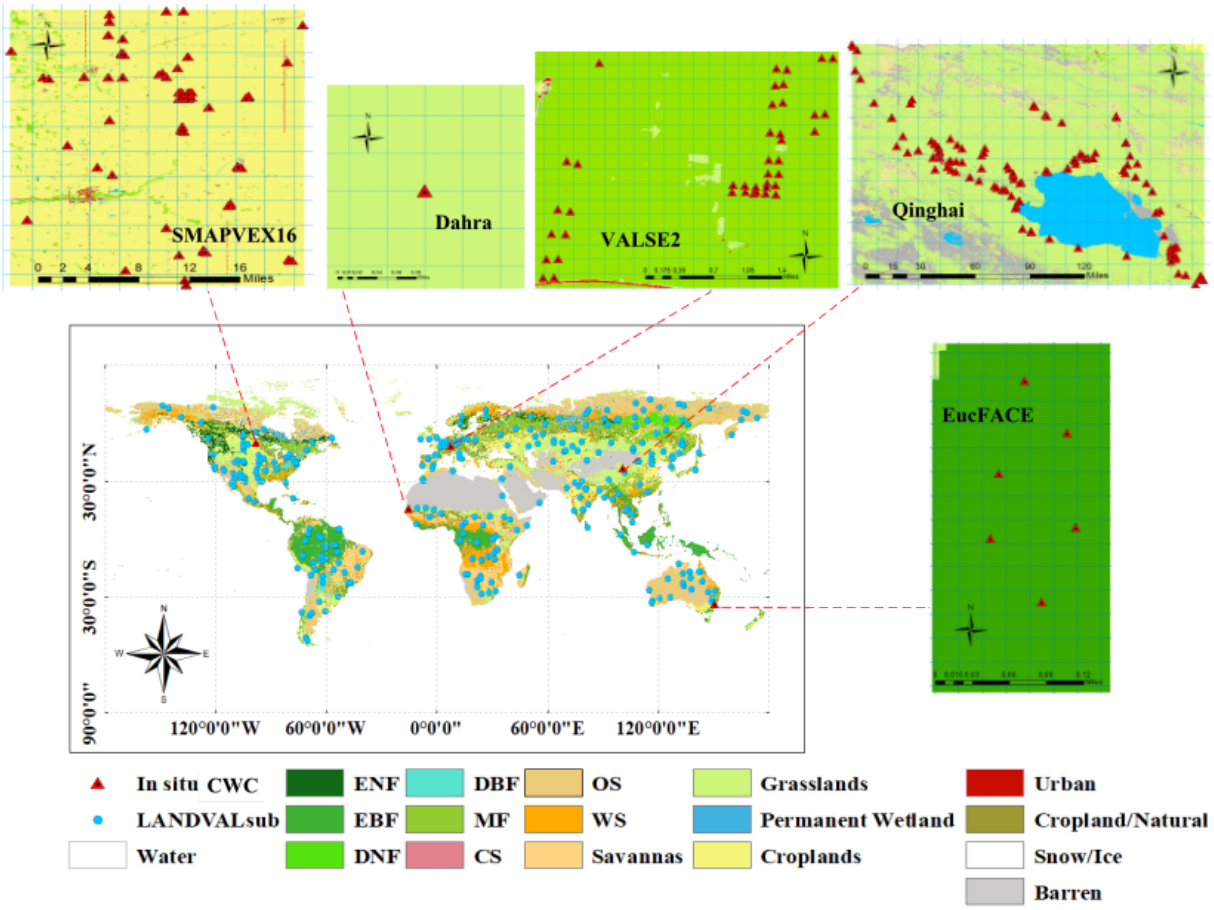


Figure 7.2: Distribution of the LANDVALSUB and ground validation sites used in this study. The 500m resolution IGBP map is shown as biome type description (Friedl et al. 2002). For the specific presentations of each ground sites, the ESA WorldCover 10m 2022 was adopted in this study (Zanaga et al. 2022). DBF: Deciduous Broadleaf Forest; WS: Woody Savannas; ENF: Evergreen Needleleaf Forest; MF: Mixed Forests; EBF: Evergreen Broadleaf Forest; CS: Closed Shrublands; DNF: Deciduous Needleleaf Forest; OS: Open Shrublands. Selection of LANDVAL sites

7.3.1.4 Selection of LANDVAL sites

The satellite product inter-comparison exercise uses the LANDVAL network of 725 sites which was specifically developed for benchmarking vegetation products such as LAI, fAPAR, or albedo *Fuster et al.* (2020). LANDVAL is a compilation of different site networks that well represent the combination of climate, vegetation types and environment at the Earth Surface. Additionally, these 3kmx3km sites were selected to be homogeneous in terms of biome type, not too close to a water body, flat, and do not correspond to urban areas. In this study, to reduce the extraction and computation requirements as well as guarantee the representativeness of diverse biome types in the world, we extracted a selection of 308 sites that we named LANDVALSUB sites Figure 7.2. Firstly, we divided the Earth into rectangles of 20°in latitude (from -60^o to 80^o) and 40^o in longitude (from -180^o to 180^o). We identified the land covers of all the LANDVAL sites present in each rectangle, and then randomly selected half of them for each biome class, except for the desert and bare soil sites. Moreover, to further check the variations of satellite CWC in low water conditions, we additionally selected 7 sites for bare soil or deserts (i.e., CWC close to zero). For each site, we then spatially averaged the CWC products at their original resolution over the 3km by 3km area. Additionally, as the MODIS product results from a synthesis over a 16-day temporal window, and considering the low revisit frequency of Landsat (16 days), we considered a 5-days tolerance for the benchmarking of the optical CWC products over the LANDVALSUB sites.

7.3.2 Method

We developed a generic and consistent algorithm to derive an operational CWC from any optical sensor. This algorithm relies on former findings related to the use of machine learning trained on radiative transfer simulations to operationally derive different vegetation biophysical variables products from medium resolution (e.g., hectometric, and kilometric) or decametric sensors We developed a generic and consistent algorithm to derive an operational CWC from any optical sensor. This algorithm relies on former findings related to the use of machine learning trained on radiative transfer simulations to operationally derive different vegetation biophysical variables products from medium resolution (e.g., hectometric, and kilometric) or decametric sensors *Weiss et al.* (2002), *Baret et al.* (2007), *Verger et al.* (2009), *Brown et al.* (2021). We set up the training database of the machine learning algorithm through the PROSAIL radiative transfer model simulations run with a priori distributions input variables partly provided by the TRY and OSSL open databases (Figure 7.3,left). Then, by considering each sensor spectral responses, and adding a realistic noise to simulate the actual satellite TOC reflectance, we tested different band and optical indices combinations as inputs to the machine learning algorithm and selected the best one evaluated on a synthetic database made of independent PROSAIL simulations (Figure 7.3, middle). Finally, we evaluated the CWC retrieval algorithm by assessing the accuracy of the Landsat-8 and Sentinel-2 products against ground measurements, inter-comparison the different optical CWC product on the LANDVALSUB site network, and comparing the spatio-temporal trends of MODIS CWC and microwave VOD products (Figure 7.3, right).

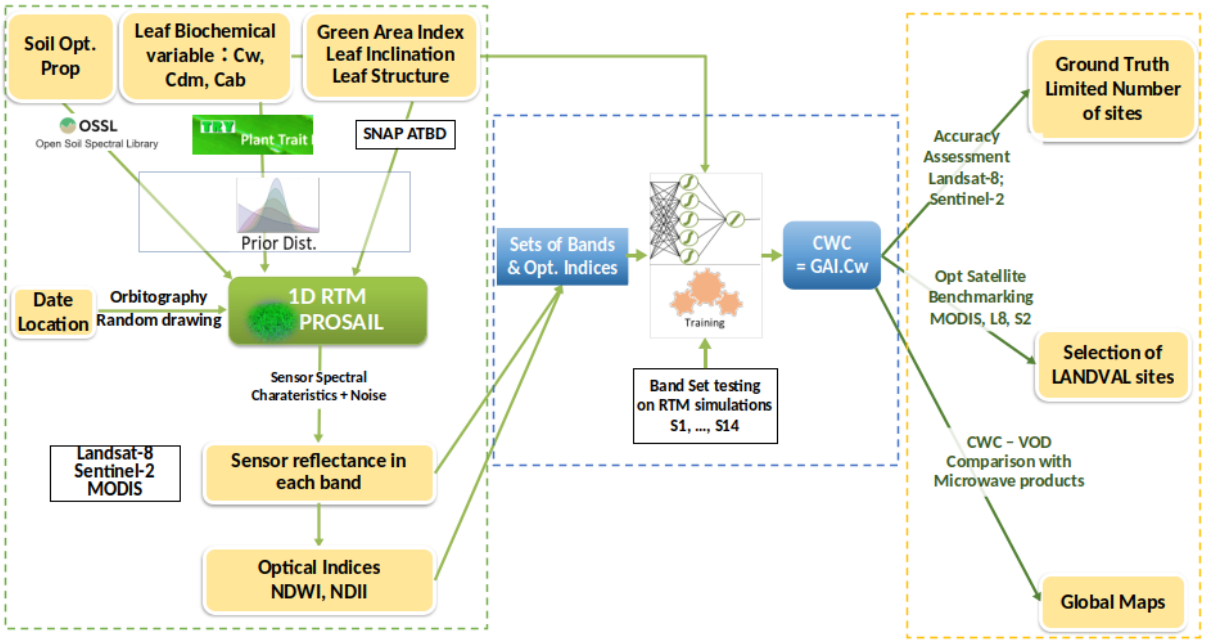


Figure 7.3: lowchart of CWC retrieval. The whole processing is separated into three main parts: (i) in green dashed frame, the parameterizing for PROSAIL model, concerning soil background, leaf traits as well as other essential variables; (ii) in blue dashed frame, the relationship building between the simulated TOC by PROSAIL and CWC; (iii) in yellow dashed frame, the assessment and discussions for the retrieved CWC from optical satellites.

7.3.2.1 Setting up the PROSAIL training database

By coupling the leaf optical properties model PROSPECT and the canopy reflectance model SAIL, PROSAIL has been adopted for several vegetation parameter retrievals from optical satellite observations *Jacquemoud et al.* (2009)and was successfully used for operational kilometric vegetation products at the global scale, despite some inconsistencies between some biome type structure and the model turbid medium assumption

Weiss et al. (2007). As compared to more complex models (e.g. 3DRTM) and besides the fast computing aspect, it presents the main advantage of having a small number of inputs which greatly eases the setting up of a training database representing the distribution of the vegetation surface reflectance at the Earth surface. Moreover, most of PROSAIL input parameters have been extensively studied and correspond to measurable variables. Additionally, CWC can be easily computed from PROSAIL inputs as the product of the green leaf area index (LAI or Green Area Index, GAI, cm2/cm2) and the green leaf water content per green area (Cw, g/cm2). We therefore took a particular attention to the distribution of PROSAIL inputs to improve the representativeness and the realism of the training database. As suggested by *Garcia-Haro et al.* (2020), we used the latest version of the TRY database to simulate 3 PROSPECT input variables: leaf chlorophyll content, leaf dry matter content, and leaf water content (LWC) instead of the relative water content (e.g. leaf water content divided by fresh biomass). Using LWC allows to keep consistency between the leaf dry matter and water content and directly simulate realistic co-distributions. Additionally, we did not consider weighting the extracted TRY distributions by PFT occurrence at the Earth surface as *Garcia-Haro et al.* (2020) showed little differences between both distributions. Therefore, we used directly the kernel density estimation (KDE) method to approximate the probability density function (PDF) of each leaf trait in TRY (Figure 7.3, Middle). The distribution of the other variables required in PROSAIL were kept the same as in the SNAP algorithm. The readers can refer to details in the ATBD document *Weiss et al.* (2023).

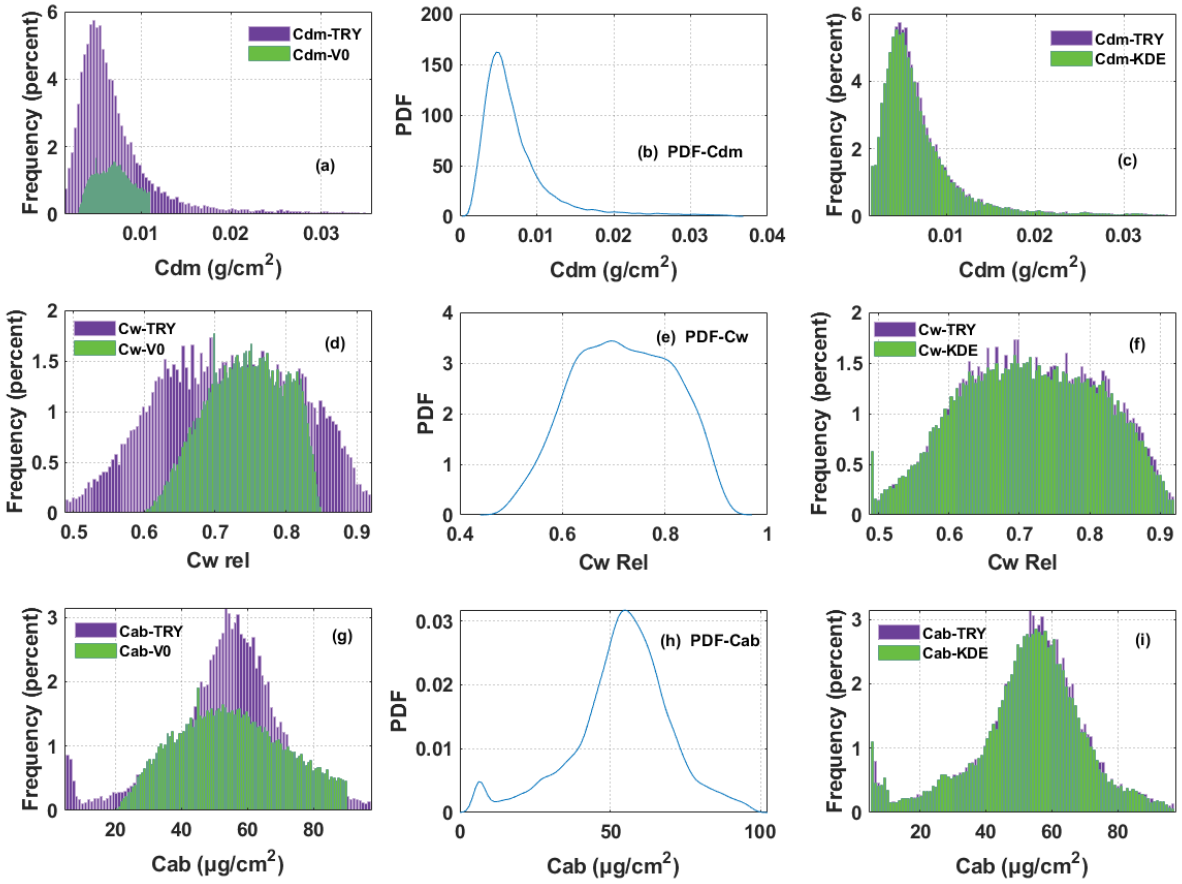


Figure 7.4: Distributions of PROSAIL inputs extracted from the TRY database: Cdm (g/cm2, top), LWC (unitless, middle) and Cab (μg/cm2, bottom). Left: former distribution in the SNAP used in the SNAP algorithm (green) compared to the distribution in TRY, middle: pdf distribution of each trait approximated by KDE, right: final distribution used for the training database (Green). Note V0 means the algorithm version used in SNAP.

Similarly, we used the OSSL soil database to simulate the soil background. However, to avoid some over-representation of similar soil spectra, we first reduced its dimension by using the soil brightness concept,

considering that the reflectance of each soil of the database in a given band can be estimated as a brightness multiplied by a reference soil spectrum, selected among an ensemble of n reference soil spectra. The n soils then constituted the database of soil reference value for building the PROSAIL training database. We used an iterative approach to determine the n soils by applying a k-means classification (with k=1,2,...n) and selecting the k soil whose spectrum was the closest from the centroid of each class. The optimized number of classes was determined by computing the RMSE between each soil spectrum and the one reconstructed using the brightness parameter and the reference value. We found that n=47 among the 35835 spectra was sufficient to well represent the range of spectral shapes within a good accuracy (RMSE around 0.01), and the related figures could be seen in Weiss et al (2023) (Weiss et al. 2023).

7.3.2.2 Training the machine learning algorithm

Many inter-comparison studies have been performed between different machine learning regression algorithms. While, the literature concurs on the importance of building a comprehensive training and realistic dataset, there is no clear consensus about which algorithm performs best. We therefore decided to keep the Neural Networks used in the SNAP algorithm as this is the only method implemented in operational contexts and comprehensively evaluated. However, as the neural networks convergence depends on the initialization of weights and biases that is done randomly, we considered an ensemble approach by training 12 neural networks and use the corresponding median value as the final estimate. The number of 12 was optimized over the testing datasets. This allows to regularize the inversion process and get more stable estimates over time and space, we set up the training database of the machine learning algorithm through the PROSAIL radiative transfer model simulations run with a priori distributions input variables partly provided by the TRY and OSSL open databases (Figure 4). Then, by considering each sensor spectral responses, and adding a realistic noise to simulate the actual satellite TOC reflectance, we tested different band and optical indices combinations as inputs to the machine learning algorithm and selected the best one evaluated on a synthetic database made of independent PROSAIL simulations (Figure 3, middle). For this part, the reader is referred to the paper as the theoretical performances of diverse band combinations were found very similar, whether using the TOC only or inclusion of optical indices. To keep spectral consistency between sensors, we used B4+B8a+B11+NDII+Angles for S2, B4+B5+ B6 +NDII+Angles for LANDSAT8 and B1+B2+B6+NDII+Angles for MODIS. Finally, we evaluated the CWC retrieval algorithm by assessing the accuracy of the Landsat-8 and Sentinel-2 products against ground measurements, inter-comparison the different optical CWC product on the LANDVALSUB site network (Figure 3,, right). For the evaluation of CWC, four error metrics including root mean square error (RMSE), unbiased RMSE (ubRMSE) or precision, Bias or accuracy, and correlation coefficient (R) were considered in this study.

7.3.3 Evaluation of the CWC algorithm performances

7.3.3.1 Direct validation of satellite CWC estimates against ground measurements

Based on the evaluation results against ground sites as shown in Figure 7.1, regarding the correlation coefficient with ground CWC, the satellite retrievals perform the best in grasslands (Dahra and Qinghai) both in terms of R and RMSE,then followed in croplands (SMAPVEX16), and forests (VALSE2 and EucFACE). As for the temporal variations of ground CWC, in spite of the same best performance in grasslands, we obtained smaller RMSE values in forests than those in croplands. In general, the satellite CWC retrievals in this study achieved acceptable performances in various vegetated conditions, which confirms the effectiveness of PROSAIL model for CWC estimations using optical data found by *Djamai et al.* (2019) and *Garcia-Haro et al.* (2020).

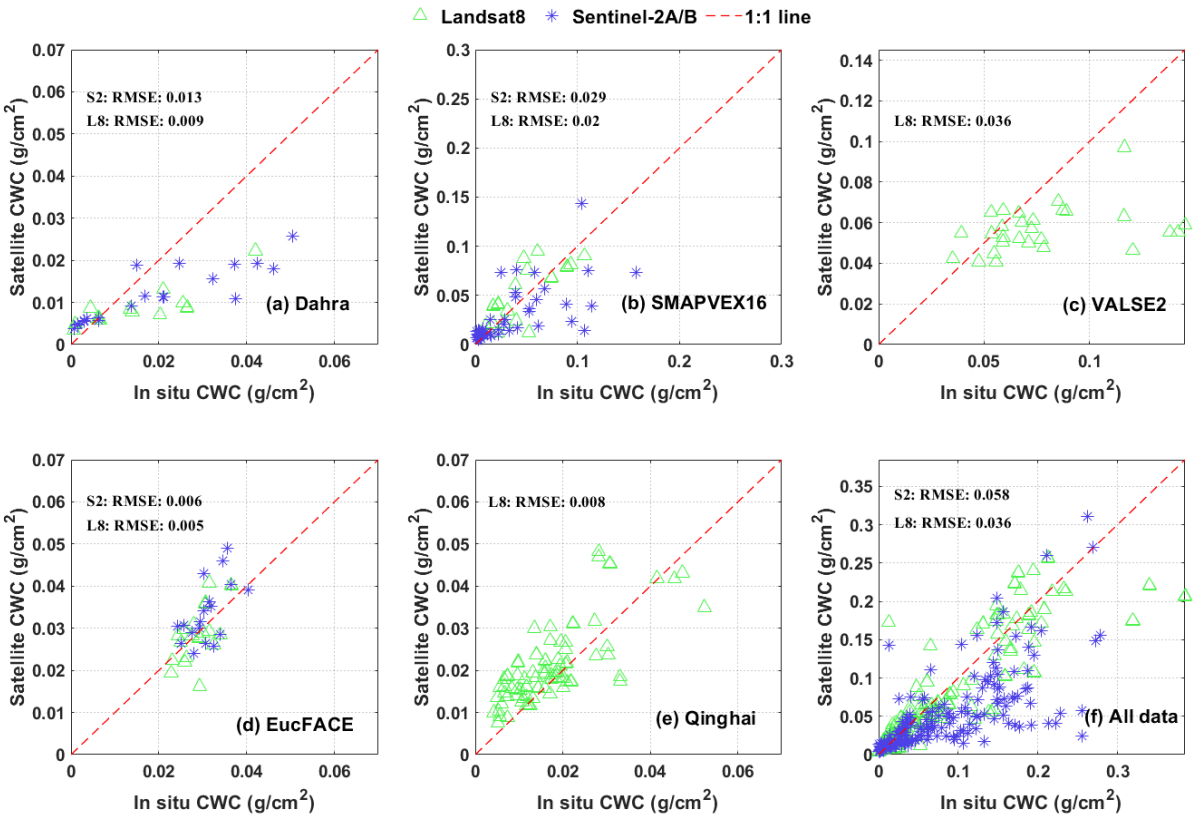


Figure 7.5: Satellite CWC estimates against ground measurements. The green triangle and blue snowflake represent Landsat8 and Sentinel-2A/B, respectively. (a) Dahra and (e) Qinghai are grassland sites; (b) SMAPVEX16 is cropland site; and (c) VALSE2 and (d) EucFACE are forest sites. (f) All data represents the scatter plot when putting data of all the sites together.

7.3.3.2 Benchmarking Sentinel-2, Landsat-8 and MODIS CWC products

The MODIS CWC product at 500m cannot be compared to ground measurements because of its large spatial resolution that hampers ground measurements at this footprint. Therefore, we performed a comprehensive comparison over several biome types by using the LANDVALSUB sites for year 2019, by averaging all the pixels within the 3km*3km area for three satellites. Figure 7.1 and Figure 1 show good consistencies between MODIS CWC and Landsat-8 or Sentinel-2 CWC for most of the biome types. The correlation is higher than 0.8, except for bare soil area (BA, e.g., desert area) and Evergreen Broadleaf Forest (EBF). The comparison between the two decametric sensors shows a better result than MODIS (Figure 8), achieving a correlation coefficient around 0.9, even for all biomes, including the BA and EBF. This indicates the strong robustness of our algorithm between the different satellites. The differences and uncertainties that are observed may be attributed to three features: (i) the scaling effect due to the nonlinearity of the relationship between CWC and reflectanceLandsat-8 (ii) the differences between the pre-processing chains of the satellites, especially regarding the cloud screening and the compositing period. This is quite well illustrated for EBF that are characterized by small clouds often mis-detected at coarse spatial resolution since the signal results from the mixing of clouds and surface reflectance. Additionally, EBF are located near the Equator and are affected by very high cloud occurrence that may impact the retrieval of Nadir reflectance from the kernel BRDF model inversion over the 16 day compositing period, while we considered the closest acquisition date when comparing Landsat-8 and Sentinel-2.

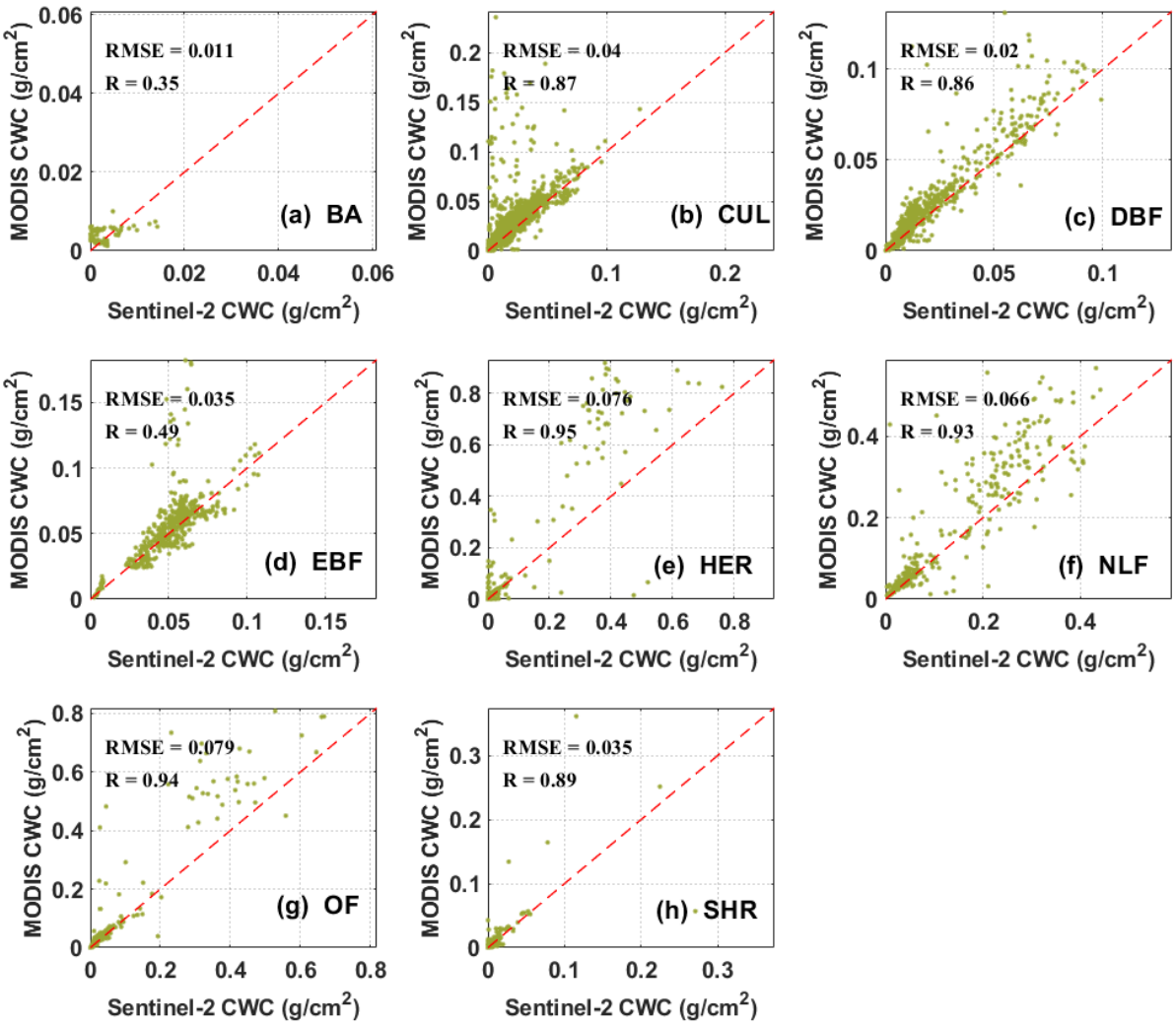


Figure 7.6: Scatter plots for CWC between high resolution Sentinel-2 and moderate resolution MODIS in diverse biome types. BA is Sparse and Bare Areas; CUL is Cultivated lands (agricultural lands); DBF is Deciduous Broadleaf Forest; EBF is Evergreen Broadleaf Forest; HER is Herbaceous (e.g., grasslands); NLF is Needle-leaf Forest; OF is Other forests, including Mixed Forest; and SHR is Shrublands.

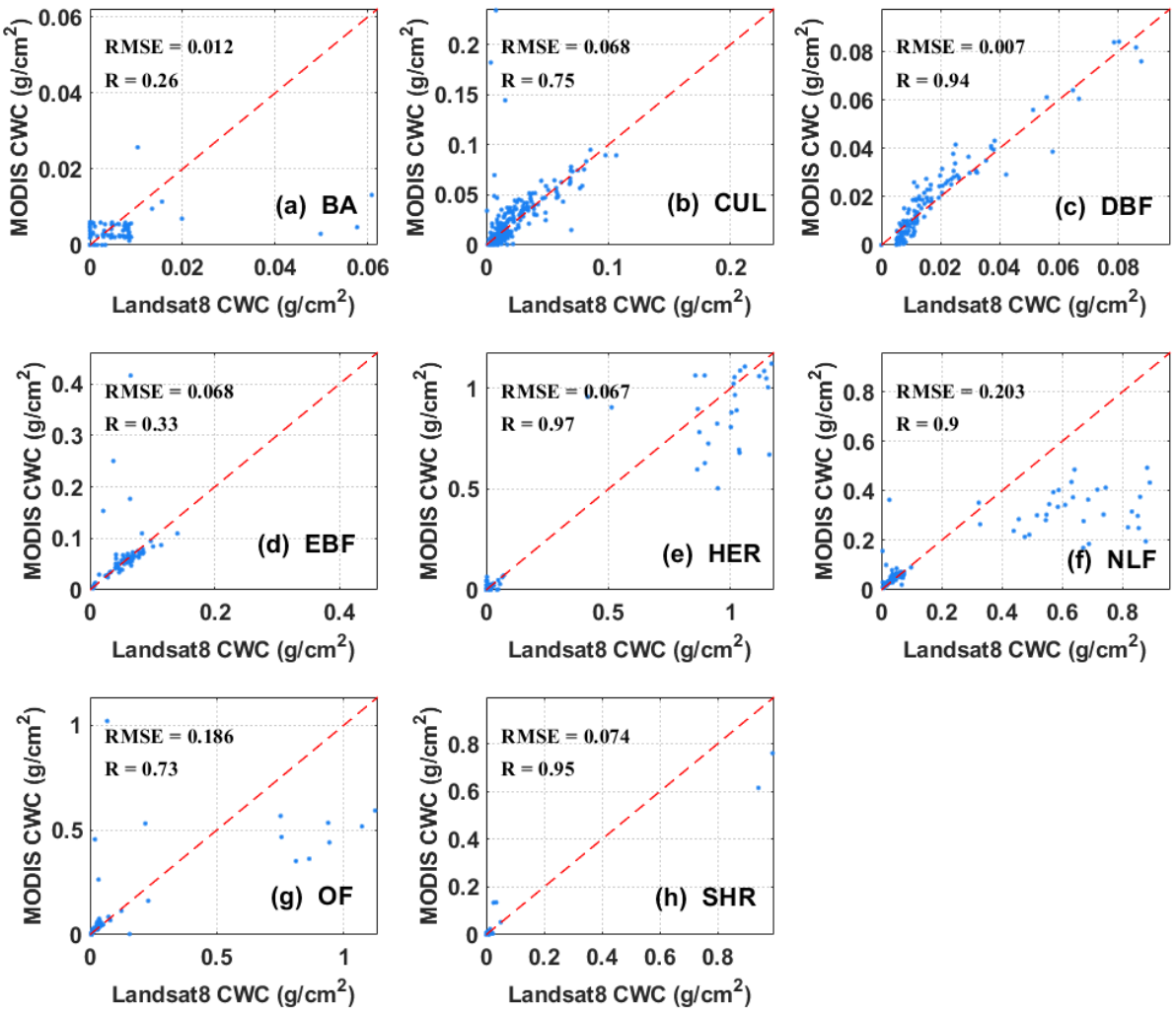


Figure 7.7: Scatter plots for CWC between high resolution Landsat8 and moderate resolution MODIS in diverse biome types. BA is Sparse and Bare Areas; CUL is Cultivated lands (agricultural lands); DBF is Deciduous Broadleaf Forest; EBF is Evergreen Broadleaf Forest; HER is Herbaceous (e.g., grasslands); NLF is Needle-leaf Forest; OF is Other forests, including Mixed Forest; and SHR is Shrublands.

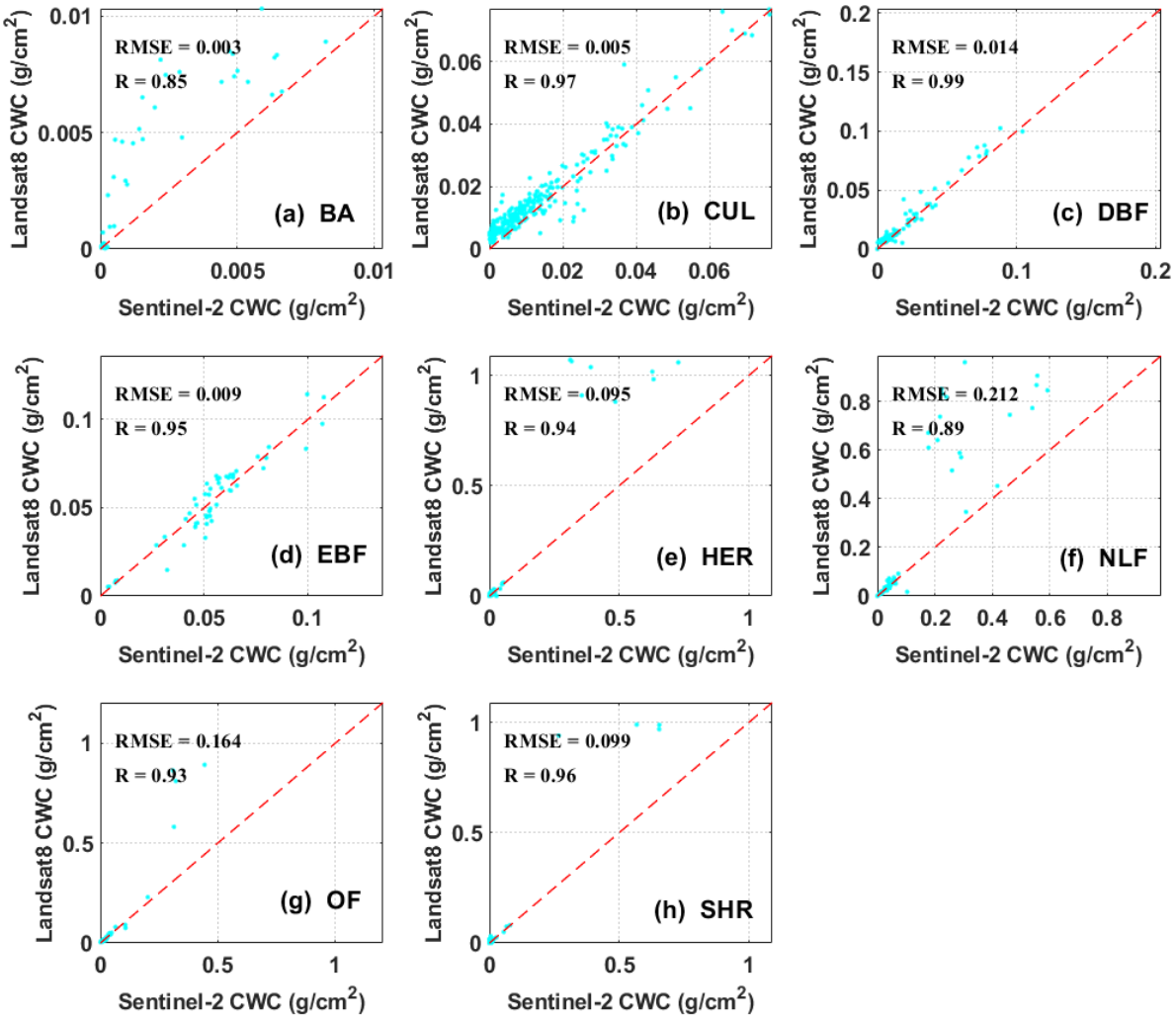


Figure 7.8: Scatter plots for CWC between high resolution Sentinel-2 and Landsat8 in diverse biome types. BA is Sparse and Bare Areas; CUL is Cultivated lands (agricultural lands); DBF is Deciduous Broadleaf Forest; EBF is Evergreen Broadleaf Forest; HER is Herbaceous (e.g., grasslands); NLF is Needle-leaf Forest; OF is Other forests, including Mixed Forest; and SHR is Shrublands.

7.4 Conclusions

In contrast to other vegetation products that are more mature and identified as essential climate variables like LAI and fAPAR, there were only few attempts to develop algorithm for CWC estimation, and no product is available yet, with very few available ground validation measurements. However, we were able to conduct both indirect comparisons between the products we developed at two spatial resolutions, and to conduct some direct validation as recommended by the CEOS LPV group. These studies also consider assessing temporal profiles of LAI and fAPAR as the spatio-temporal variations of these variables are quite well known, smooth and continuous (apart from natural hazards or strong human intervention), and available through different vegetation functioning or land surface models as a function of climate, soil and environmental conditions. These temporal variations are much less documented for CWC, which is susceptible to vary less smoothly, depending on the density and frequency of precipitation events or water stress, related to soil conditions. However, we compared the seasonal behaviour of our proiduct with other similar ones (e.G. VOD). More detailed is provoded in the paper submitted to remote sensing of environment and these behaviours are analysed in more details within WP10. While the performances of the proposed CWC product appear satisfying for most biome

types in terms of expected spatio-temporal behavior the direct validation exercise, although limited to the several available sites, show uncertainties that could be attributed to three main issues related to (i) the representativeness of the training database (ii) the accuracy and footprint of the ground measurements, and (iii) the saturation effect for dense canopies. For the representativeness, even the global efforts concerning leaf traits of diverse biome types were delivered into PROSAIL model in this study, the possible limitations and bias of TRY concerning diverse biome types, locations and temporal samplings, would also partially contribute to the uncertainties. For example, with not very high representation for crop species in TRY as those for other vegetation species *Kattge et al.* (2020), the uncertainties of CWC evaluations in croplands could also be partially explained. Moreover, since this study considered all the kinds of biome together by using PROSAIL model, in which the canopy is represented only by green components and a turbid medium. This may induce a high simplification of biome representation. In terms of ground sites, the diverse measurement strategies could clearly explain the different results shown in 7.5. The in situ data in Dahra, EucFACE and Qinghai, were obtained within the area ideally matching with one Landsat8/Sentinel-2 pixel (30m/20m). In this case, less uncertainties were taken into the CWC evaluations due to the relieved spatial mismatching between satellite pixel and ground site, comparing to SMAPVEX16 designed for SMAP mission with coarse grids (McNairn et al. 2017). Moreover, the environmental heterogeneity of ground data would also slightly contribute to the uncertainties. For the SMAPVEX16 site, although fully covered with croplands, diverse crop types including Oat, Wheat, Canola, Black bean, Soybean and Corn with different vegetation features (e.g., canopy structures) would increase the spatial heterogeneity in the same study area (even within the same satellite pixel). This may rise the difficulties of optical modeling for vegetation compared to that in homogenous area (e.g., the full grasslands for Dahra). Based on the discussions concerning the spatial mismatching and heterogeneity issue, it is fairly reasonable to understand the satisfying performances in Dahra, EucFACE and Qinghai, but more uncertainties for SMAPVEX16. Finally, the definition of CWC with regards to what the optical satellite is sensitive to should be also taken into account. In general, the spectral saturation issue in high-density vegetation traits would inherently limit the optical signals to completely model the whole canopy (Allen and Richardson 1968; Tucker 1977; Mutanga et al. 2023) and contribute to some uncertainties against the ground destructive measurements of the whole vegetation canopy. The denser the vegetation, the stronger the saturation issue as well as the more uncertainties would be introduced into the CWC assessments. This could partially explain why we get the better results in grasslands, then followed by croplands, and forests. In other words, the CWC retrievals based on PROSAIL model could satisfyingly represent the ground CWC of whole canopy in low and most medium densely vegetated areas (more refer to all the water of the canopy green components), but for high-dense vegetation traits, only exhibiting the leaf water of canopy considering green components mainly come from canopy leaf, also with the saturation issues due to the penetration capacity of optical signals. In this case, the integration with microwave observations should be promising.

Chapter 8

Optical Datacube (WP 8-9)

8.1 Introduction and purpose

The purpose of this section is to detail the setup of the Optical datacubes processing chains and their configurations. Biophysical products such as LAI, Fapar, Fcover have been downloaded from different sources and aggregated at the same spatial resolution to compare them to the VOD data sets derived from passive microwave measurements and to assess the complementary contribution of their content in the understanding of the VOD variability. The following sections represents the activities developed for the work package 8 and 9 of the project.

8.2 Input/Output data

Table 8.1 below summarizes the information about input optical datasets that were downloaded and used for the creation of the Optical Datacubes 1 and 2.

The output data are Optical Datacubes 1 and 2 with following specifications :

- **Datacube #1** : MODIS Climate Monitoring Grid, latitude/longitude projection, Global coverage [-180°W; +180°E] and [-90°S; +90°N], 0.05° spatial resolution (image size 3600H x 7200W)
- **Datacube #2** : EASE2 25km grid, latitude/longitude projection, coverage [-180°W; +180°E] and [-83.5°S; +83.5°N], 0.25° spatial resolution (image size 1388W x 584H)

Table 8.1: Input data for the Optical Datacubes #1 and #2

	Satellite	Product title	Provided by	Temporal range	Temporal resolution	Grid / projection, spatial resolution and coverage
1	AVHRR	LAI, FAPAR, FCOVER	POSTEl, Theia, CNES	1981-2020	10 days	Geographic lat/lon projection, 0.05°,[-180°W; +180°E], [-90°S; +90°N]
2	Proba V, Sentinel-3 OLCI	LAI, FAPAR, FCOVER (v1), NDVI (v2)	Copernicus Land Service	2014 - present	10 days	Plate carrée, 1/112° (1km), [-180°W; +180°E], [-60°S; +75°N]
3	VGT1, VGT2, Proba V	LAI, FAPAR, FCOVER (v2), NDVI (v3), Top Of Canopy (TOC) Reflectance (v1)	Copernicus Land Service	1999 - 2020	10 days	Plate carrée, 1/112° (1km), [-180°W; +180°E], [-60°S; +75°N]
4	MODIS Terra	MOD13C1 Vegetation Index (VI) (v6.1)	United States Geological Survey (USGS)	2000 - present	Multi-day	Geographic lat/lon projection, 0.05°, [-180°W; +180°E], [-90°S; +90°N]
5	MODIS Terra and Aqua (combined)	MCD43C4 BRDF (v6.1)	United States Geological Survey (USGS)	2000 - present	Daily product	Geographic lat/lon projection, 0.05°, [-180°W; +180°E], [-90°S; +90°N]
6	MODIS Terra and Aqua (combined)	MCD15A2H LAI, FAPAR (v6.1)	United States Geological Survey (USGS)	2002 - present	Multi-Day	Sinusoidal grid (tiles of 10°x10° at the equator), 500m, [-180°W; +180°E], [-90°S; +90°N]
7	MODIS Terra and Aqua (combined)	MCD18C2 PAR (v6.1)	United States Geological Survey (USGS)	2002-present	Daily product	Geographic lat/lon projection, 0.05°, [-180°W; +180°E], [-90°S; +90°N]

Based on the Datacubes specifications and the source data projections, the collections that had to undergone processings were as follows:

- projection for the Datacube 1 : Copernicus Land Service vegetation variables and the MCD15A2H vegetation variables
- projection for the Datacube 2 : all of the collections

8.3 Datacube #1 reprojection algorithm

Based on the Table 8.1, certain optical datasets were identified for the reprojection in order to match the projection and spatial resolution of the Datacube #1. This first geometrical reprojection was done mainly

with the `gdalwarp` tool. `gdalwarp` is an image mosaicing, reprojection and warping utility (Fig. 8.1). This program can reproject data to any supported projection. Multiple resampling methods are offered by the tool, and the resampling kernel size also depends on the chosen method.

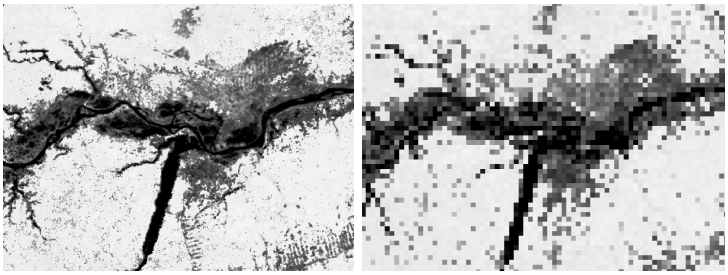


Figure 8.1: Example of the projection with `gdalwarp` from a more fine spatial resolution (left) to a more coarse spatial resolution (right)

Amongst multiple resampling methods offered by `gdalwarp`, some of these methods can induce information loss dû to the kernel size and the resampling method itself (Fig. 8.2). Lost information is the information coming from pixels that do not fall into the resampling kernel.

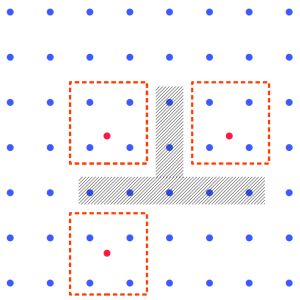
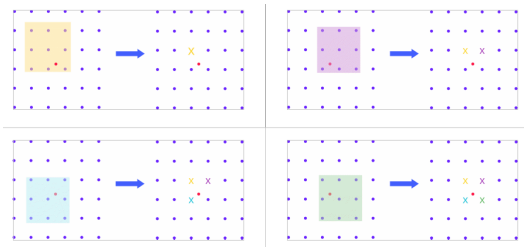


Figure 8.2: `gdalwarp` 2x2 bilinear kernel with the hatched zones showing the pixels where the information has been lost during the resampling

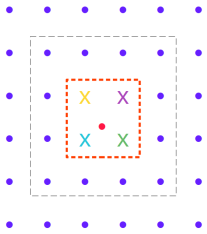
This problem can be solved either by:

1. changing the resampling method and taking a bigger kernel which would take into account a larger number of pixels at once (cubic resampling with 3x3 kernel, for example)
   or
2. adding an averaging convolution step (as a low-pass filter) before the projection.

The second method is visualised on the Fig. 8.3. It shows that adding the convolution step with a 3x3 mean kernel allows to use the bilinear resampling algorithm without losing any pixel that falls in-between the resampling kernels.



(a) 1st step : convolution of the neighbourhood for the final pixels in the resampling kernel



(b) 2nd step : projection of the data. Red square is the `gdalwarp` kernel

Figure 8.3: Bilinear `gdalwarp` convolution with the preliminary mean-convolution step

The example above is a demonstration of a particular example but the explanation of the convolution step can be generalized. When the difference between input and output spatial resolutions is too big, the change of the resampling method with a bigger kernel won't necessarily help, and aliasing artefacts may occur int final result. The robust solution in this case is to add a convolution step before projection. This averaging convolution step has to summarize data of the neighbourhood of each pixel in a mean value, and prevent an aliasing in the final image, as it will play the role of the low-pass filter. The convolution has to bring the input spatial resolution closer to the final one, but not over-filter the data. This anti-aliasing concept is the same that the one used in the processing of 1D signals. On top of the possible information loss, the aliasing can also be the reason of the reshaped or modified forms or curves which can falsify the original data (Fig. 8.4).

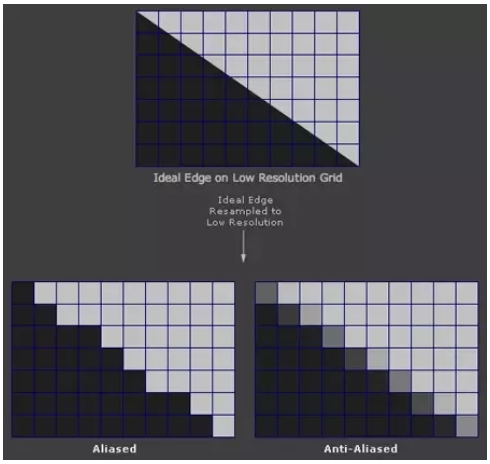


Figure 8.4: Example of aliased and anti-aliased results on a test image

Considering the resolutions of the input data for the Datacube #1, the reprojection of the data contained two main steps:

1. convolution of the data to the "intermediate" spatial resolution and,
2. the reprojection itself

These steps were the same for every reprojected collection with some slight differences in the configuration of the convolution kernel size which depends on the gap between source and output spatial resolutions.

8.3.1 Convolution to the intermediate resolution

The averaging kernel is defined as follows (Eq. 8.1), where dimensions N and M are defined based on the input/output spatial resolutions:

$$N\underbrace{\begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}}_{M} \cdot \frac{1}{NM} \tag{8.1}$$

After that, the convolution method is a usual kernel convolution algorithm, but with coefficients that are allows to compute the mean value of the neighbourhood (Fig. 8.5):

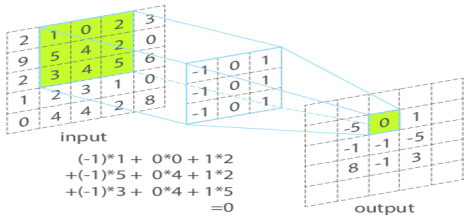


Figure 8.5: 2D image convolution concept

Two kernel sizes were defined for the processing:

1. 5x5 kernel for the Copernicus 1km products
2. 7x7 kernel for the MODIS MCD15A2H 500m tiles

For the majority of the data, main variables are "physical" quantitative data (reflectances, vegetation indices), and the convolution can be applied as it is. However, for some other additional variables, like quality flags, the data is not something physical or continuous; it won't make much sense to compute a mean value or do a projection on such qualitative information. Thus, for this type of data, a special approach was developed in this project and is discussed below.

8.3.1.1 Special case : convolution of the non-physical qualitative variables

For the quality/classification flags, the convolution is not meant to compute the mean value of the neighbourhood, but the percentage of a chosen flag in the region. Fig. 8.6 shows an example of this approach. For a given flag description or mask value, the kernel compute the contribution percentage of this flag in the final projected pixel. Then, this percentage can be projected with `gdalwarp` the same way as other variables. The size of the kernel is the same as the one that was used for the physical variables of the same product collection.

0	WATER
1	LAND

0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

(a) Original image with flag value LAND/WATER per pixel

0	5	10	20	25	30	40	45	50	60	65	70	80	85	90	100
% of land pixels in the neighbourhood															

0	0	0	0	0	0	0	0	0	0	0	0	0	22	44	100
0	0	0	0	0	0	0	0	0	0	0	0	11	44	78	100
0	0	0	0	0	0	0	0	0	0	11	22	44	67	89	100
0	0	0	0	0	0	0	0	0	11	33	56	78	89	100	100
0	0	0	0	0	0	0	11	22	44	67	89	100	100	100	100
0	0	0	0	0	11	22	44	56	78	89	100	100	100	100	100
0	0	0	0	11	33	56	78	89	100	100	100	100	100	100	100
0	0	0	0	11	44	78	100	100	100	100	100	100	100	100	100
0	0	0	0	11	33	67	89	100	100	100	100	100	100	100	100
0	0	0	0	0	22	56	89	100	100	100	100	100	100	100	100
0	0	0	0	11	33	67	89	100	100	100	100	100	100	100	100
0	0	0	0	22	56	89	100	100	100	100	100	100	100	100	100
0	0	0	0	22	56	89	100	100	100	100	100	100	100	100	100
0	0	0	0	11	33	56	100	100	100	100	100	100	100	100	100

(b) Convolved image with LAND pixels percentage as the result

Figure 8.6: Convolution to the intermediate resolution in a case of a qualitative variable

8.3.2 Reprojection with `gdalwarp`

As it was already mentioned above, the projection step was done via the `gdalwarp` utility. This tool is also helpful for merging tiles and image parts together or for reprojection between different spatial references. In the scope of this project the functionality of merging tiles was used for the MODIS MCD15A2H collection, where original products are computed on the sinusoidal grid (Fig. 8.7).

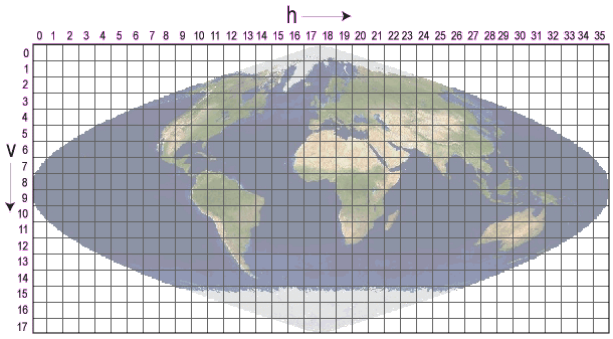


Figure 8.7: MODIS sinusoidal tile grid

`gdalwarp` is a command line tool, and it accepts multiple input options that allow to tune the projection to one's needs. Below is the verbatim of the `gdalwarp` help function.

```
Usage: gdalwarp [--help-general] [--formats]
    [-s_srs srs_def] [-t_srs srs_def] [-to "NAME=VALUE"]* [-novshiftgrid]
    [-order n | -tps | -rpc | -geoloc] [-et err_threshold]
    [-refine_gcps tolerance [minimum_gcps]]
    [-te xmin ymin xmax ymax] [-tr xres yres] [-tap] [-ts width height]
    [-ovr level|AUTO|AUTO-n|NONE] [-wo "NAME=VALUE"] [-ot Byte/Int16/...]
    [-wt Byte/Int16] [-srcnodata "value [value...]"] [-dstnodata "value [value...]"]
    -dstalpha [-r resampling_method] [-wm memory_in_mb] [-multi] [-q]
    [-cutline datasource] [-cl layer] [-cwhere expression]
    [-csql statement] [-cblend dist_in_pixels] [-crop_to_cutline]
    [-of format] [-co "NAME=VALUE"]* [-overwrite]
    [-nomd] [-cvmd meta_conflict_value] [-setci] [-oo NAME=VALUE]*
    [-doo NAME=VALUE]*
    srcfile* dstfile
```

```
Available resampling methods:
    near (default), bilinear, cubic, cubicspline, lanczos, average, mode,
    max, min, med, Q1, Q3.
```

In our case the input parameters was defined as follows:

- `-t_srs` target spatial reference, was extracted from MODIS products which were already in a Datacube #1 spatial reference and defined by

```
GEOGCS["Unknown datum based upon the Clarke 1866 ellipsoid",
   DATUM["Not specified (based on Clarke 1866 spheroid)",
      SPHEROID["Clarke 1866",
         6378206.4,
         294.9786982138982,
         AUTHORITY["EPSG","7008"]]],
   RIMEM["Greenwich",0],
   UNIT["degree",0.0174532925199433]]
```

- `-te`, georeferenced extents of output file, was defined by "-180 -90 180 90" as the products are of global coverage
- `-tr`, output spatial resolution, was defined by "0.05 -0.05"
- `-r`, resampling method, was defined by "bilinear"
- `-dstnodata` and `-srcnodata`, no-data masking values respectively in the output and input data, was defined by "-999" for both because the convolution step provides the no-data pixels with -999 value and for the harmonization of the data this value was also kept for the Datacubes as the no-data value
- option `-co` defined with `WRITE_BOTTOMUP=NO`, this is a NetCDF driver option which allows to set the y-axis order for export. In order to export images with the "north-up" convention, this option has to be set to NO.

The input data for the `gdalwarp` utility is the convolved data, and the output data is the final projected data that is concatenated afterwards along `time` dimension into the Datacube #1. Source data that was already in the Datacube #1 projection and on the same grid was directly concatenated into the Datacube #1 format without `gdalwarp` or low-pass filter processing.

8.4 Datacube #1 validation

The first Optical Datacube contains the main data of the product, projected onto a grid of 0.05° spatial resolution (image size 3600H x 7200W), as well as some metadata and additional variables which give additional information on the products used for the Datacube aggregation. Global and variables' attributes are validated by manual checking. Variables' values are validated using global map visualisation and comparisons of time series "first Optical Data Cube VS source data" for given coordinates on the globe. In this section, we will describe the validation methodologies for two different scenarios :

1. same spatial resolution between source data and Datacube #1
2. source data is on the different grid and with different spatial resolution than Datacube #1

8.4.1 Products with grid shared by source data and Datacube #1

8.4.1.1 Methodology

For the case where no reprojection is needed, variables are validated using:

- comparison of global map from first optical datacube and source data, generally for one moment in January and one in August. Absence of artefacts and spatial coherence between source data and first optical data cube are checked.
- comparison of time series data (over a year) from the first optical datacube and from the source data for a specific position. In this case, when the grid is shared by source data and first optical datacube, a pixel in the source data corresponds to the same pixel in the first optical datacube. Therefore, for a given period and position, the variability of variables in this datacube must be the same as the one in the corresponding optical database.

For this second step, time series over three hundred LANDVAL sites were systematically created for all of the collections, in order to assess the Datacube #1 results. The distribution of these sites is showed on Fig. 8.8. These reference sites are well known for their high temporal stability and are used by CNES for the absolute calibration of remote sensing sensors, thus allowing us to have a strong basis of locations where we can test the results of our Datacube processing. For this validation five sites were selected amongst the three hundred in order to represent a diverse range of cases in terms of time series patterns and variable range values allowing a more quick and efficient study :

- Aguascalientes (21.7°N/-102.32°E)
- Aire Adour (43.7°N/0.25°E)
- Arm_Cart_SGP (36.64°N/-97.5°E)
- BAC_LIEU (9.28°N/105.73°E)
- Belmanip_00075 (41.58°N/-77.85°E)

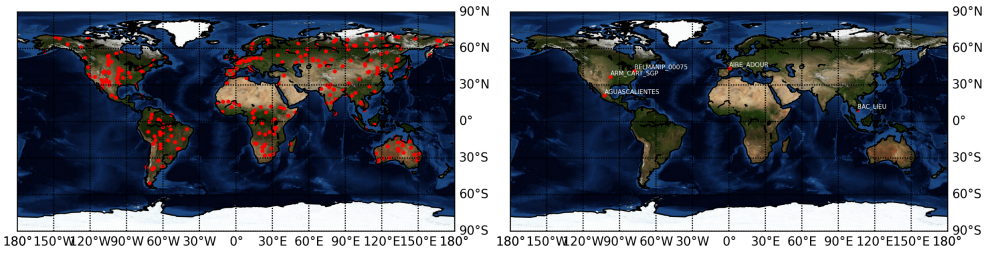


Figure 8.8: Map of the 308 LANDVAL ground validation sites (left) and the five selected station for the validation (right)

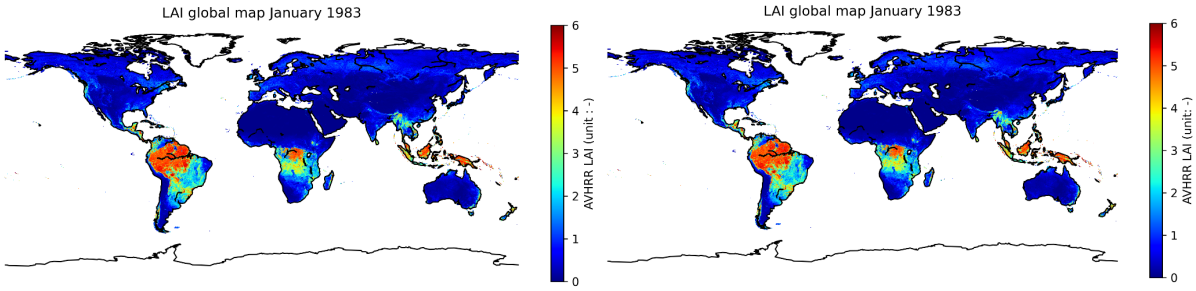
8.4.1.2 Results

- AVHRR LAI product

On Figure 8.9 no artefacts are observed, and there is a good match between the two global maps.

On the Figure 8.10, the variability of the LAI over the year 1983 is exactly the same for the first optical data cube and the source data as expected. Same validation methodology was applied to the FCOVER and FAPAR products of the AVHRR collection. The global maps comparison and the analysis of the time series provided the same conclusion as the one demonstrated for LAI product : good match is observed between maps and

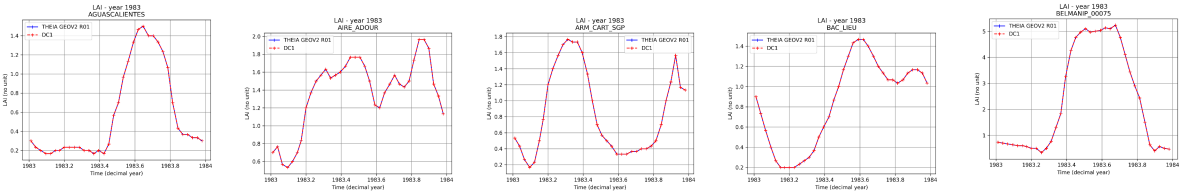
the variability of the data is exactly the same between Datacube #1 and the source data for the mentioned variables of the AVHRR collection. The comparison was also declined for another date in the summer season to observe seasonal changes, and this comparison also demonstrated a good match between the source and the Datacube # 1 data. The projection of the data was validated.



(a) Datacube #1 AVHRR LAI map

(b) Original AVHRR LAI map

Figure 8.9: Global visual comparison between AVHRR LAI original and harmonized data on January 5, 1983



(a) AGUASCALIENTES site (21.7°, -102.32°)

(b) AIRE ADOUR site (43.7°, 0.25°)

(c) ARM CART SGP site (36.64°, -97.5°)

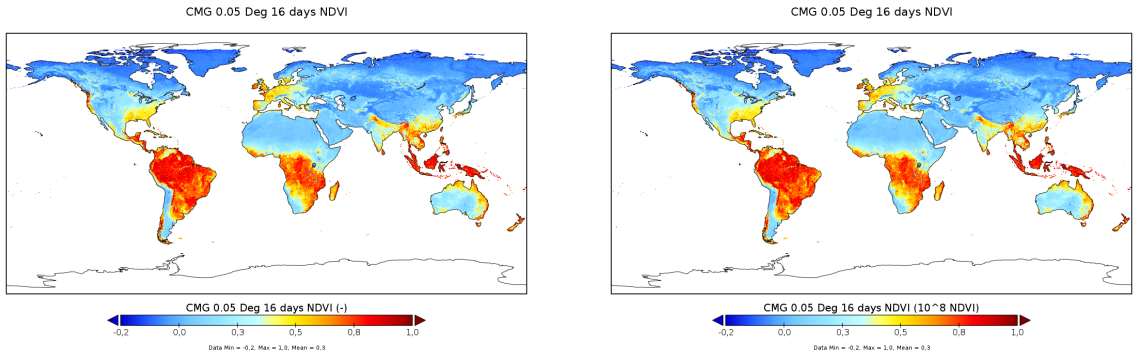
(d) BAC LIEU site (9.28°, 105.73°)

(e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.10: LAI time series (year 1983) comparison between first optical data cube (in red) and source data (in blue) for the five selected stations.

- MOD13C1 NDVI product

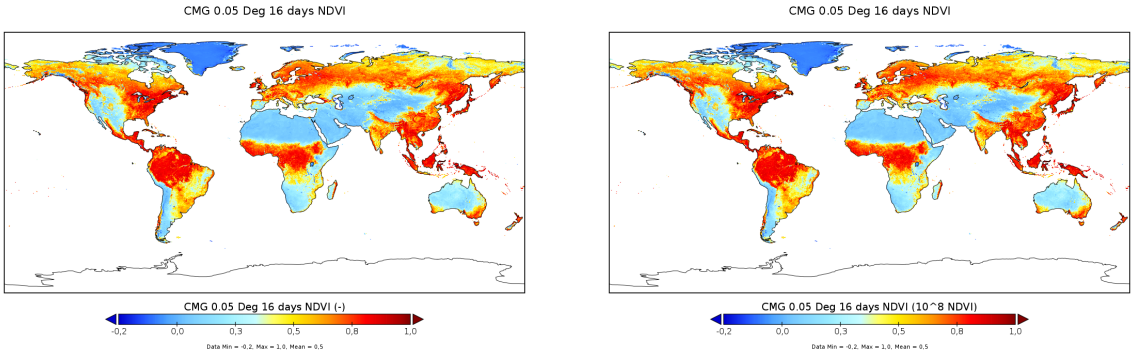
For the MOD13C1 NDVI product, good matches were observed on the global maps for the winter/summer seasons, and good match were also observed for the time series of the five selected stations. The projection of the data was validated.



(a) Datacube #1 MOD13C1 NDVI map

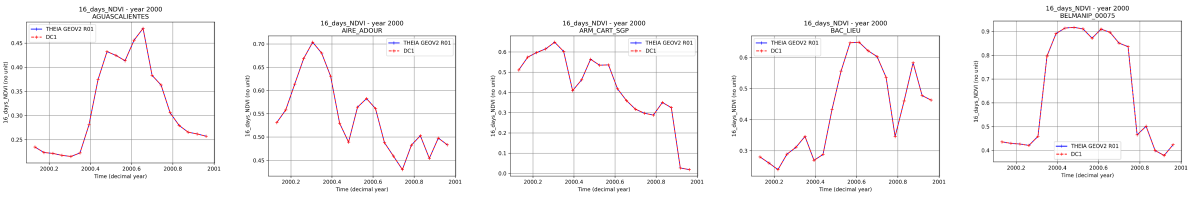
(b) Original MOD13C1 NDVI map

Figure 8.11: Global visual comparison between MOD13C1 NDVI original and harmonized data on a February 18, 2000



(a) Datacube #1 MOD13C1 NDVI map (b) Original MOD13C1 NDVI map

Figure 8.12: Global visual comparison between MOD13C1 NDVI original and harmonized data on a September 5, 2000

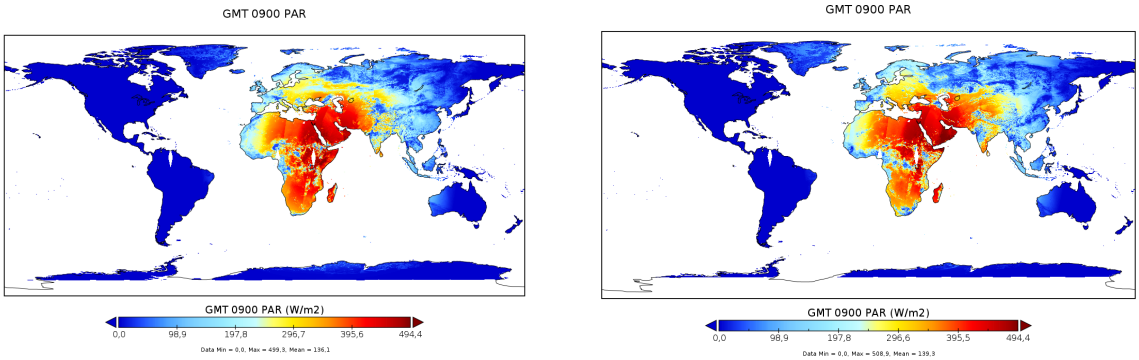


(a) AGUASCALIENTES site (21.7°, -102.32°) (b) AIRE ADOUR site (43.7°, 0.25°) (c) ARM CART SGP site (36.64°, -97.5°) (d) BAC LIEU site (9.28°, 105.73°) (e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.13: MOD13C1 NDVI time series (year 2000) comparison between first optical data cube (in red) and source data (in blue) for the five selected stations

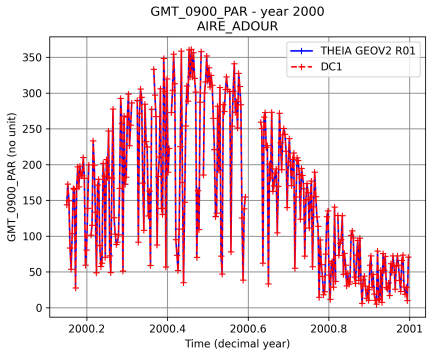
- MCD18C2 Photosynthetically Active Radiation (PAR) product

Good matches were observed on the global maps for the winter/summer seasons, and good match were also observed for the time series of the five selected stations. For the illustration purposes, only plots of one of variables are presented below. The projection of the data was validated.

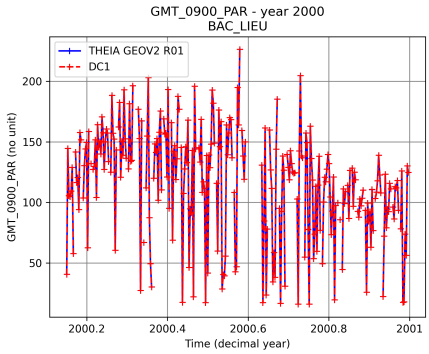


(a) Datacube #1 MCD18C2 GMT 09:00 PAR map (b) Original MCD18C2 GMT 09:00 PAR map

Figure 8.14: Global visual comparison between MCD18C2 GMT 09:00 PAR original and harmonized data on an August 25, 2002



(a) AIRE ADOUR site (43.7°, 0.25°)

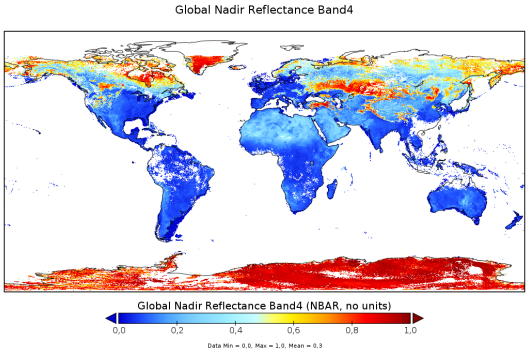


(b) BAC LIEU site (9.28°, 105.73°)

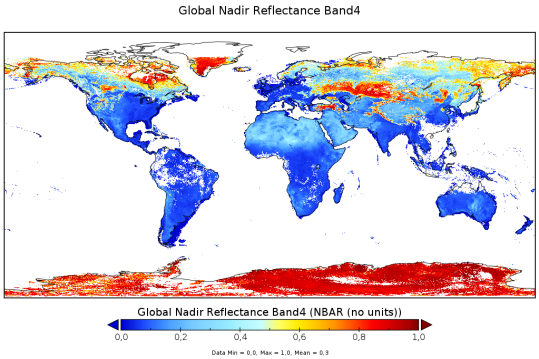
Figure 8.15: MCD18C2 PAR time series (year 2000) comparison between first optical data cube (in red) and source data (in blue) for the selected stations

- MCD43C4 Nadir Reflectance product

Good matches were observed on the global maps for the winter/summer seasons, and good match were also observed for the time series of the five selected stations. For the illustration purposes, only plots of one of variables are presented below. The projection of the data was validated.

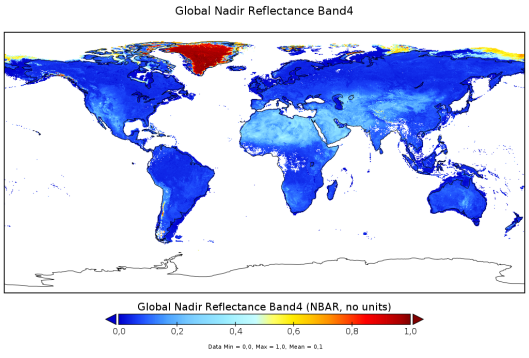


(a) Datacube #1 MCD43C4 Global Nadir Reflectance (Band 4) data

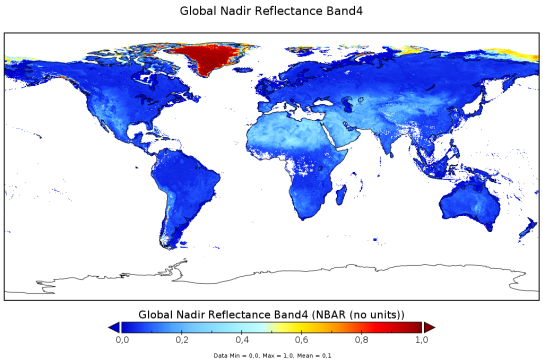


(b) Original Global Nadir Reflectance (Band 4) data

Figure 8.16: Global visual comparison between MCD43C4 Global Nadir Reflectance (Band 4) original and harmonized data on an February 25, 2000

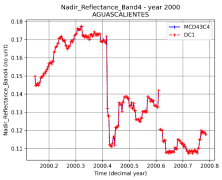


(a) Datacube #1 MCD43C4 Global Nadir Reflectance (Band 4) data

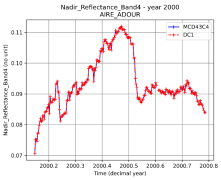


(b) Original Global Nadir Reflectance (Band 4) data

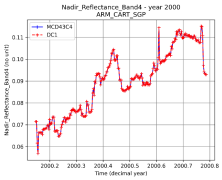
Figure 8.17: Global visual comparison between MCD43C4 Global Nadir Reflectance (Band 4) original and harmonized data on an July 29, 2000



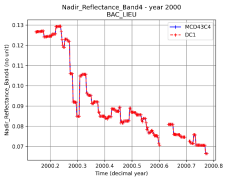
(a) AGUASCALIENTES site (21.7°, -102.32°)



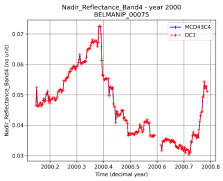
(b) AIRE ADOUR site (43.7°, 0.25°)



(c) ARM CART SGP site (36.64°, -97.5°)



(d) BAC LIEU site (9.28°, 105.73°)



(e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.18: MCD43C4 Global Nadir Reflectance (Band 4) time series (year 2000) comparison between first optical data cube (in red) and source data (in blue) for the five selected stations

8.4.2 Products with different grids and projections

8.4.2.1 Methodology

In this specific case, quantitative variables (such as LAI, FCOVER...) and qualitative variables (such as quality flag) require validation using different methodologies:

- *quantitative variables* are validated using the same methodology as described in the previous section (global map and time series comparisons)
- *qualitative variables* are manually validated through pixel-to-pixel comparisons

As described in the section 8.4.1.1, a comparison is performed between the global map generated from the first optical datacube and the source data (when global map is available). Nothing changes for this method.

However, in order to perform a time series analysis and comparison with data with different spatial resolutions, following methods were applied:

1. find the nearest neighbour pixel to the ground validation site in the original source data (via minimal geodesic distance), this pixel is represented by the red square on Fig. 8.19
2. identify the Datacube #1 pixel that includes the source station pixel and retrieve its center; this pixel is indicated by the yellow circle on Fig. 8.19

3. retrieve from the original source data the 5x5 km² neighborhood around the DC1 found pixel. The size of this neighbourhood in pixels will depend on the original data spatial resolution. In the example of Fig. 8.19 the 5x5 neighborhood is 25 Copernicus pixels because the original Copernicus spatial resolution is 1km.

4. calculate the mean value and standard deviation of pixel values, as well as the number of pixels with available data *within the selected area around the DC1 central pixel*

5. compare time series from the DC1 center pixel, mean +/- std distribution and the original source data pixel

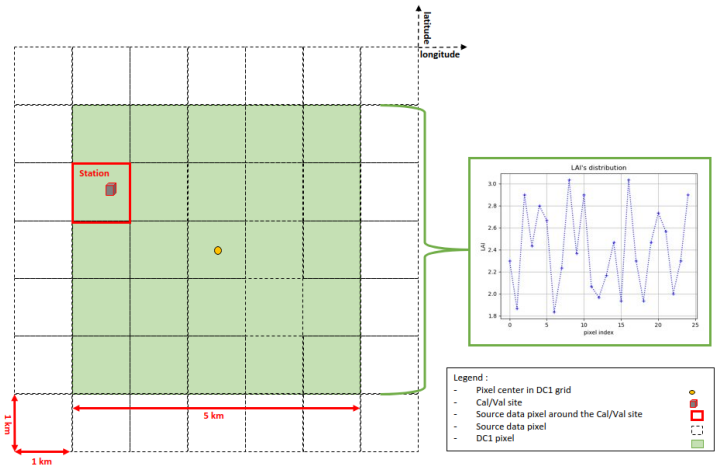


Figure 8.19: Diagram of time series comparison methodology (example for a Copernicus product with 1x1 km² spatial resolution)

An example of such method visualization (applied to the Copernicus LAI product) is presented on Figure 8.20. On the left plot, the legend is as follows:

red dashed line : LAI value from the Datacube #1 pixel center

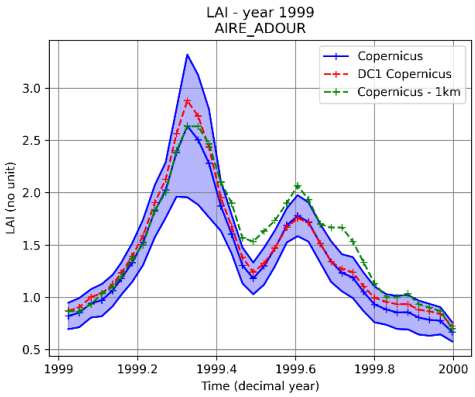
green line : LAI value from the source Copernicus data (nearest neighbour pixel to the station)

blue line : the average (mean) value of the 25 source pixels surrounding the DC1 center pixel

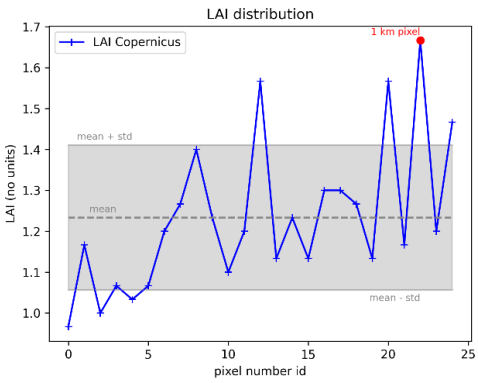
shaded blue area : mean +/- standard deviation of the 25 source pixels surrounding the DC1 center pixel

To validate the first Copernicus optical data cube we expect to have the red and green line within the blue area.

If it is not the case, the distribution of a variable is examined more closely to determine if it can be explained by the source data distribution in the 5x5 km² neighbourhood (for example, at the decimal year 1999.7 on Fig. 8.20). This case is illustrated on the right plot of the Figure 8.20. The blue curve represents LAI pixel values distribution from the 25 1km pixels that fall into the neighbourhood around the DC1 center pixel (green zone on Fig. 8.19). The central grey dashed line corresponds to the distribution mean and the grey shaded area represents the mean +/- std of the LAI distribution. The red point corresponds to the source data pixel containing the LANDVAL site (red square on Fig. 8.19). This pique explains the less strong match between the Datacube #1 pixel values and the station 1km pixel values around the dates of decimal year 1999.7 (left plot on Fig .8.20).



(a) Copernicus LAI time series (year 1999) comparison between first optical data cube and source data at the position of Aguascalientes station



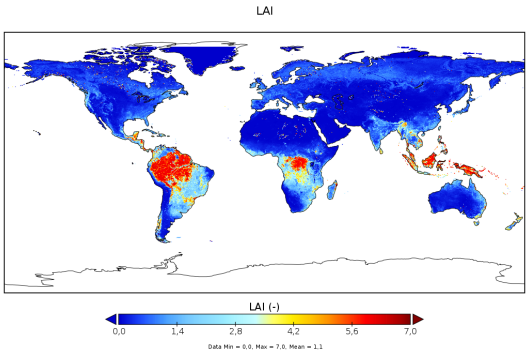
(b) LAI distribution for 1km pixel resolution at 1999.7 decimal year

Figure 8.20: Example of time series analysis for the Copernicus 1km LAI dataset

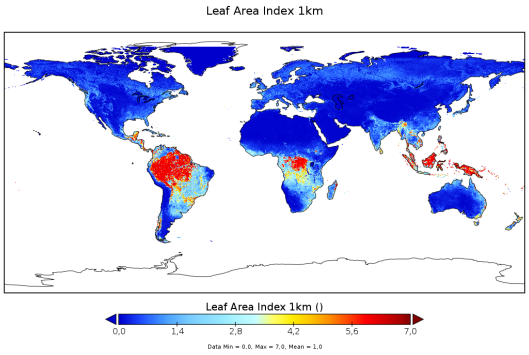
In the following subsection, we present detailed comparisons between the first optical datacube and source data for each of the collections that have undergone the reprojection. Firstly, we compared global maps, followed by a comparison of time series. The results showed a strong agreement between the first optical data cube and source data.

8.4.2.2 Results

- LAI Copernicus Global Land Service product

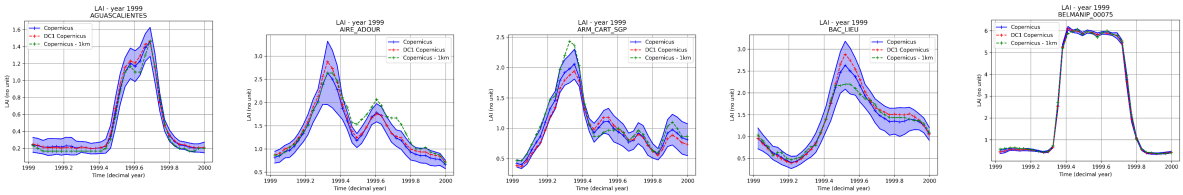


(a) Datacube #1 LAI Copernicus data



(b) Original LAI Copernicus data

Figure 8.21: Global visual comparison between Copernicus LAI original and harmonized data on January 10, 1999



(a) AGUASCALIENTES site (21.7°, -102.32°)

(b) AIRE ADOUR site (43.7°, 0.25°)

(c) ARM CART SGP site (36.64°, -97.5°)

(d) BAC LIEU site (9.28°, 105.73°)

(e) BELMANIP 75 site (41.588198°, -77.852451°)

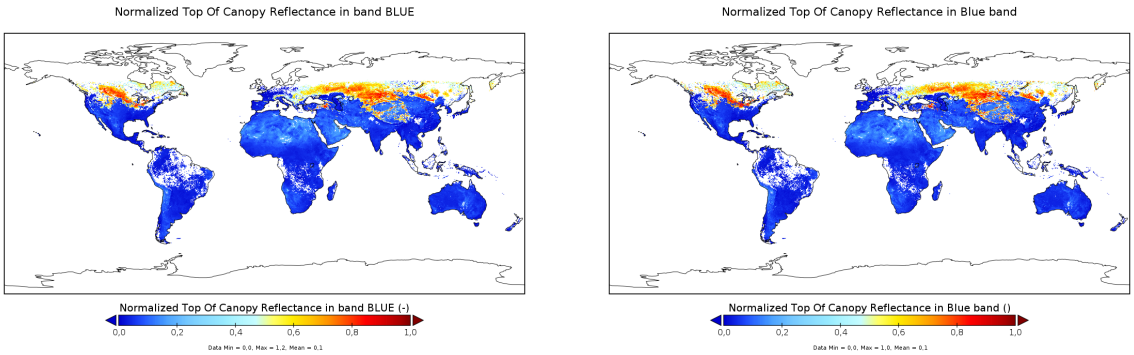
Figure 8.22: Copernicus LAI time series (year 1999) comparison between first optical data cube (in red) and source data (in blue) for the five selected stations

We can observe that on all five plots of Figure 8.22 the red curve is well within the shaded blue area. Upon examining the values of the 1 km resolution pixel, it is evident that for the "ARM_CART_SGP" and "BELMANIP_00075", there are data points that fall outside the blue zone. As discussed in section 8.4.2.1, this can be verified and analysed by plotting the distribution of the 25 Copernicus pixels included within the first optical data cube's pixel. As it was seen on Figure 8.20, the value of the 1 km Copernicus pixel (red point) can fall outside the mean of the distribution, causing the occasional mismatch between the 1km pixel value and the Datacube #1 pixel value for a given station.

Same validation was performed on the FCOVER and FAPAR products of the Copernicus collection. Similar results were observed whether studying the global maps distribution or analyzing time series plots. The projection was validated.

- Top-of-Canopy Reflectance (TOCR) Copernicus Global Land Service product

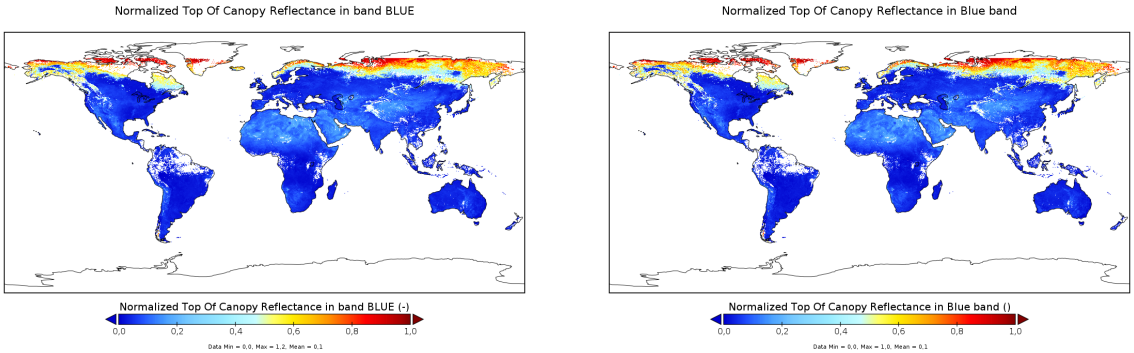
Good match and no artefacts were observed on the global maps. During the time series analysis, the Datacube # 1 values are well within the mean ± std value range. The Copernicus 1 km pixel value sometimes falls outside of that range due to the data reprojection and the reasons explained in the section 8.4.2.1. The projection was validated.



(a) Datacube #1 TOCR Copernicus data

(b) Original TOCR data

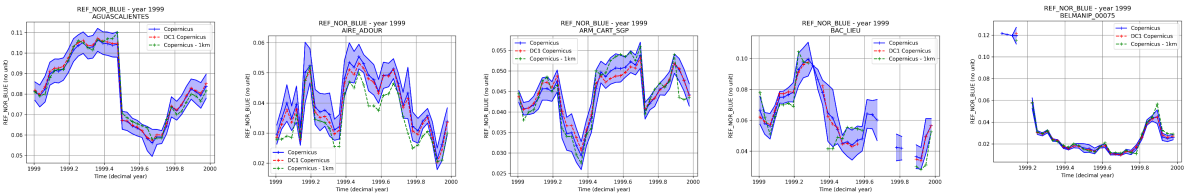
Figure 8.23: Global visual comparison between Copernicus Global TOCR original and harmonized data on a January 3, 1999



(a) Datacube #1 TOCR Copernicus data

(b) Original TOCR data

Figure 8.24: Global visual comparison between Copernicus Global TOCR original and harmonized data on a May 3, 1999



(a) AGUASCALIENTES site (21.7°, -102.32°)

(b) AIRE ADOUR site (43.7°, 0.25°)

(c) ARM CART SGP site (36.64°, -97.5°)

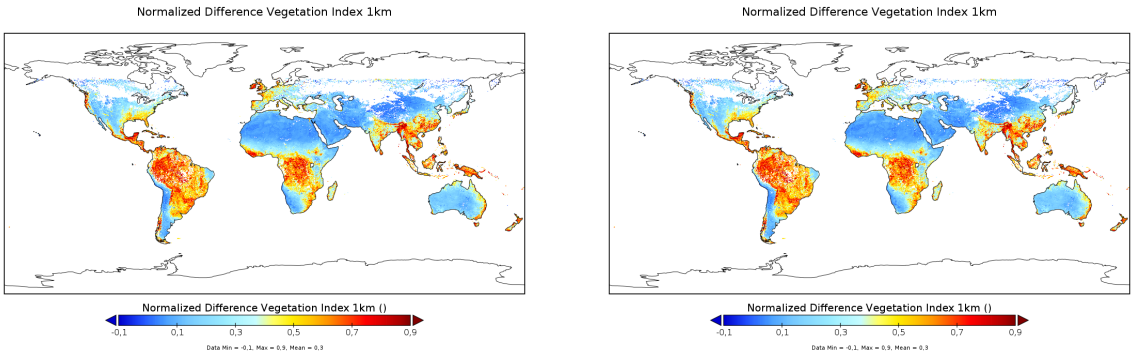
(d) BAC LIEU site (9.28°, 105.73°)

(e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.25: Copernicus TOCR time series (year 1999) comparison between first optical data cube (in red) and source data (in blue) for the five selected stations

- NDVI Copernicus Global Land Service product

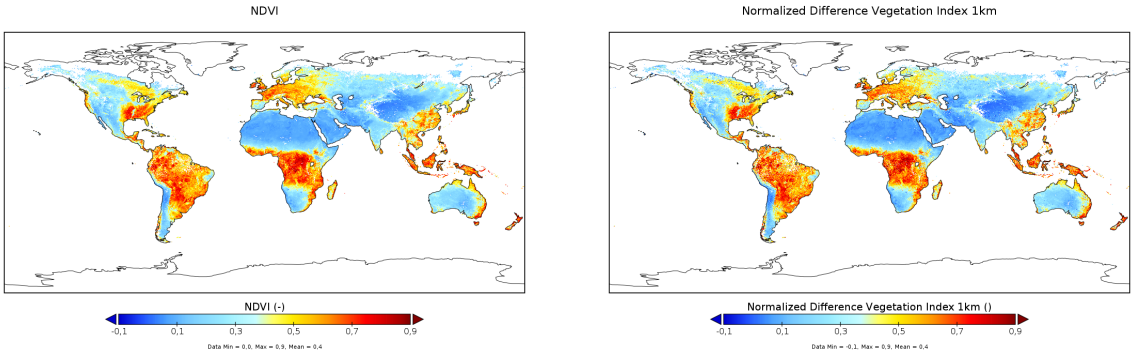
Good match and no artefacts were observed on the global maps. During the time series analysis, the Datacube # 1 values are well within the mean ± std value range. The Copernicus 1 km pixel value falls sometimes outside of that range due to the data reprojection and the reasons explained in the section 8.4.2.1. The projection was validated.



(a) Datacube #1 NDVI Copernicus data

(b) Original NDVI data

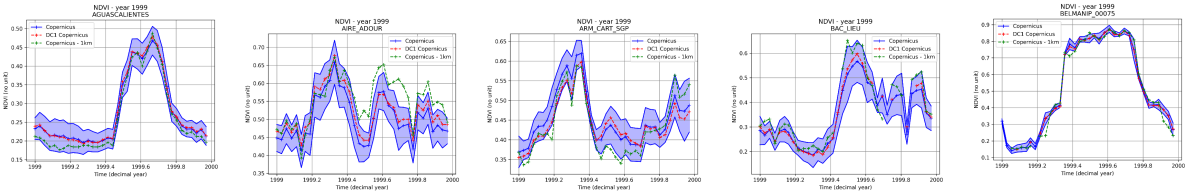
Figure 8.26: Global visual comparison between Copernicus Global NDVI original and harmonized data on a January 1, 1999



(a) Datacube #1 NDVI Copernicus data

(b) Original NDVI data

Figure 8.27: Global visual comparison between Copernicus Global NDVI original and harmonized data on a May 1, 1999



(a) AGUASCALIENTES site (21.7°, -102.32°)

(b) AIRE ADOUR site (43.7°, 0.25°)

(c) ARM CART SGP site (36.64°, -97.5°)

(d) BAC LIEU site (9.28°, 105.73°)

(e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.28: Copernicus NDVI time series (year 1999) comparison between first optical data cube (in red) and source data (in blue) for the five selected stations

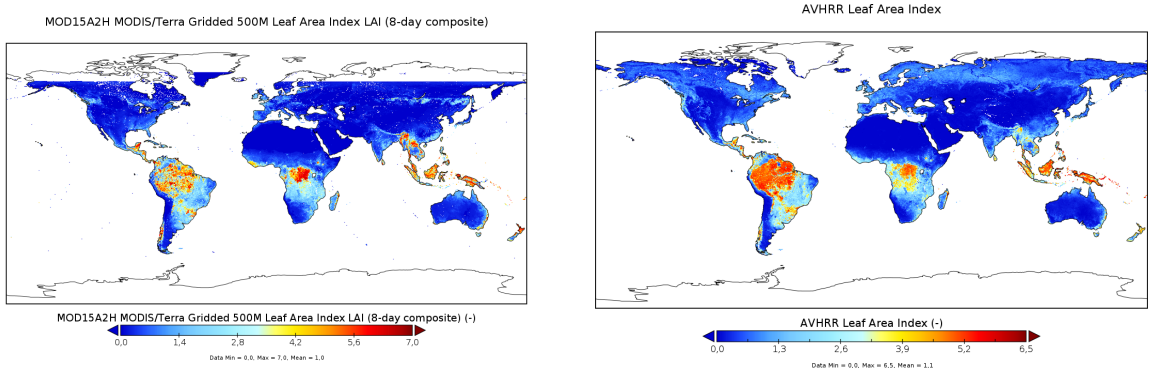
- MCD15A2H LAI product

The MCD15A2H FPAR (Combined Fraction of Photosynthetically Active Radiation) and LAI (Leaf Area Index) product is an 8-day composite dataset with 500 meter pixel size. Due to the fine spatial resolution, the MCD15A2H products use the Sinusoidal grid tilling system (Fig. 8.7). In consequence, after merging the tiles together into a global map, we used the methodology presented in section 8.4.2 for this product. We will present global map and time series comparison. For each time series comparison the legend is the same as explained in the section 8.4.2.1.

The original MCD15A2H data is provided on the very fine Sinusoidal grid tilling system, hence only the analysis of the global map form the Datacube #1 was possible (Fig. 8.29). The use of the others collections' LAI products can help to asses the coherence of the observed tendencies. Based on the comparison on Fig. 8.29, we can see no artefacts and incoherent data for the MCD15A2H reprojected product. Zones with higher values are less homogeneous on the MODIS map compared to the same zones on the AVHRR map but it is due to the more fine resolution of the source MODIS data; the more fine resolution brings us more details while the AVHRR data gives a more averaging result with its more coarse resolution. Nevertheless, the tendencies between the two maps are coherent and follow the same pattern.

We can observe on Figure 8.30 that in all five figures, except for Aguascalientes station, the red curve is well within the shaded blue area. Upon examining the values of the 500 m resolution pixel, it is evident that there are data points that fall outside the blue zone. It can be explain by the 500m pixel distribution.

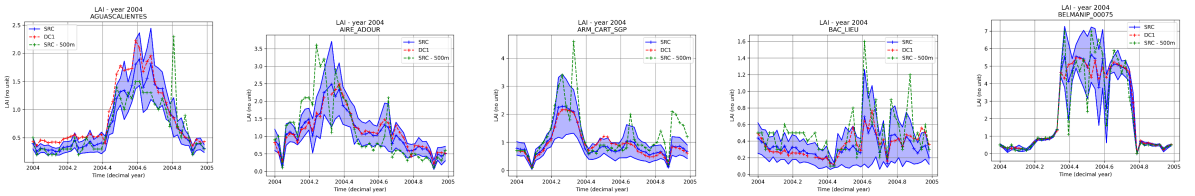
Based on the global map observations and time series analysis, no artefacts and incoherent data were detected for this product. The projection was validated.



(a) MCD15A2H LAI global map from the first optical data cube on January 1, 2020

(b) AVHRR LAI global map from the first optical data cube on January 5, 2020

Figure 8.29: Comparison of the global LAI maps from the MCD15A2H and AVHRR Datacubes #1



(a) AGUASCALIENTES site (21.7°, -102.32°)

(b) AIRE ADOUR site (43.7°, 0.25°)

(c) ARM CART SGP site (36.64°, -97.5°)

(d) BAC LIEU site (9.28°, 105.73°)

(e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.30: MCD15A2H LAI time series (year 1999) comparison between first optical data cube (in red) and source data (in blue) for the five selected stations

8.4.2.3 Conclusion

In this section, we have presented detailed comparisons between the first optical data cube and source data for each data cube. Firstly, we compared global maps, followed by a comparison of time series. As expected to validate the first optical data cube, the results showed a strong agreement between the first optical data cube and source data.

8.5 Datacube #2 reprojection algorithm

In order to use the project optical data presented above to improve our understanding of the vegetation opacity data sets retrieved from SMOS measurements and other low frequency passive microwave measurements, it is necessary to produce this optical data at 0.25° spatial resolution (SMOS spatial resolution). To do this, a Second Optical Datacube at 0.25° spatial resolution were created from the Datacube #1 presented above.

For this second datacube, the input is the first datacube itself, where all of the data has already been harmonized and projected onto the same intermediate grid. In order to obtain the Datacube #2, the first datacube data has to undergone a new re-projection where the objective is to smooth the data to the 0.25° spatial resolution via a kernel of a weighted average. This kernel is meant to represent the behaviour of the SMOS antenna. The weights of the kernel are calculated following a predefined formula elaborated by the CESBIO colleagues of this project. The section aims to detail different steps of this reprojection algorithm.

8.5.1 Kernel computation

The mathematical approximation of the SMOS antenna coefficients that multiply received signal follows the formula 8.2 :

$$w = \frac{\left(sinc(\frac{0.0733 \cdot d}{\pi})\right)^{1.4936}}{1 + d^{2.1030} \cdot 2.5748118104204e - 04} \tag{8.2}$$

Where d is the distance between the central point right underneath the antenna and any other point in the neighbourhood.

For a point with coordinates (x_0, y_0) in 2D dimension, we define the distance to another point (x, y) by the formula 8.3:

$$d = \sqrt{(x_0 - x)^2 + (y_0 - y)^2} \tag{8.3}$$

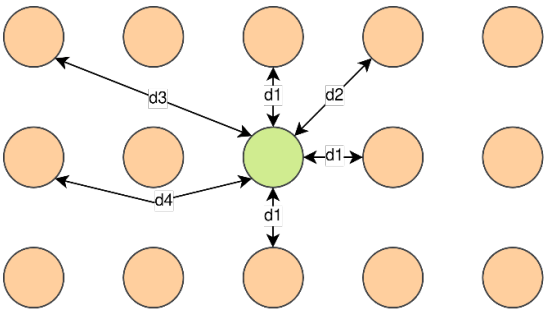
For a latitude/longitude point, we can define the distance respecting the curvature of the globe by the formula

$$d = R \cdot 2 \cdot arctan\left(\frac{\sqrt{sin^2(d_{lat}/2) + cos(lat_0) \cdot cos(lat_n) \cdot sin^2(d_{lon}/2)}}{\sqrt{1 - (sin^2(d_{lat}/2) + cos(lat_0) \cdot cos(lat_n) \cdot sin^2(d_{lon}/2))}}\right) \tag{8.4}$$

where :

- d_{lat} and d_{lon} are respectively the latitude and longitude distance in radians between two points
- lat_0 and lat_n are respectively latitude values of the reference point and the second point

The kernel is applied to the Datacube #1 data on the Datacube #1 grid. For any point of the grid we can compute a matrix with distances to this point's neighbours (Fig. 8.31).



(a) Distances from the central point to the neighbours on a regular grid

d3	d2	d1	d2	d3
d4	d1	0	d1	d4
d3	d2	d1	d2	d3

(b) Matrix with the calculated distance values

Figure 8.31: Notion of distance matrix in the kernel computation methodology

Then, we can apply the formula 8.2 from above to compute kernel weights for each of the neighbour points based on computed distances. An example of such distance matrix and associated kernel is presented on Figure 8.32. In this example the neighbourhood is 140 km² (x and y varying from -70 to 70km from the central point).

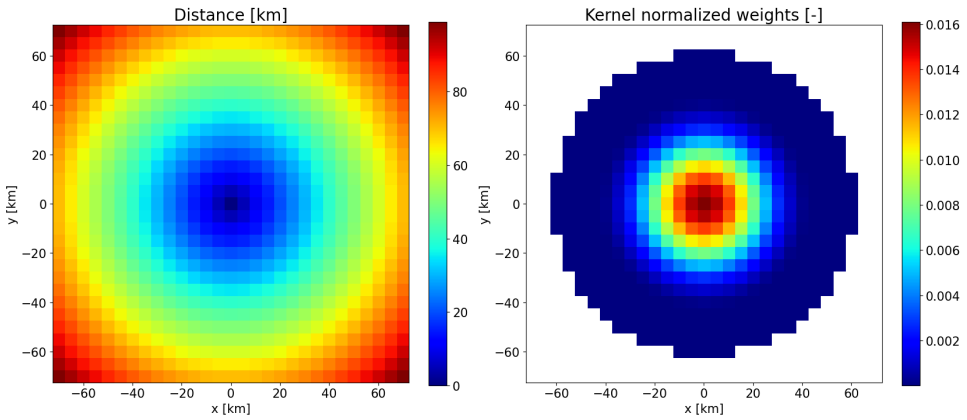


Figure 8.32: Example of the distance matrix and the associated kernel matrix with normalized weights

Giving the formula 8.3, we can already notice that the kernel weights depends only on the latitude/longitude distance between points and their latitudes, and does not depend on points' longitude values. Hence, for two points of the same latitude on a regular grid, the distances between these points and their respective identical neighbourhood will be the same. Thus, the distance matrix and the kernel could only be computed once for each of the different latitude values of the grid in order to speed up the data processing.

Regarding the kernel coefficients formula 8.2, there are three important additional steps.

First of all, as it can be seen on the right plot of Figure 8.32, there are blank pixels where the weights are masked, even if the distance values are present. This is due to the cut-off distance that is defined by the user, and beyond which the weights are considered null. In our case this distance is 61.5 km (thus maximum allowed neighbourhood of 123x123 km²).

Second, the mathematical aspect of the formula allows for the imaginary numbers. It was decided to not take into account the eventual imaginary part of the coefficients, and resolve this problem by replacing all of the imaginary numbers by the minimum real-only number present in the kernel.

Lastly, the weights can be normalized to make the convolution easier and to adapt it for a real physical variable use-case. To do so, we divide each coefficient by the sum of all of the coefficients, so the sum of the final weights makes 1 in total.

In summary, the kernel computation should go through following steps in this order:

1. compute the distance matrix
2. compute the weights following the formula
3. replace imaginary numbers by the minimum of the real-only coefficients
4. replace coefficients beyond the cut-off distance by 0
5. normalize weights

8.5.2 Convolution algorithm

The convolution algorithm contains the kernel calculation and the convolution itself which follows the classical convolution method (Figure 8.33), also used in the Datacube #1 production chain. In our case, the input is the Datacube #1 data, the filter is the SMOSE kernel with normalized weights and the resulting values will constitute the Datacube #2.

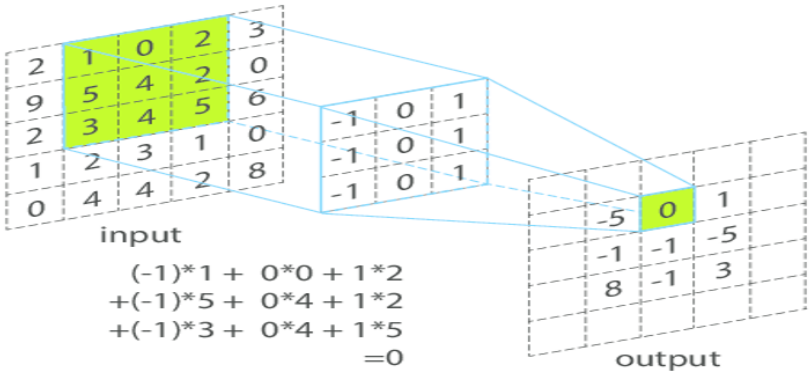


Figure 8.33: 2d image convolution concept

For each of the SMOS EASE grid points, the algorithm is as follows:

1. for a point of the SMOSE EASE grid (which is the grid of the Datacube #2) we find the nearest neighbour in the Datacube #1 grid (which is constant for every Datacube 1 product). This found nearest neighbour will then be the central starting point on the Datacube #1 grid for further computations (Fig. 8.34).

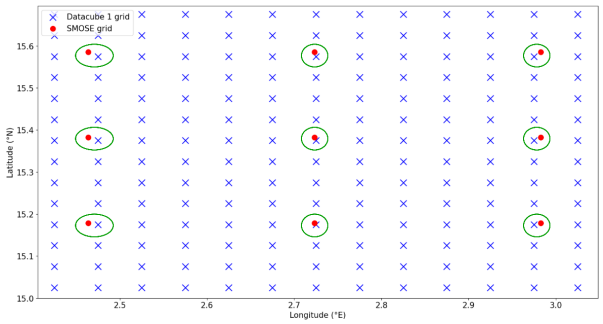


Figure 8.34: Datacube 1 (blue crosses) vs SMOSE (red points) grid, nearest neighbours are circled in green

2. from the found central point, we compute the distance matrix via the formula 8.3 and cut-off any pixels with distance > cut-off distance as they won't be useful in the processing
3. based on the distance matrix, we compute the weights as described in the previous section (coefficient 0 if distance is still > cut-off distance, no imaginary numbers, normalized coefficients)
4. we perform the classical convolution between the computed kernel and the data sub-extract of the Datacube #1 which is overlayed by the kernel and centered on the found nearest neighbour
5. the resulting value corresponds to the product re-projected data value on the SMOSE point coordinates from which we started the step n°1

8.5.3 Invalid/no data pixels contribution

In case when the data overlaid by the kernel contains some no-data pixels or fill values, these pixels are not taken into account during the convolution, the resulting value contains only valid values multiplied by the corresponding coefficients. On Figure 8.35 no-data pixels from the left are propagated to the plot on the right as they must not be taken into account. Nevertheless, the kernel is always computed without taking into account valid/invalid pixels as its purpose is to simulate the instrument behaviour that does not adapt its functioning depending on the seen/received information.

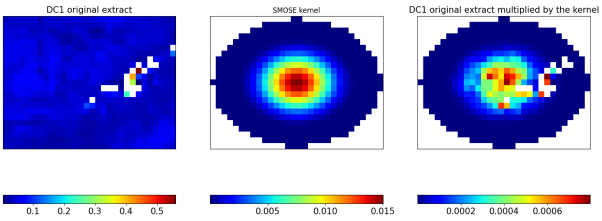


Figure 8.35: Example of the data convolution with no data pixels

8.5.4 Special case n°1 : uncertainties/error variables

In the case of the variables containing uncertainties or error estimation values, the convolution concept remains the same except that the kernel coefficients have been changed so as to calculate the classic average of the kernel covered data. The invalid data is treated in the same manner as explained in the section 8.5.3. The valid data that remains after applying kernel and filtering the invalid data (Fig. 8.36) is averaged to the final value. Kernel's averaging coefficients are calculated on the basis of the valid data only.

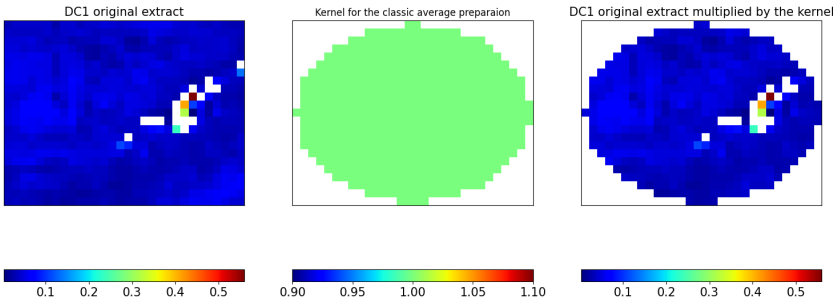


Figure 8.36: Data masking for the average computation of the uncertainties/error estimation variables

8.5.5 Special case n°2 : quality flags

The case of quality flags and other qualitative variables were treated separately as it required a different approach from the physical/quantitative variables presented above. Two scenarios were studied for this part :

1. no projection were performed between source data and Datacube #1. In this case, qualitative variables have not yet undergone any processing.
2. projection between source data and Datacube 1 were performed

8.5.5.1 No projection between source data and Datacube 1

In the case of source data being already in the Datacube #1 projection, no data processing were performed. This data did not undergo any changes. Thus the question of how to project the qualitative data onto the SMOSE grid was addressed.

The retained solution was the same as for the case where we had to reproject the source data into the Datacube 1 grid (section 8.3.1.1): for each flag or qualitative data that we are interested in, we want to compute the percentage of contribution of such pixels to the final SMOS pixel, using the same neighbourhood that in the case of a physical variable (as explained in previous paragraphs). An example is presented on Figure 8.37.

Here we have a flag that represents whether a pixel in the Datacube #1 AVHRR product is recognized as Evergreen Broadleaf Forest or not. When we couple the data extract with a kernel, we end up with the same

neighbourhood as in the physical variable reprojection case. Then, in the remaining data we calculate the contribution as :

$$c = \frac{N_{flag\_pixels}}{N_{valid\_pixels}} \tag{8.5}$$

The resulting value is then affected to the SMOS current pixel, in the same manner as for the physical variable convolution.

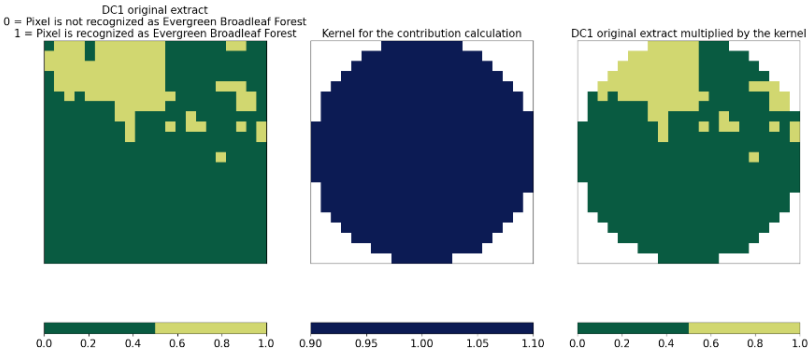


Figure 8.37: Example of a quality flags vs kernel overlay

8.5.5.2 Projection between source data and Datacube 1

In the case where the data needed reprojection between the source grid and the Datacube 1 grid, the quality flags processing has already been assessed. The variables in the Datacube 1 product already represent the contributions of the flags to the Datacube 1 pixels.

Here, we applied the same technique as for the uncertainties and errors data: we compute the classic average of the contributions in the same kernel neighbourhood as for the physical variables.

8.6 Datacube #2 validation

The second Optical Data Cube contains the main data of the product, projected onto a grid of 0.25° spatial resolution (image size 1388W x 584H), as well as some metadata and additional variables which give information on the products used for the aggregation of the data cube.

Global and variable attributes are validated by manual checking. Other variables are validated using global map visualisation and with comparisons of time series from the second Optical Data Cube and the first optical data cube for a given position.

8.6.1 Methodologies

8.6.1.1 Global map comparison

A comparison is performed between the global map generated from the second optical data cube at 0.25° resolution and the first optical data cube at 0.05° resolution. The aim is to check the absence of artefacts.

8.6.1.2 Time series comparison

For a more detailed and accurate comparison between the datacubes, following processing were performed :

1. find the DC2 nearest neighbour pixel to the ground validation site (via the minimal geodesic distance)
2. find the DC1 nearest neighbour pixel to the DC2 pixel center and retrieve its value with the 25 adjacent DC1 pixels
3. calculate the mean and standard deviation of the 25 DC1 pixels, as well as the number of pixels with available data within the selected area
4. find the DC1 nearest neighbour pixel to the ground station
5. compare the time series from DC2, DC1 and from the DC1 25 pixel neighbourhood

It's important to note that in this method we average the 25 nearest pixels and don't use the convolution kernel, which can lead to significant differences in heterogeneous areas. This point will be taken into account in the analysis of the results.

For each time series comparison, the legend is as follows:

red dashed line : variable value from the DC2 center pixel (validation site nearest neighbour)

green dashed line : variable value from the DC1 center pixel (validation site nearest neighbour)

blue line : average value of the 25 DC1 pixels surrounding the center of the DC2 pixel

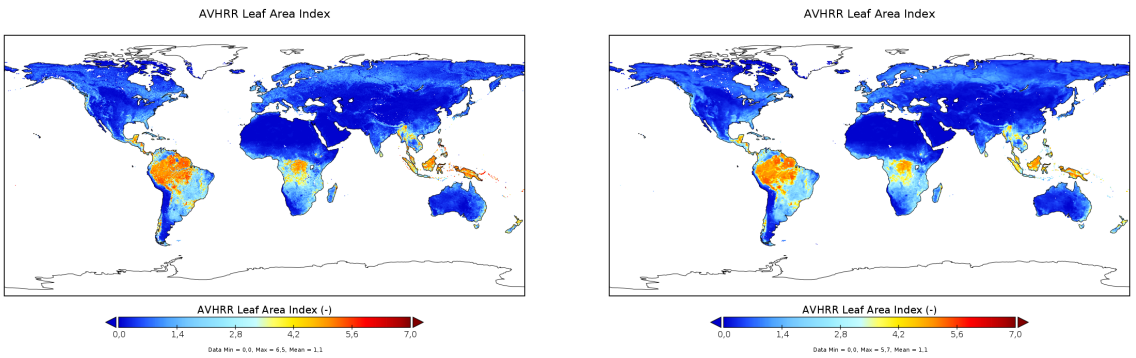
shaded blue area : mean +/- std value of the 25 DC1 pixels surrounding the center of the DC2 pixel

For this validation activity the time series for the three hundred LANDVAL sites were also systematically extracted from the Datacube # 2 (see section 8.4.1.1). The analysis were performed especially on the same five selected ground sites as for the Datacube # 1 validation, but can be easily derived to other LANDVAL stations for which the time series were already generated.

8.6.2 Results

- LAI AVHRR products from POSTEL, Theia (CNES)

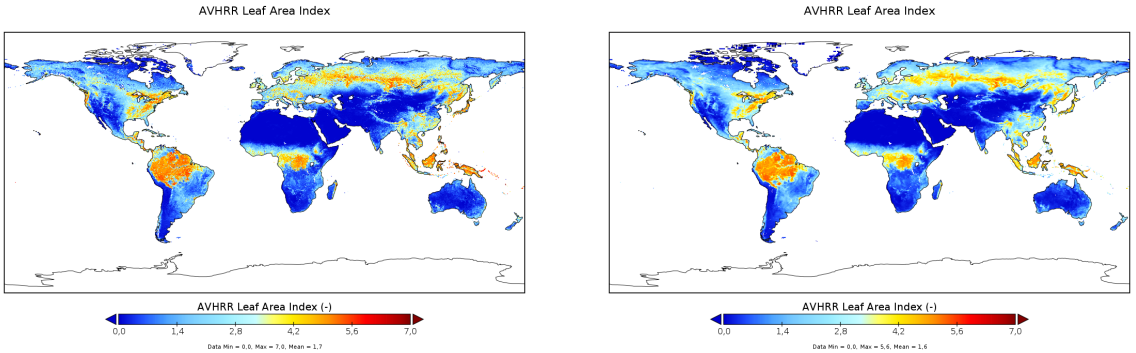
No artefacts or incoherence were observed between the Datacube # 1 and 2 plotted data.



(a) Datacube #1 AVHRR LAI data

(b) Datacube #2 AVHRR LAI data

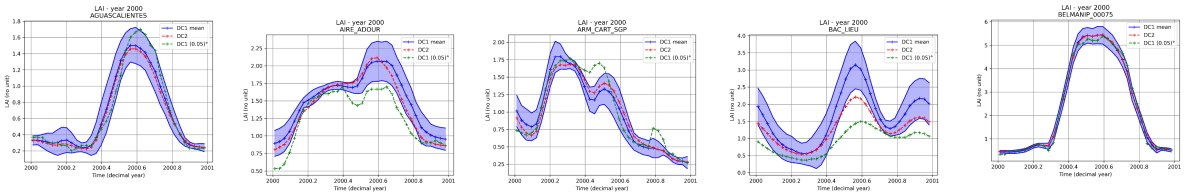
Figure 8.38: Global visual comparison between AVHRR LAI Datacube #1 and 2 data on a January 5, 2000



(a) Datacube #1 AVHRR LAI data

(b) Datacube #2 AVHRR LAI data

Figure 8.39: Global visual comparison between AVHRR LAI Datacube #1 and 2 data on a June 30, 2000



(a) AGUASCALIENTES site (21.7°, -102.32°)

(b) AIRE ADOUR site (43.7°, 0.25°)

(c) ARM CART SGP site (36.64°, -97.5°)

(d) BAC LIEU site (9.28°, 105.73°)

(e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.40: AVHRR LAI time series (year 2000) comparison for the five selected stations

We can see that for each station the patterns are the same for each curve. The green curve, representing the DC1 pixel containing the validation site, is more variable and sometimes falls outside the blue shaded zone. This is explained by the fact that it represents a pixel and may therefore represent an extremum of the value distribution.

In the case of the "Bac Lieu" station, the blue curve is well above the red curve in the middle of the year. This is due to the use of the convolution kernel for the red curve and the use of a simple average over 25 pixels for the blue curve. Convolution is used on many more than 25 pixels, and if the area is very heterogeneous, this can imply significant differences between the simple average and the use of the convolution kernel.

To verify this hypothesis, let's take the following example:

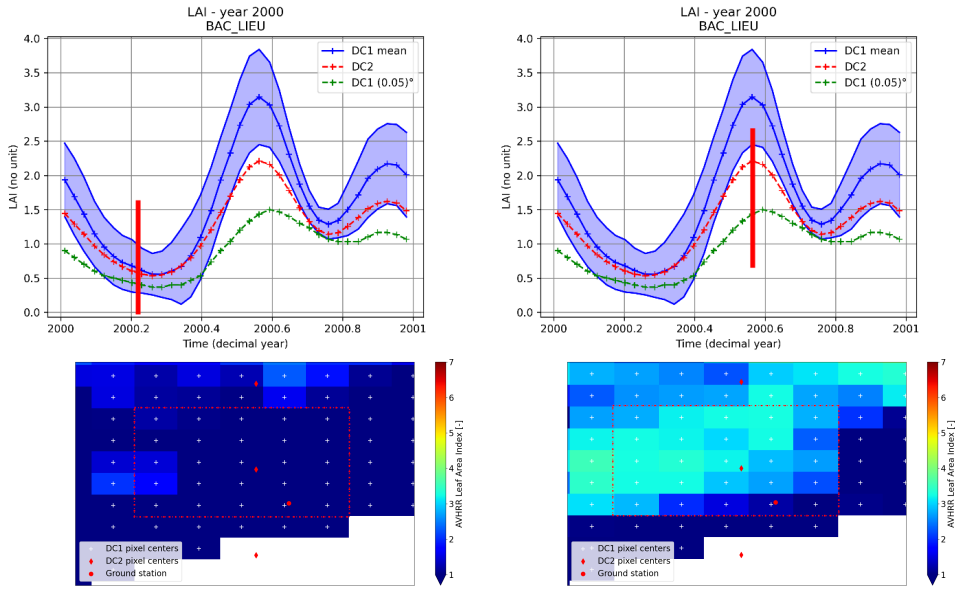


Figure 8.41: LAI time series for the "Bac Lieu" validation site and the corresponding pixels values

Figure 8.41 presents the LAI time series for the "Bac Lieu" validation site and the corresponding pixels values. The upper plots represent the time series and the red vertical bar represents the time for which we will plot the geographical area with LAI values. For the bottom plots the legend is:

white cross : DC1 grid pixel center

red diamond : DC2 grid pixel center

red dot : LANDVAL station

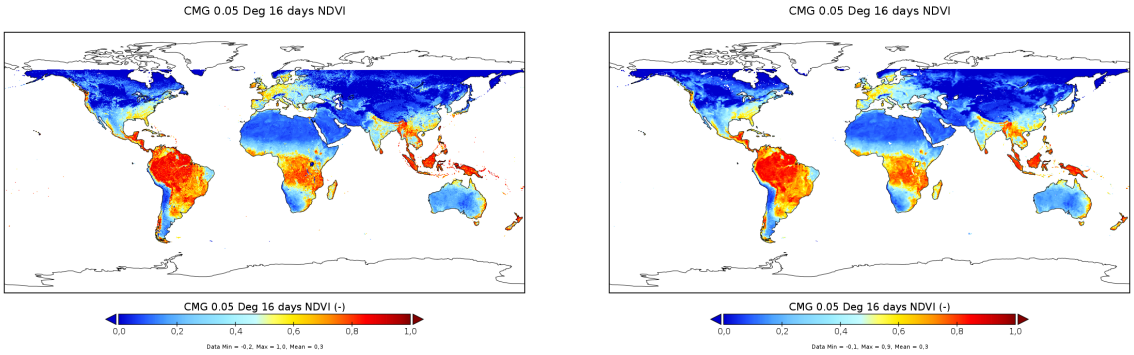
red rectangle : 25 pixel neighbourhood around the DC2 pixel containing the ground station

We can see that at t = 2000.3 the LAI values are homogeneous in the area of interest, so there will be little difference between using the simple average and using the convolution kernel. This is not the case for t = 2000.5. In this case, the simple average zone includes higher values, but low values are nearby. The latter are taken into account when using the convolution kernel and will therefore lower the average. This is why, in this case, the red curve is below the blue area. This type of analysis was used to study similar cases for other variables. However, for the sake of clarity and concision, this example will be the only one detailed in this document.

For the AVHRR collection, the similar analysis was also performed on the FAPAR and FCOVER products, resulting in a good match between the two Datacubes data. The projection was validated.

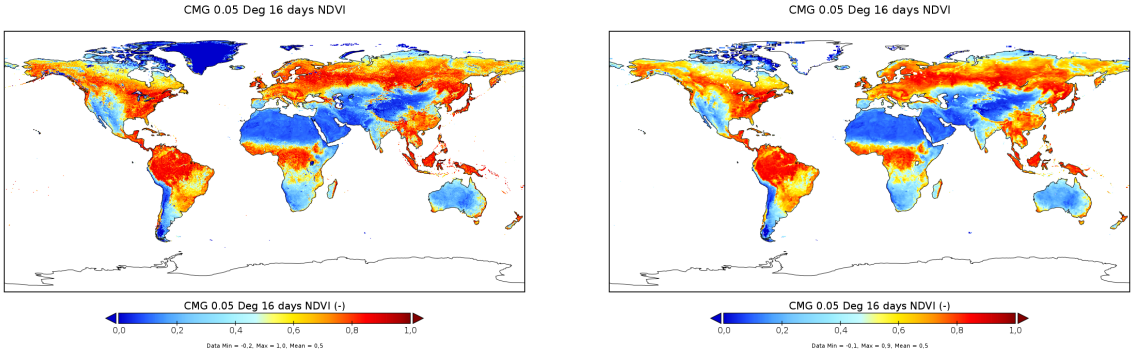
- MOD13C1 NDVI products from USGS

The MOD13C1 product introduces two types of variables to validate : physical variable (NDVI) and quality variables (status or quality flags). First of all, a good match was observed between the Datacubes plots for the physical NDVI variable (Figures 8.42 and 8.43 below).



(a) Datacube #1 MOD13C1 NDVI data (b) Datacube #2 MOD13C1 NDVI data

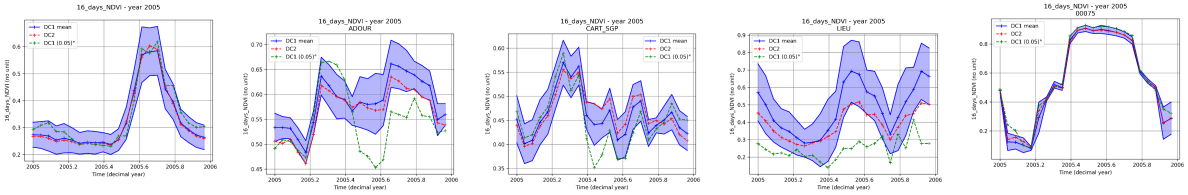
Figure 8.42: Global visual comparison between MOD13C1 NDVI Datacube #1 and 2 data on a January 1, 2005



(a) Datacube #1 MOD13C1 NDVI data (b) Datacube #2 MOD13C1 NDVI data

Figure 8.43: Global visual comparison between MOD13C1 NDVI Datacube #1 and 2 data on a June 26, 2005

The time series for the NDVI were also analyzed in the scope of this product (Fig. 8.44). The curves are following similar tendencies, except some cases where the disparities can be explained as before by the different averaging kernels and the heterogeneity of the station surroundings.



(a) AGUASCALIENTES site (21.7°, -102.32°) (b) AIRE ADOUR site (43.7°, 0.25°) (c) ARM CART SGP site (36.64°, -97.5°) (d) BAC LIEU site (9.28°, 105.73°) (e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.44: MOD13C1 NDVI time series (year 2005) comparison for the five selected stations

Second, the assessment of the quality flag reprojection was performed. In the Datacube # 1 this MODIS data had not yet undergone any projection, so the quality flags in this Datacube are the native ones. Pixel values match certain flag or status in the original data, which is :

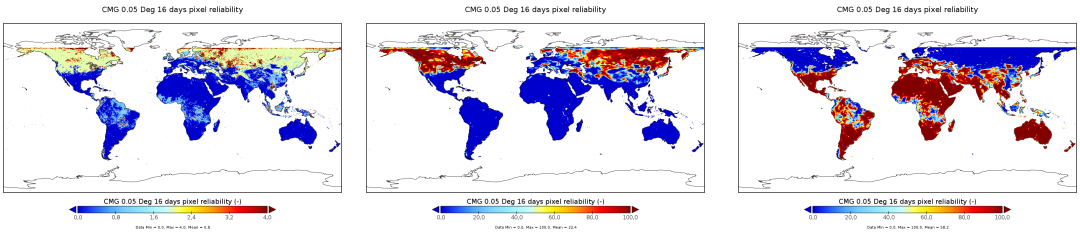
- 0 = good data

- 1 = marginal data
- 2 = snow/ice
- 3 = cloudy
- 4 = estimated data

In the Datacube # 2 on the other hand, these flags were projected into their respective contribution percentages in the final pixels. Thus, the Datacube # 2 has the "16 days pixel reliability" variable with 5 layers, corresponding to the 5 possible states. Each layer gives the percentage of the number of pixels in the corresponding state that contributed to the final Datacube # 2 pixel :

- 0 = percentage of good pixels
- 1 = percentage of marginal pixels
- 2 = percentage of snow/ice pixels
- 3 = percentage of cloudy pixels
- 4 = percentage of estimated pixels

On Figures 8.45 and 8.46 we can observe a good coherence between the original flags and the projected data. Indeed, for the January 1st, 2005 we can see that all of the yellow/green pixels are tagged as snow/ice pixels with a value of 2, and when projected (middle plot on the Fig. 8.45) this area has the biggest percentage of the snow/ice pixels. Same works for the pixels tagged as "good" pixels (blue pixels on the left plot); on the projected data (right plot) we see that the majority of this surface have significant values of the percentage of "good" pixels. Same observations and analysis have been applied for the data on the June 26, 2005 presented on Figure 8.46.

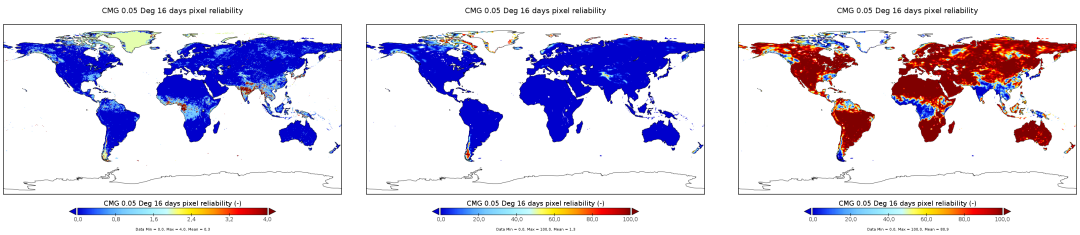


(a) Datacube #1 MOD13C1 pixel reliability flag

(b) Datacube #2 MOD13C1 percentage of snow/ice data

(c) Datacube #2 MOD13C1 percentage of good data

Figure 8.45: Comparison of the MOD13C1 native and projected quality flags in Datacubes # 1 and 2 on January 1, 2005



(a) Datacube #1 MOD13C1 pixel reliability flag

(b) Datacube #2 MOD13C1 percentage of snow/ice data

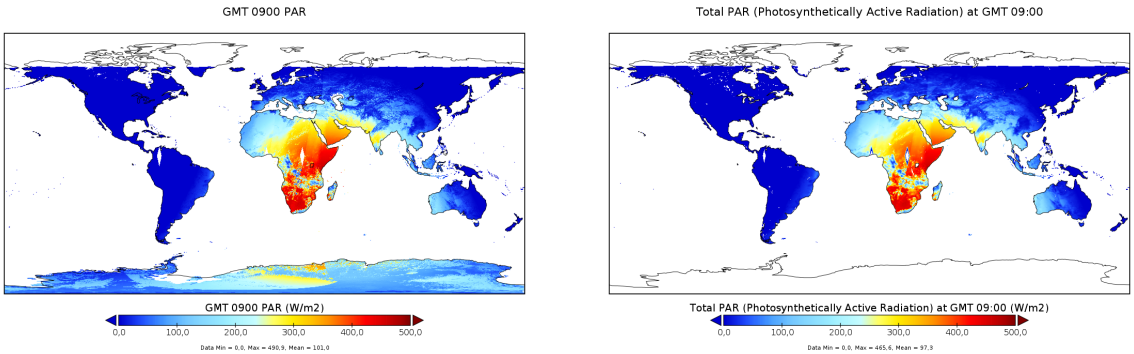
(c) Datacube #2 MOD13C1 percentage of good data

Figure 8.46: Comparison of the MOD13C1 native and projected quality flags in Datacubes # 1 and 2 on June 26, 2005

Based on the observations and analysis from above, the projection of this collection was validated.

- MCD18C2 PAR products from USGS

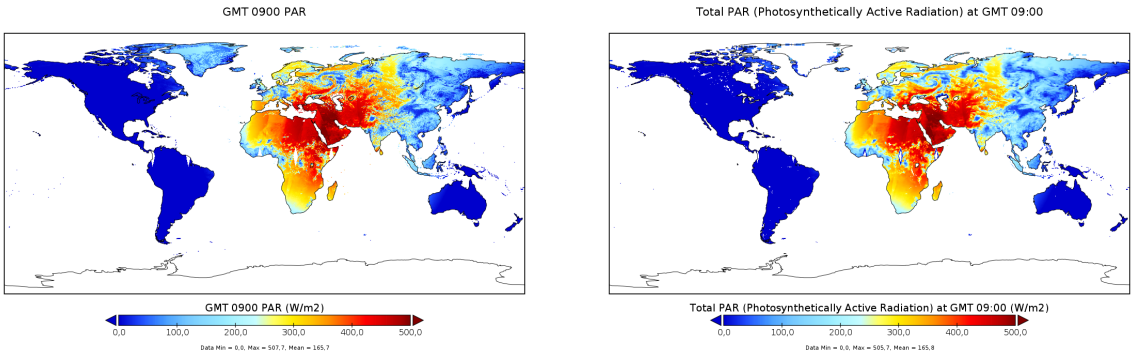
The comparison of the global maps of the PAR variable were performed for every variable subclass (GMT0900, 1200 etc) but only one of the example is presented here for the sake of the concision. Here, a good match is observed between the global maps of the Datacube # 1 and # 2.



(a) MCD18C2 PAR Datacube #1 data

(b) MCD18C2 PAR Datacube #2 data

Figure 8.47: Global visual comparison between MCD18C2 Datacube #1 and 2 data on a January 1, 2005

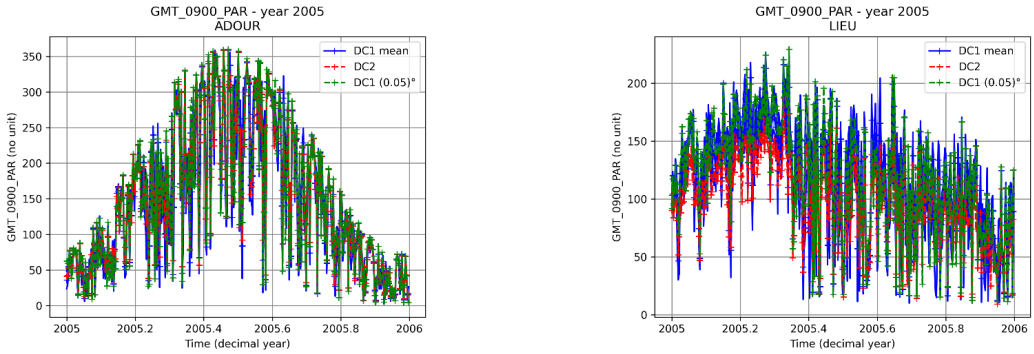


(a) MCD18C2 PAR Datacube #1 data

(b) MCD18C2 PAR Datacube #2 data

Figure 8.48: Global visual comparison between MCD18C2 Datacube #1 and 2 data on a July 1, 2005

For the time series analysis we can also see that for each station the patterns are the same for each curve. The green curve, representing the DC1 pixel containing the validation site, is more variable and sometimes falls outside the blue shaded zone. This is explained by the fact that it represents a pixel and may therefore represent an extremum of the value distribution.



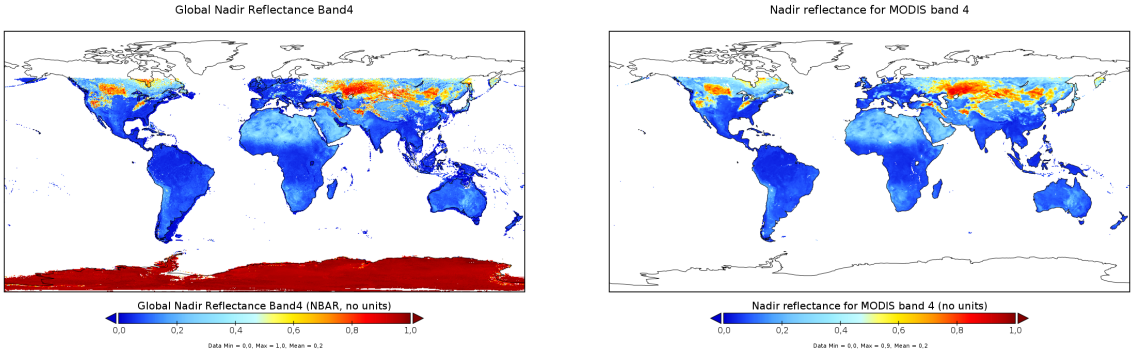
(a) AIRE ADOUR site (43.7°, 0.25°)

(b) BAC LIEU site (9.28°, 105.73°)

Figure 8.49: MCD18C2 GMT0900 PAR time series (year 2005) comparison for two of five stations

- MCD43C4 Nadir Reflectance products from USGS

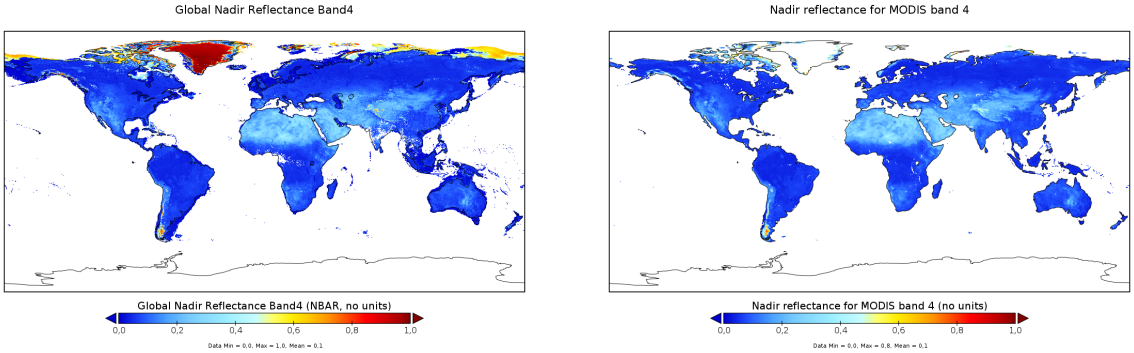
For the MCD43C4 collection no artefacts were observed while comparing global maps of the data. The very high or low latitude zones (Greenland and South pole) were not projected onto the Datacube # 2 grid because the Datacube # 2 grid does not have points in these areas.



(a) MCD43C4 Nadir Reflectance Datacube #1 data

(b) MCD43C4 Nadir Reflectance Datacube #2 data

Figure 8.50: Global visual comparison between MCD43C4 Datacube #1 and 2 data on a January 1, 2005

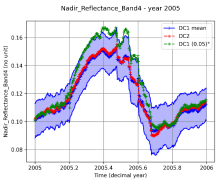


(a) MCD43C4 Nadir Reflectance Datacube #1 data

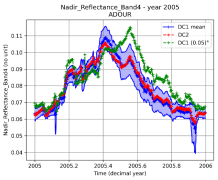
(b) MCD43C4 Nadir Reflectance Datacube #2 data

Figure 8.51: Global visual comparison between MCD43C4 Datacube #1 and 2 data on a July 1, 2005

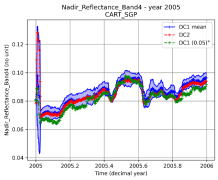
Similar tendencies between the data were also observed on the time series plots (figures below). The projection of this collection was validated.



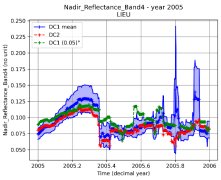
(a) AGUASCALIENTES site (21.7°, -102.32°)



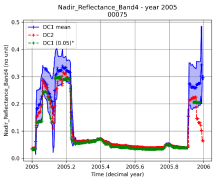
(b) AIRE ADOUR site (43.7°, 0.25°)



(c) ARM CART SGP site (36.64°, -97.5°)



(d) BAC LIEU site (9.28°, 105.73°)

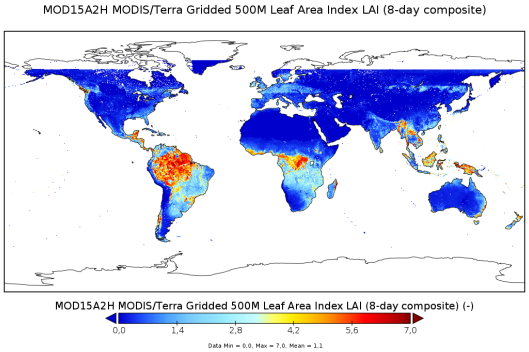


(e) BELMANIP 75 site (41.588198°, -77.852451°)

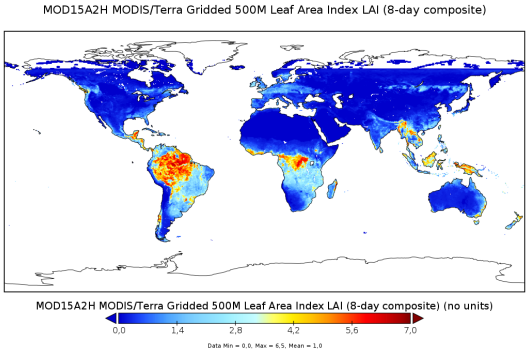
Figure 8.52: MCD43C4 Nadir Reflectance time series (year 2005) comparison for the five selected stations

- MCD15A2H LAI products from USGS

No artefacts were observed on the global maps for the MCD15A2H collection; good matches between the time series of the data for the five selected stations. The projection was validated.

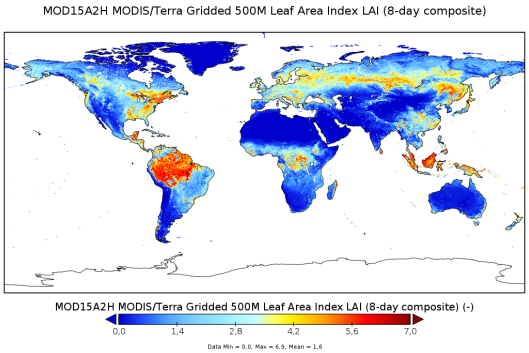


(a) MCD15A2H LAI Datacube #1 data

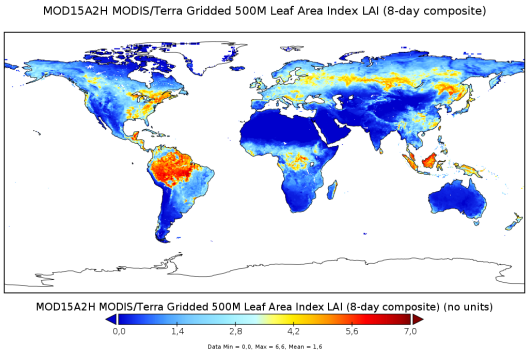


(b) MCD15A2H LAI Datacube #2 data

Figure 8.53: Global visual comparison between MCD15A2H LAI Datacube #1 and 2 data on a January 1, 2005

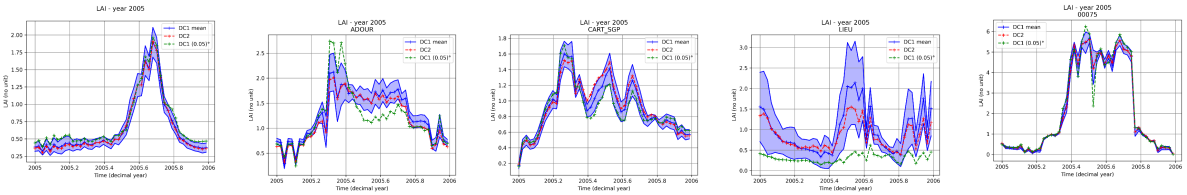


(a) MCD15A2H LAI Datacube #1 data



(b) MCD15A2H LAI Datacube #2 data

Figure 8.54: Global visual comparison MCD15A2H LAI Datacube #1 and 2 data on a June 26, 2005



(a) AGUASCALIENTES site (21.7°, -102.32°)

(b) AIRE ADOUR site (43.7°, 0.25°)

(c) ARM CART SGP site (36.64°, -97.5°)

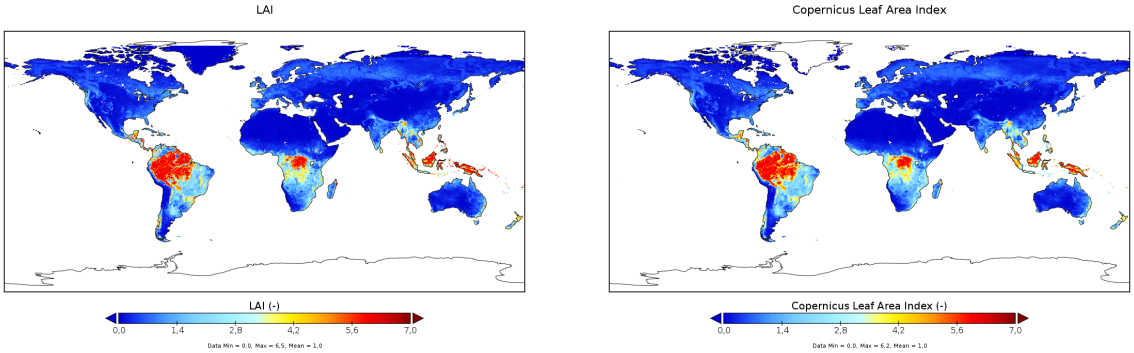
(d) BAC LIEU site (9.28°, 105.73°)

(e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.55: MCD15A2H LAI time series (year 2005) comparison for the five selected stations

- Copernicus Global Land Service LAI products

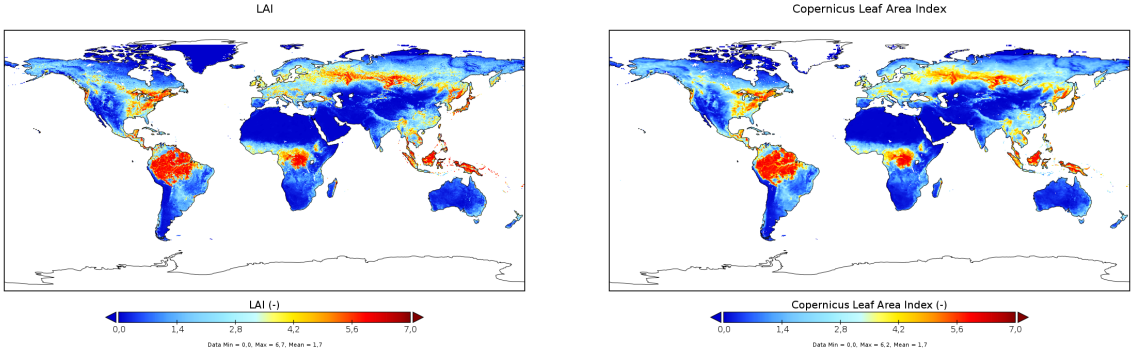
Global map and time series analysis were performed for all of the Copernicus products (LAI, FAPAR, FCOVER, NDVI, TOCR). For the sake of clarity and concision, only the LAI example is presented in the document. Good match and no artefacts were observed during the global map comparison; correct tendencies were observed on the data time series. Same results were demonstrated during other products validation. The projection of this collection was validated.



(a) Copernicus LAI Datacube #1 data

(b) Copernicus LAI Datacube #2 data

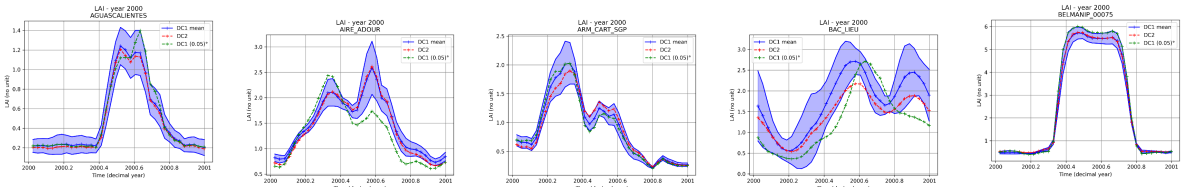
Figure 8.56: Global visual comparison between Copernicus LAI Datacube #1 and 2 data on a January 10, 2000



(a) Copernicus LAI Datacube #1 data

(b) Copernicus LAI Datacube #2 data

Figure 8.57: Global visual comparison between Copernicus LAI Datacube #1 and 2 data on a June 30, 2000



(a) AGUASCALIENTES site (21.7°, -102.32°)

(b) AIRE ADOUR site (43.7°, 0.25°)

(c) ARM CART SGP site (36.64°, -97.5°)

(d) BAC LIEU site (9.28°, 105.73°)

(e) BELMANIP 75 site (41.588198°, -77.852451°)

Figure 8.58: Copernicus LAI time series (year 2000) comparison for the five selected stations

8.6.3 Conclusion

In this section, we have presented detailed methodologies of production and validation of Optical Datacubes #1 and #2 for each of the studied optical collections. In the validation, we first compared global maps, followed by a comparison of time series. The results showed a strong agreement between the two datacubes, thus validating the projection of the data onto the two grids (intermediate grid and EASE2 grid).

Chapter 9

Comparative Data Analysis (WP 10)

9.1 Introduction

The WP10 consists in comparing several datasets used or produced in the project. Ideally, CWC should have been compared to VOD and optical indices but due to delays in the design and production this was not possible. Therefore, the comparison was done with different frequencies VOD, optical indices and soil moisture. The goal was to get more insight into how the remote sensing observations can provide information on the hydraulic behaviour of vegetation.

Vegetation Optical Depth (VOD), derived from microwave observations, has gained significant attention in remote sensing studies due to its sensitivity to water content, in contrast to optical indices. Previous crop-related studies using VOD have predominantly concentrated on phenological analysis at specific microwave frequencies and regional scales. This study examines the seasonal dynamics of multi-frequency VODs and NDVI in comparison to that of soil moisture (SM) across global croplands. The time lags between the peaking periods of different VODs, the NDVI, and the SM signal were used to describe the vegetation's response to water availability. An unsupervised machine learning algorithm (K-means clustering) was applied to classify these temporal relationships across global croplands. The analysis identified eleven distinct clusters, six of which could be related to specific regions and crop types, enabling the analysis of their response to microwave radiation as a function of their potential hydric behaviors. Regions such as North America's corn and wheat belts are distinctly identified. A systematic lag between higher frequency VODs and NDVI was observed across all clusters, ranging from 15 to 49 days. On the contrary the relationship between L-Band VOD (LVOD) and NDVI exhibited significant regional variations: In northern North America, LVOD peaked before NDVI, while in areas characterized by dry/warm weather, they were almost synchronous. Time series similarities and relative time shifts were also computed using the dynamic time warping method for a limited number of points. The comparison with the previous time lag analysis shows agreement with respect the average time lags for the corresponding clusters but not for all of them. This WP provides new insights into cropland water-use strategies linking isohydric (constant water levels) and anisohydric (varying water levels) behaviors to specific crop types and local environmental conditions. This enhances our understanding of passive microwave signals over vegetated areas in agricultural systems worldwide.

9.2 Data

9.2.1 SMOS LVOD and SM

The SMOS satellite mission (*Kerr et al.*, 2010) is the second Earth Explorer mission, and was launched in 2009. Its main objective is to measure soil moisture (SM) and sea surface salinity (SSS) worldwide. The satellite carries a 2D interferometric radiometer and measures Earth's microwave emission at 1.4 GHz (L-Band) at

multiple angles of incidence, ranging from 0 to around 60 degrees. The ascending (descending) orbit crosses the equator at 6:00 (18:00).

The SMOS LVOD and SM are derived from multiangular brightness temperature (TB) observations via the implementation of the L-MEB (L-Band Microwave Emission of the Biosphere) model (*Wigneron et al.*, 2007b; *Kerr et al.*, 2012b). The retrieved values for both variables are derived simultaneously through an iterative process aimed at minimizing the difference between the observed TB and modeled TB *Kerr et al.* (2012b). The level 3 products *Al Bitar et al.* (2017b) (version 330) for LVOD and SM are utilized in this study.

LVOD is a unitless measure that parameterizes the extinction effects (absorption and scattering) of the vegetation layer on surface microwave emissions; lower values represent sparser/lower vegetation and higher values the opposite. SM represents the water content of the superficial layer of soils (around 5cm depth) and is expressed as water volume per unit soil volume (m^3/m^3). Both products are provided by the CATDS (Centre Aval de Traitement de Données SMOS); the dataset covers daily observations since 2010 up to the present. They are projected onto the EASE (Equal-Area Scalable Earth) grid version 2 (*Brodzik et al.*, 2012b), with pixels of 25x25 km at approximately 30 degrees latitude. In this study, datasets covering the period between 2011 and 2020 are utilized. Even though MW at the L-Band are protected for scientific use, they are still sensitive to Radio Frequency Interference (RFI). In that sense, both the LVOD and SM were filtered. Any pixel with a RFI probability appearance larger than 0.2 was excluded. Additionally, pixels where the χ^2 (a variable that measures the goodness of fit in modeled TB with respect to observed TB) surpasses 2 K are not taken into account.

9.2.2 VODCA VOD

The global long-term Microwave Vegetation Optical Depth Climate Archive (VODCA) (*Moesinger et al.*, 2020b) is a harmonized multisensor time series of VOD products. The dataset comprises three time series for the C- (6.9 GHz), X- (10.7 GHz), and Ku-band (19 GHz), derived from level 2 multisensor products (SSM/I, TMI, AMSR-E, AMSR-2, and WindSat). All VOD values are retrieved using the Land Parameter Retrieval Model (LPRM) (*Owe et al.*, 2008). The C-band VOD (CVOD) time series covers the period 2002-2018, the X-band (XVOD) 1997-2018, and the Ku-band (KuVOD) 1987-2017. The merging strategy is based on Cumulative Density Function (CDF) matching, with the AMSR-E product serving as the scaling reference. Temporal aggregation is achieved by overlapping the available observations. The resulting products have a spatial sampling of 0.25°. As part of the harmonization procedure, inaccurate data in the original time series were removed, considering: RFI (using the level 2 RFI flags from the original datasets), land surface temperature (all observations below 0°C were removed), and unrealistic VOD values (negative VOD values were removed). All VODCA time series were re-gridded to the reference SMOS L3 grid (EASE grid v2) by averaging the original pixels into 25x25 km EASE grid cells. In this study, observations from 2011 on-wards are used.

9.2.3 NDVI MODIS

The NDVI from MODIS (Moderate Resolution Imaging Spectro-radiomete, product MYD13C1) (*Tucker*, 1979; *Huete et al.*, 2002) is used as proxy of vegetation health and photosynthetic activity. The original dataset is a 16 day composite with a spatial resolution of 1 km. Similar to the VODCA time series, the dataset is projected onto the SMOS L3 EASE grid v2. Data spanning from 2011 to 2020 are used in this study.

9.2.4 Crop mask

The global crop extent product (GFSAD1KCD) from the Global Food Security-support Analysis Data (GFSAD) (version v001) (*Teluguntla et al.*, 2016) is used in this study. The dataset is produced by the USA's National Aeronautics and Space Administration (NASA) and the United States Geological Survey (USGS) at a spatial resolution of 1 km. It provides cropland extent, water management practices (irrigated or rainfed), and crop dominance (circa 2010). Crop dominance is determined based on datasets from three separate studies on global crop type distribution. Water management practices are derived from the map of irrigated/rainfed

cropland area produced by the International Water Management Institute. The dataset is taken into the EASE grid v2; the dominant value from the finer resolution map (per pixel) is used as the representative value within the coarser EASE grid.

9.2.5 Climate classification mask

The Köppen-Geiger climate classification map (Version 2), provided by GloH2O and developed by (*Beck et al.*, 2023) at 1 km spatial resolution, is used. The map is representative of the period between 1991 to 2020. It is derived primarily from three distinct temperature datasets and four precipitation datasets. The classification comprises five primary climatic groups: A) tropical, B) arid, C) temperate, D) cold (continental), and E) polar climates , each defined by specific temperature and precipitation thresholds.

The original 1 km map was reprojected onto the EASE Grid v2. This transformation involved selecting the predominant value from the higher resolution map to represent the area within each pixel of the EASE grid.

9.2.6 IGBP land surface classification

The International Geosphere-Biosphere Program (IGBP) land cover classification is used (*Loveland et al.*, 2000). Specifically, the dataset produced by *Broxton et al.* (2014b) at 0.5km spatial resolution, which uses the MODIS MCD12Q1 products to create a single map representative of the period 2001-2010. This map is projected into the EASE grid v2 by assigning the predominant class of the original classification into each resulting SMOS L3 pixel. Here, we focus on IGBP cover class 12, which represents croplands. According to *Loveland and Belward* (1997) and *Sulla-Menashe and Friedl*, class 12 corresponds to regions *'with temporary crops followed by harvest and a bare soil period'* where the cultivated area corresponds to $\geq$60 % of the pixels surface. Perennial woody crops are not included in this classification. To exclusively focus on regions dominated by croplands, only pixels classified as IGBP land cover class 12 (across all previously mentioned datasets) were used. This masking, in conjunction with the removal of spurious values in the SMOS dataset (as detailed in Subsection 9.2.1), results in approximately 8000 pixels containing available information classified under IGBP class 12.

9.3 Methods

9.3.1 Scaling and noise removal from time-series

The seasonal behavior of six variables across croplands was studied. The first four served as indicators of water content across vegetation's structural components: LVOD, CVOD, XVOD, and KuVOD. Given the greater sensitivity of lower-frequency microwaves to larger vegetation components, LVOD was considered as a proxy for water content in primary plant structures such as stems, while higher frequencies served as descriptors of smaller components such as small branches and leaves.

The fifth variable, the SM, obtained from the SMOS mission, provided information on water availability for plant uptake. The sixth variable, the NDVI, served as an indicator of croplands' phenology through vegetation greenness.

Before analysis, time-series were pre-treated. To minimize the influence of outliers in the seasonal signal, all values beyond the 0.05 and 0.95 quantiles were removed before smoothing. Given that the original time-series are affected by noise, a 30-day moving average per year was then computed to increase the signal-to-noise ratio.

Once the time-series of each year were smoothed individually, an interannual daily mean was computed. This involved aggregating the daily values across all available years to obtain a single time-series for each pixel; days with less than 3 observations in the period 2011-2020 were removed, just as yearly time-series with less than 180 days.

Finally, and given that the selected variables changed in scale and magnitude, a normalized version of each

variable was used to facilitate the comparison of their seasonal behaviors. The normalization was based on the maximum and minimum values of the interannual mean (per pixel), as follows:

$$\text{Normalized } X_i = \frac{X_i - \max(\mathrm{X})}{\max(\mathrm{X}) - \min(\mathrm{X})}$$

where X is the mean time-series and X_i the value for a given day. Even after normalization, time series can still exhibit noise. To preserve the primary features of the seasonal behavior for each variable and eliminate the influence of local minima or maxima, a smoothing spline is employed to represent the seasonality per pixel.

9.3.2 Descriptors of crops hydraulics

After normalizing the time series, the subsequent step involves identifying periods that reflect plant responsiveness to water availability. Periods characterized by optimal conditions for plant growth are determined using SM, a direct measure of water availability for plant uptake. Simultaneously, periods when vegetation is thriving and nearing peak development are assessed using the VODs and the NDVI. The simultaneous analysis of these periods aims to explore the potential synergies between water availability and plant development.

Considering that these critical moments of optimal conditions are dependent on local factors, such as soil type and meteorological conditions, a local threshold was established for all normalized time-series to standardize the global analysis. *Liu et al.* (2023) compared different approaches for estimating phenological metrics. Their study demonstrated that, when compared to ground data, local thresholds are more accurate for estimating the start of the season (and the end of the season) using NDVI.

This threshold method has been previously used to estimate land surface phenology metrics using vegetation indices - VIs (such as NDVI or LAI) (*White et al.*, 2009; *Garonna et al.*, 2014; *Karkauskaite et al.*, 2017; *Liu et al.*, 2023; *Li et al.*, 2021). *Karkauskaite et al.* (2017) established a local threshold of 20% to determine the start of the growing season in northern latitudes, which is equivalent to 0.2 of the normalized NDVI time series in our study. *Garonna et al.* (2014) used a 50% threshold, while *Liu et al.* (2023) used both 20% and 50%. Additionally, *Li et al.* (2021) used multiple thresholds, defining 75% as the moment of peak growing season.

This document employs then a local threshold of 0.75 to ensure the capture of the period of peak vegetation growth/development conditions.

Hence, a normalized SM value exceeding 0.75 represents the periods where superficial moisture levels have surpassed at least three-fourths of their annual range. Similarly, for the VOD, it indicates that water retention for a given plant component is nearing its relative maximum, and for NDVI, represents the greening-up of vegetation, representing full development and health. Periods when the normalized time series of each variable exceed this threshold will be referred to as the peaking period (PP) for that variable.

Next, considering our objective to understand the seasonal responses of both passive microwaves and optical indices to water availability, the onset of each PP will be used as the primary descriptor of VODs and NDVI reactivity to water availability, hereafter called the 'Start of' (SO) the peaking period. Thus, the day of the year (DOY) when the SM surpasses the 0.75 threshold is defined as SM_{SO}, for LVOD it will be $LVOD_{SO}$, and so forth.

Finally, given the spatial scale of the study area, which encompasses different climatic regions, the DOY for the SM_{SO}, and the other variables, is expected to vary significantly from one location to another. To mitigate the inherent variability in the SO dates, our analysis focuses on the relative timing of VOD's and NDVI's responses to water availability, rather than their absolute calendar dates.

Consequently, the relative lag in days between SM and the different plant descriptors is computed. These lags will be identified as 'Δ + Descriptor'. For instance, the lag between the $LVOD_{SO}$ and the SM_{SO} is computed as $LVOD_{SO}$ - SM_{SO} and is referred to as ΔLVOD, and similarly for the other variables. The PP for the SM is defined as the longest period, of consecutive days, during which the threshold of 0.75 is exceeded. For the remaining variables, the PP is determined as the longest period, with a consecutive number of days above the threshold, closest to that of the SM.

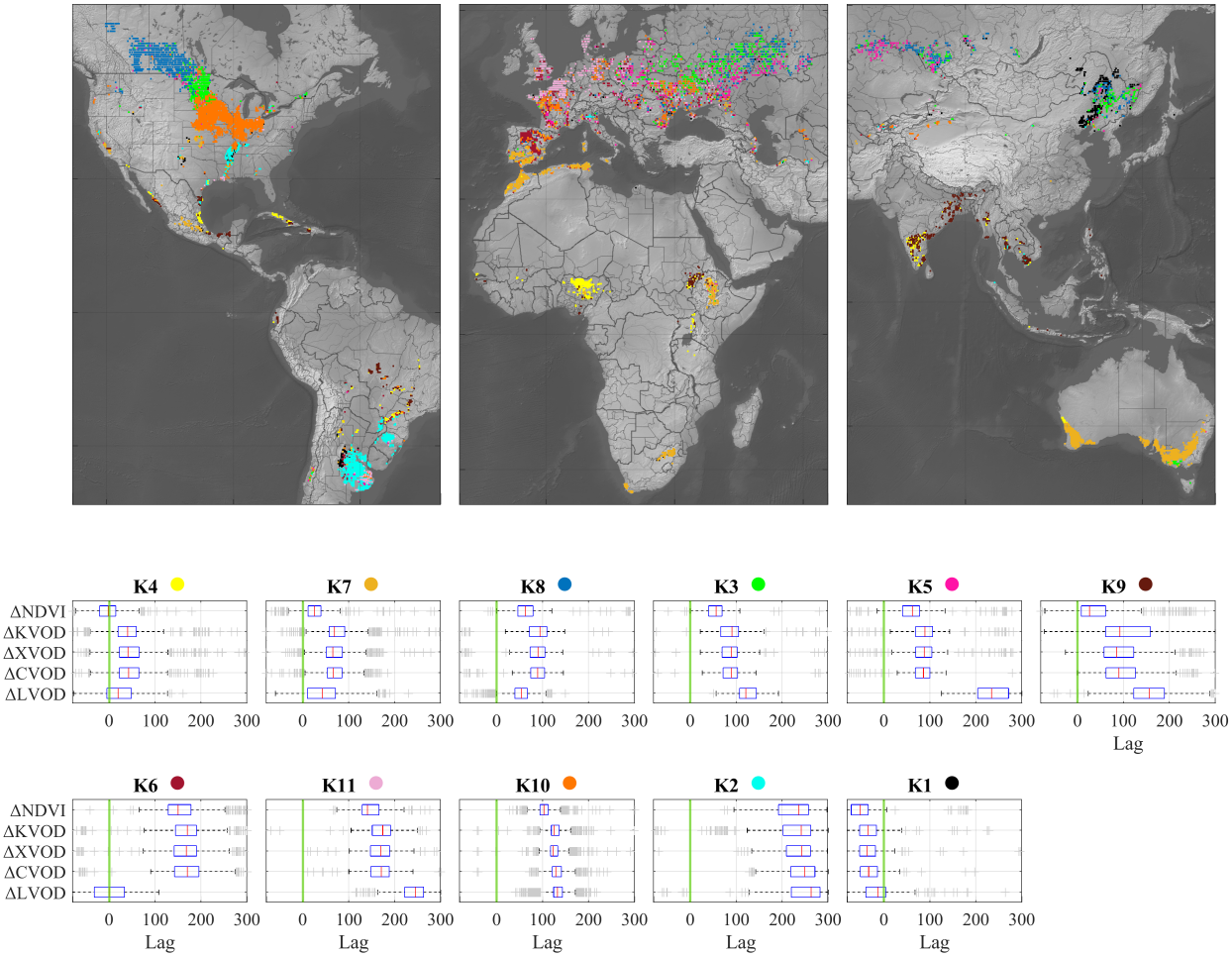


Figure 9.2: Global mean behavior of Δ values per cluster (N=11), lags are given in days. The SM_{SO} is represented as a green vertical line in each box-plot. The order of box-plots has been deliberately modified so cluster with similar behaviors are adjacent to each other.

9.3.3 Group Identification through K-Means Clustering

Given the extensive scale of the study area and the possible representation of each pixel by multiple attributes (e.g., ΔLVOD, ΔCVOD, ΔNDVI), a straightforward pixel-wise classification approach is not feasible. To extract the relationships and patterns within our dataset, an unsupervised machine learning classification algorithm was employed. Specifically, the K-means clustering algorithm was selected for this purpose. K-means clustering operates on the principle of partitioning an input dataset into 'K' distinct clusters based on shared similarities within the data. Initially, the algorithm assigns centroids randomly. Subsequently, each data point is linked to the nearest centroid, thereby forming the initial clusters. The centroids are then recalculated based on the mean values of the data within each cluster, and the distances of each point to its new centroid are recomputed. This iterative process continues until the positions of centroids exhibit insignificant changes.

Finally, due to the sensitivity of the k-means method to the initial set of centroids, 1000 draws were performed, and the one resulting in the smallest intra-cluster sum of squared distances (CSSD) was selected.

9.3.4 Attributes and cluster selection

The K-means clustering algorithm needs two primary inputs: first, a dataset comprising relevant attributes that describe the phenomenon under investigation, and second, a predetermined number of groups (N) in which the dataset will be divided.

The proposed attribute matrix includes six variables. The first four, which are the ΔVODs at different frequencies, describe the reactivity of water storage in various plant components to SM availability. The fifth variable, the ΔNDVI, provides insights into the reactivity of the leafy components of plants to water availability. Finally, to adequately account for external environmental factors that significantly influence both water availability and plant development, the mean temperature (T) is added as the sixth and final attribute.

On the other hand, selecting the optimal N usually depends on prior knowledge of the dataset. However, to our knowledge, no comprehensive classification addressing the hydraulic behaviors of crops at a global scale exists. Consequently, defining an initial number of possible classes is difficult, making any number of clusters theoretically acceptable. In this study, we chose to assess clustering quality by varying the number of N within a specific range.

The climate classification by Köppen-Geiger (*Beck et al.*, 2023) divides the study area into four main climatic zones: A - Tropical, B - Arid, C - Temperate, and D - Continental. Here, we opted to establish the minimum N at 4 (to cover at least the possible climatic regions of the study area) and observe the quality of clustering with increasing values. With too few clusters (2 or 3), the model might lose information and fail to capture essential patterns in the data. Conversely, with too many clusters, the resulting classification may lose interpretability. Clustering quality was evaluated based on the mean Silhouette value (a measure of cluster cohesion) and the geographical distribution of clusters.

9.3.5 Dynamic time warping

In addition to the methods discussed in Sect. 9.3.2, a more automatic method was used to compare different time series for a selected number of points. The goal is also to infer the complementarity of VOD (derived from L, X and C bands) and vegetation phenology indices (derived from optical data).

We choose to perform time-series analysis using a statistical approach for analyzing data that has been structured through time and detect patterns, trends, and anomalies. In time-series analysis, there are numerous approaches for searching for similarities, including the Euclidean distance, intercorrelation computation, dynamic time warping (DTW), and shape-based methods like the Fourier transform and Symbolic Aggregate ApproXimation (SAX).

In this activity, the degree of similarity of different VOD and optical indices time series will be analyzed through three metrics:

- Dynamic Time Warping (DTW) distance

- Pearson correlation
- time shift between time series

The DTW distance is a non-linear, non-time dependent measure of the alignment – similarity – between two time series. Contrary to the classical methods for time series comparisons (i.e. Euclidean distance), which perform a point-to-point comparison, the DTW is able to perform a suitable matching of the time series even if they are out of phase in the time axis or have different lengths (Jeong et al., 2011) and thus could be used to identify anomalies in time series. Pearson correlation and time shift estimation will bring additional information to the analysis about the correlation between time series and their similarities. The time shift estimation was performed via the cross-correlation function which is a measure of similarity of two series as a function of the displacement of one relative to the other. Thus, finding the maximum of the cross-correlation function and its corresponding displacement value means finding the best overlay between the time series and estimating the shift between them.

We describe the comparative analysis of the time series produced, for several locations on the earth's surface, from microwave VOD data (SMOS, AMSRE, AMSR2) and optical AVHRR LAI data variable available in the Datacube #2 archive. In our analysis we were targeting the LANDVAL in-situ sites that correspond to the CROPLAND class from the IGBP classification. The list of such sites was determined beforehand by cross-referencing the list of the LANDVAL sites and the global classification map provided by the IGBP.

To complete this study we performed following steps :

1. Create VOD time series from SMOS database for the filtered CROPLAND sites
2. Analyse obtained time series visually in order to select sites with consistent and continuous data only, sites with few data points or with no apparent patterns were set aside for a latter study
3. Process VOD time series with moving average method to smooth the data and to highlight the underlying pattern
4. Create VOD and AVHRR LAI time series for the final filtered CROPLAND sites
5. Overlay time series for each station, analyze it visually and perform statistical parameters computation

First, in order to filter down the list of CROPLAND sites and to select stations with only viable and consistent data, visual analysis of the time series was performed.

After the visual analysis, the list was filtered down from 120 to 27 CROPLANDs stations. On those stations a pre-processing step was done to prepare the data for the latter comparisons and statistics computation. To smooth the data and to highlight the underlying patterns in the time series, the moving average method was used. The smoothing step allows to make time series cleaner which then allows a more precise and coherent statistics computation when comparing time series between them. This step was then performed on all of the selected stations and for all of the VOD time series (SMOS L, AMSRE C/X, AMSR2 C/X). After performing the smoothing step, we can also plot time series from different satellites together to analyse their similarities and disparities if exist.

After the selection of stations and pre-processing of the VOD data, time series can be plotted with the AVHRR LAI variable for the comparison and statistics computation. To analyse similarities between time series in a more statistical way with quantitative metrics, we used metrics such as Dynamic Time Warping

9.4 Results

9.4.1 Pixel Seasonality from pulse diagrams

The seasonality of the selected variables was systematically analyzed across global croplands. A discernible seasonal pattern emerges across all variables, characterized by a peak occurring approximately between June and September for vegetation descriptors (VODs and NDVI); the temperature time-series on the contrary

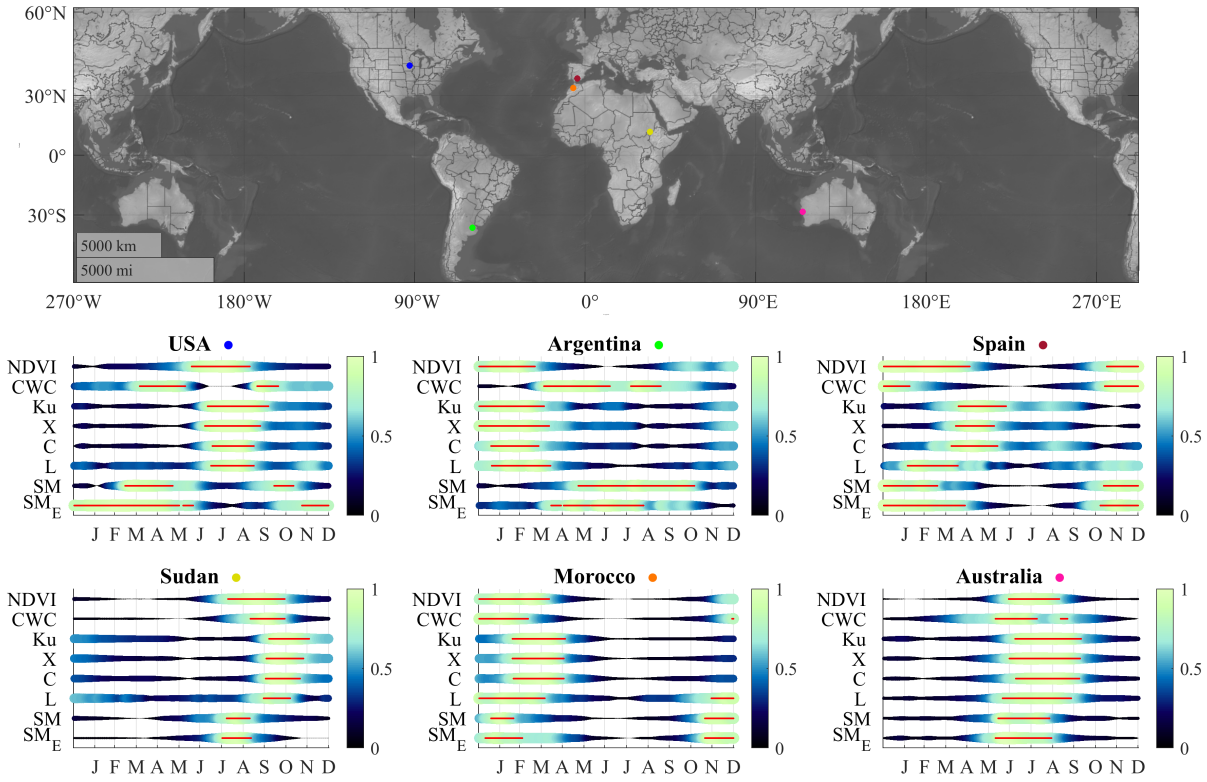


Figure 9.3: Pulse diagrams for single cropland-pixels in USA, Argentina, Spain, Sudan, Morocco and Australia. Mean pulses computed from observations between 2011 and 2020. Red lines correspond to the PP for each variable.

reaches its peak in the first months of the year. The first variable to ascend and reach its maximum after the SM is the NDVI, followed by the VODs. Overall, all frequencies of VOD exhibit synchrony throughout the year.

While figures of this nature offer a comprehensive depiction of the overall behavior of the selected variables, the inherent disparities in scale, ranges, and units prevent a more intuitive interpretation. As described in subsection 9.3.1, in this study we have chosen to represent these seasonal patterns as relative pulses (normalized time-series). Values closer to 1 indicate proximity to the annual maximum, while values close to 0 represent proximity to the minimum. In figure 9.3 (Australia) the size and color of each pulse represent the magnitude of the normalized value: brighter and larger dots indicate values closer to 1, while darker and smaller dots indicate the opposite. The PP of each variable are highlighted in red.

The new time series representation offers enhanced interpretability and facilitates graphical comparison across various geographic locations. Pulse diagrams for different locations are depicted in figure 9.3. Notably, consistent patterns emerge across different regions, as clearly observed in Sudan, Morocco and Australia figure 9.3.

In these locations, the dynamics of $LVOD_{SO}$ and $NDVI_{SO}$ closely track variations in SM_{SO}. However, in Morocco, LVOD starts its peak development nearly synchronously with NDVI, whereas in Sudan, the $NDVI_{SO}$ occurs first. In both regions, the occurrence of higher frequency VODs lags behind SM by approximately 45 days. A comparable trend to that observed in Morocco is discernible in Australia, albeit with a delay of approximately six months. This temporal shift can be attributed to seasonal variations linked to the differing latitudinal positions of the two regions.

At northern latitudes (USA, Spain), the pattern changes drastically. Vegetation descriptors tend to lag several months behind SM_{SO}. While clear similarities/differences in behavior can be observed in different regions, performing a pixel-by-pixel classification is impractical due to the data volume and complex, unique patterns

of individual pixels. Therefore, the utility of the classification algorithm lies in its capacity to group similar patterns efficiently without manual intervention.

9.4.2 Attribute selection

As mentioned in Section 9.3.4, the attributes selected to describe VODs and NDVI reactivity to SM include the delta values (the lag of each variable SO with respect to SM_{SO}) and T. However, when examining the seasonal signal of higher frequency VODs across multiple locations, including but not limited to those shown in figure 9.3, a strong similarity was observed. The time series of these VODs demonstrated coordinated increases and decreases, indicating synchronized seasonal patterns.

Additionally, significant correlations (as R coefficients) between ΔCVOD, ΔXVOD, and ΔKVOD were observed at a global scale; ranging from 0.84 to 0.91. Consequently, only one of the higher frequency VODs was retained. Given that the C-Band provides information on medium-sized components of vegetation, and considering that larger components are already represented by LVOD and leafy vegetation by the NDVI, ΔCVOD was retained as representative of the higher frequency VODs.

Finally, to simplify the selected features and address the remaining correlations between ΔNDVI and ΔCVOD, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the attributes, namely ΔLVOD, ΔCVOD, ΔNDVI, and T. The first principal component (PCA1) accounts for around 54% of the total variance in the dataset, followed by 25% for PCA2 and 14% for PCA3. Therefore, the first three principal components were selected as the final attributes for the clustering algorithm, collectively explaining 93% of the variance in the original dataset.

9.4.3 Unsupervised classification

The geographical distribution of the produced clusters (N=11) is shown in figure 9.2. The lower part of the figure presents the mean Δ values for the VODs and the NDVI per cluster, with the green vertical line representing SM_{SO}. The order of clusters was deliberately adjusted so that clusters with similar Δ behaviors are adjacent.

The first group of clusters comprises K4 (yellow), K7 (golden), and K8 (blue). As observed in subsection **??**, increasing the number of N during the classification process has little effect in their geographical distribution, suggesting distinctive patterns and clear differences in their behaviors. Elements in this group are characterized by a distinct pattern: ΔLVOD tends to behave almost in synchrony with ΔNDVI, while higher frequency ΔVODs are consistently larger than the latter two (see corresponding box-plots in figure 9.2).

Both K4 and K7 are spatially close to each other, located in Australia, Africa, the southern Iberian Peninsula, and tropical America. As presented in Subsection **??**, K4 is a branch of K7 that appears when N reaches 5. Among all eleven clusters, K4 and K7 present the lowest Δ values across all variables. For K4, $NDVI_{SO}$ is almost in synchrony with SM_{SO}, with a mean ΔNDVI close to zero (-2$\pm$42 days). $LVOD_{SO}$ starts later in time with a ΔLVOD of approximately 19$\pm$42 days. In K7, the same pattern is observed, but both ΔLVOD and ΔNDVI increase by about 23 days (42$\pm$41 and 25$\pm$28 days, respectively). ΔCVOD is approximately 42$\pm$50 and 66$\pm$33 for K4 and K7, respectively.

K8 on the other hand is located at northern latitudes, mostly in North America, with lesser extents in Europe and Asia. Overall, K8 increases its Δ values to 54$\pm$30, 89$\pm$28, and 63$\pm$32 for LVOD, CVOD, and NDVI, respectively. It presents three main differences compared to K4 and K7: its geographical position (further north), an increase in Δ values of roughly 20 days (relative to K7), and the tendency of $LVOD_{SO}$ to be closer to SM_{SO} than $NDVI_{SO}$, which, as observed in figure 9.2, is not often the case for other clusters.

The second group corresponds to clusters K3 (green), K5 (magenta), and K9 (brown). All the clusters in this group can be considered subdivisions of the main four clusters at N=4. K3 is the most stable of the three and is the first to appear (at N=5) as a subdivision of the original blue and orange clusters (see N=4 in figure **??**). K5 and K9, on the other hand, are the last clusters to be created, at N=11 and N=10, respectively. K5 is primarily a branch of both the blue and orange clusters at N=4, while K9 is a subdivision of the golden cluster at N=4.

All Δ values in the second group follow a stair-like pattern, with ΔNDVI at the top (56$\pm$24, 62$\pm$46, and 26$\pm$65 days for K3, K5, and K9, respectively), followed by ΔCVOD (89$\pm$27, 86$\pm$45, and 90$\pm$53 days), and finally ΔLVOD (121$\pm$26, 234$\pm$38, and 156$\pm$54 days). Though their Δ values for NDVI and CVOD are similar in magnitude, the main source of their differences lies in their geographical locations and the ΔLVOD behavior (which ranges from 121 to 234 days). K3 and K5 are both located at northern latitudes, while K9 is in the pan-tropical region.

Elements in the third group, comprising K6 (crimson) and K11 (pink), are derivations of the orange cluster at N=4 (see figure **??**). These clusters are primarily located in western Europe and are close to each other, though K11 is situated further north. Two main attributes define the elements in G3: first, the Δ values for NDVI and CVOD surpass 120 days (which is not observed in the previously mentioned clusters), and second, the ΔLVOD can either start at the same time as SM_{SO} (ΔLVOD = 0$\pm$57 days) or much later (245$\pm$34 days). Specifically, ΔNDVI values are 150$\pm$51 days for K6 and 140$\pm$31 days for K11, while ΔCVOD values are 170$\pm$56 days and 171$\pm$32 days, respectively.

The final group corresponds to clusters K1, K2 and K10. These clusters demonstrated stability with increasing N and are part of the original four clusters observed when N=4 ; except for K1 that appears when N reaches 7. Geographically, they are located in America and Northern China, although K10 can have some elements in Europe. In North America, these clusters are adjacent to each other, whereas in South America, K2 dominates the region, with K1 and K10 being less numerous. The main feature of these clusters is the short lag between the Δ values across all attributes, resulting in overall synchrony of all variables. The main differentiating factor between K2 and K1 with respect to K10 is the displacement of all Δ values with respect to SM_{SO}: around 100 days in K10 and over 200 days in K2. K1 and K2 exhibit a similar behavior, differing only in lag direction. The negative lag in K1 could be considered equivalent to the positive lag in K2 due to the time series' cyclical nature.

9.4.4 Assessment of Unsupervised classification using external data

We used the crop mask by *Teluguntla et al.* (2016) and the Köppen-Geiger climate classification (*Beck et al.*, 2023), both introduced in Section 9.2, to cross-validate our results. The geographical distribution of classes from these products aligns closely, likely due to the dependence of agricultural practices on regional climate conditions.

In North America, particularly between Canada and the USA, K8 (blue) and K10 (orange) are the dominant clusters. K8 predominantly represents wheat-barley crops [RA: Wh.Ba.(D)], while K10 represents corn-soy crops [RA: Co.So.(D)]. Both clusters share a similar continental (cold) climate without a dry season, but K8 experiences slightly colder summers than K10.

K8 exhibits an 80% agreement with both the GFSAD1KCD and Köppen-Geiger maps, while K10 shows around 70% agreement. These percentages are computed from the fraction of pixels of each cluster that fall into the crop/climate classifications.

In K8, LVOD is in synchrony with changes NDVI, whereas in K10, it is aligned with CVOD. The water column (represented by the VODs) and the greening of plants (represented by the NDVI) presented a quicker reaction to water availability in K8, with delta values ranging between 50 and 90 days, compared to over 100 days in K10.

Another major cluster in America is K2 (cyan), primarily located in the Mississippi River Basin and the Argentinian Pampas. It shows approximately 83% similarity with the climatic class [Cfa], characterized as a humid subtropical climate. In northern regions, the dominant crop is irrigated wheat-rice [IR: Wh.Ri.(D)], with about 41% agreement, while in the south, it consists mostly of a mix of rain-fed crops [RA: Mix3], accounting for 53%.

In these regions both the VODs and the NDVI seasonal behavior was in synchrony. The CVOD starts its season about 15 days later than the NDVI, this difference increases to 27 days with the LVOD. K2 exhibits the largest Δ values among all clusters (above 180 days).

The fourth main cluster is K3 (green), which appears in North America and Europe, with a small fraction in northern China. It represents areas with mixed crops [RA: Mix3] in continental climates, with warm summers

and no dry season [Dfb]. This cluster covers the transitional regions where the LVOD has a high reactivity to water availability to areas where its response tends to be slower.

K7 (golden) is the predominant cluster in southern latitudes, primarily found in Australia and Africa, with a small presence in the southern Iberian Peninsula (Figure 9.2). In Australia and Africa, it corresponds mainly to wheat-barley crops [RA: Wh.Ba.(D)], with approximately 69% and 55% agreement respectively, while in Europe, it is mainly related to mixed crops [RA: Mix3]. K7's climates feature hot, dry summers in Europe and Africa, and cold semi-arid weather in Australia. Its behavior is similar to K8, with LVOD nearly synchronous to NDVI (followed by CVOD). Overall, the Δ values are shorter than in K8, falling below 50 days (for LVOD and NDVI), indicating higher reactivity in the water column to SM variations.

The sixth of the main clusters is K4 (yellow), which is characteristic of warm weather and primarily located in the pan-tropical region. Approximately half of its pixels are associated with regions dominated by Wheat-Rice-Soy [RA: Wh.Ri.So.(D)], and the other half with areas having a low fraction of croplands [MinorFrac]. K4's behavior is similar to K7 and K8, but it exhibits the lowest Δ values among all clusters, indicating higher sensitivity in the water column to water availability.

Europe and Northern Asia, on the other hand, show the most significant changes when the number of clusters is increased, encompassing seven out of the eleven possible clusters (as shown in figure 9.2). This suggests a high degree of variability and complexity in the attributes used for the classification in these regions.

Some clusters in Europe and Asia exhibit behaviors close to those of the other regions, such as K5, K9, and K11 being similar to K3, possibly indicating areas of transition characterized by highly heterogeneous crops. However, other clusters present clear distinctive patterns. For example, in K6, LVOD is strongly synchronous with SM, but NDVI is not.

9.4.5 Dynamic time warping

The computed metrics of the CROPLANDs stations are presented on a latitude/longitude map on the figure 9.4. The distance and shift values are the median values computed for each station across the five VOD datasets (SMOS L band, AMSRE C and X bands, AMSR2 C and X bands), the AVHRR LAI dataset being the one to which we compare the VOD time series.

On this map we can see that the majority of the stations present same order of magnitude in their respective metrics. The distance values are around 20-40 in average, and the shifts are up to a maximum of 50 days in most of the cases. No apparent spatial variability was found analyzing given points.

We can also overlay these results to the CROPLANDs clusters that were found in Sect. 9.4.3 in order to see of our spatial variability matches cluster distributions. We can see that for some classes we observe similar metrics for the stations inside, such as BELMANIP_00089 and DIRECT_00001_AGRO stations from the class n°10. However we also have sites in the Australian region that have different metrics while being in the same n°7 cluster.

9.5 Discussion

In this study, we aimed to understand the seasonal behavior of PMW optical depth and NDVI in response to water availability over croplands. This analysis resulted in a well-structured geographical classification of vegetation based on the sensitivity of these variables to SM variations, revealing the potential hydric behavior of plants in different cropland areas.

9.5.1 Cluster Interpretation

As shown in Section 9.4, some of the clusters exhibit a clear agreement with the geographical distribution of the different classes from crop type and the Köppen-Geiger maps. Thus, it seems possible to relate the joint response of PMW and NDVI to SM as a function of agricultural practices and climate conditions. The classification found in Section 9.4.3, explained in terms of time lags, clearly depicts well-known regions.

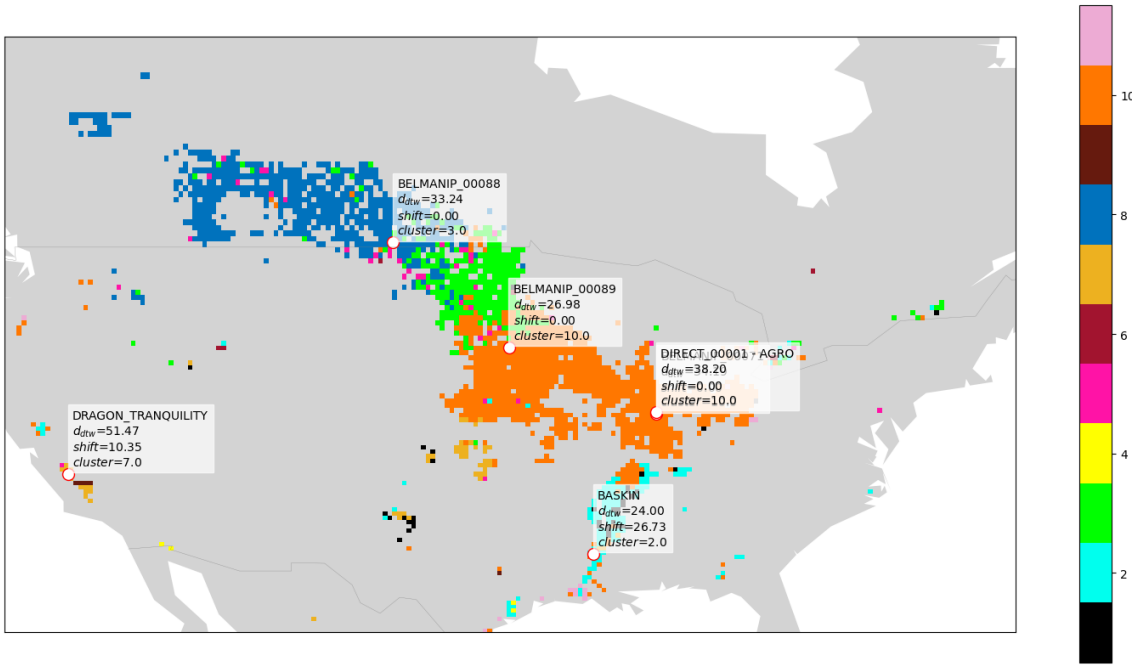


Figure 9.4: CROPLANDs stations with their metrics overlaid on clusters in North America.

For example, the cluster K10 presented strong spatial agreement with the USA's corn belt, as observed in figure 9.2. This region is characterized by strong synchronization among all VODs, which consistently lag behind the NDVI.

Furthermore, cluster K8 predominantly spans North America's wheat belt, while K7 encompasses the southern coast of Australia, the Mediterranean region, South Africa, and the majority of croplands in Ethiopia. Both clusters exhibit a distinctive synchrony between LVOD and NDVI (as shown in figure 9.2), and are predominantly found in regions where wheat is a major crop. A similar behavior is observed in K4, situated within the semi-arid Sahel belt; although it contains wheat crops, it also presents a significant fraction of non-agricultural low vegetation. We consider, therefore, that K4 and K7 represent warm arid ecosystems in Africa and Australia, predominantly featuring wheat crops. In contrast, K8 corresponds to cold continental regions in North America, which are also associated with wheat cultivation (USA's wheat belt).

Interesting behaviors were observed in other clusters, particularly K2 and K1, where the VODs and NDVI reached their peak development with a significant temporal offset with respect to SM's. The K2 cluster is found in three distinct regions: the Mississippi River basin (part of the USA's rice belt), the Argentinian pampas and China's Northern rice region. Regions like Europe, known for having a mosaic of different crop types, indeed showed the widest variety of clusters, underlying the high degree of variability in the attributes used for the clustering over this region.

While the clustering identified several classes, only the six mentioned before corresponded to well-defined geographical regions linked to specific cropland types. Therefore, in the next section, a potential interpretation of the response of PMW VODs and NDVI to water availability in those regions is given. Other regions exhibiting interesting patterns, like those represented by the remaining clusters require further research.

9.5.2 Temporal lags between PMWs and spectral indices

Previous studies on PMW applications for crops have mainly focused on the sensitivity of microwaves to plant phenology within the United States (*Jones et al.*, 2012), particularly in the region known as the USA's corn-belt (*Patton and Hornbuckle*, 2013; *Lawrence et al.*, 2014b; *Hornbuckle et al.*, 2016; *Guan et al.*, 2017). A systematic lag between the peaks of LAI and the LVOD, with LAI appearing around 19 days earlier, was observed by (*Lawrence et al.*, 2014b) in some locations of this region. The mean seasonal behavior of the

XVOD over the corn-belt was studied by (*Guan et al.*, 2017), they observed that the NDVI started its growing season before the XVOD by around 1 month. Our results align with (*Guan et al.*, 2017) regarding the lags between XVOD and NDVI. The observed lags were around 25 days with a comparable response in LVOD of approximately 28 days. The observed lags are slightly longer than those reported by *Lawrence et al.* (2014b) by around 6 days.

Additionally, *Vreugdenhil et al.* (2017), using active CVOD over Australia, observed that LAI reached its growing season 1 to 2 months before CVOD over croplands; here we observed a similar behavior and estimated it at around 37 days. Field experiments with L-Band radiometers in Europe estimated that the NDVI appeared around two months before the LVOD over a mixed parcel in eastern France, where approximately 55% of the surface was identified as croplands (*Barrée et al.*, 2021). The absence of data specifically corresponding to that region in our study prevents direct validation of these observations.

Other research has observed this lag between optical and PMW indices across different cover types, besides croplands (*Jones et al.*, 2011, 2012). *Jones et al.* (2011) observed that in regions with low vegetation, the KVOD lags behind various VIs, although no estimation of actual time lags was provided. A similar approach was adopted by Jones et al. in a subsequent study (*Jones et al.*, 2012), but focused exclusively on North America. In their study, across various cover types, they observed that in southern latitudes, the XVOD reached its growing season, after the NDVI by approximately 5.7 weeks (around 40 days). We estimated this temporal lag to be approximately 26 days. They also noted that in northern latitudes of North America, the XVOD appeared before the NDVI by about 7.5 weeks (around 53 days). Our observations, however, revealed contrasting patterns over croplands in this region: the NDVI peaked about 26 days before the high-frequency VODs. Possible explanations for the differences between our findings and those of Jones et al. include the methods used to estimate the start of the season and the exclusive focus on croplands in our study, as opposed to various cover types in theirs.

In this study, these observations were extended to croplands around the world, not just in North America or Australia, and to multiple frequencies. A systematic lag between NDVI and higher frequencies VOD in all clusters was observed, ranging from 15 to 49 days. Furthermore, the current analysis revealed that, over croplands, the LVOD does not consistently lag behind the NDVI. Unlike higher frequency VODs, LVOD exhibits variable timing in reaching its peak development relative to NDVI.

9.5.3 Physical interpretation of the vegetation parameters

The delayed response of high-frequency VODs compared to NDVI, can be attributed to the sensitivities of both variables to different aspects of vegetation. NDVI responds to leaf greening, while the VOD is sensitive to the water stored in the different components of vegetation. Across the entire study area, photosynthetic activity (NDVI) reaches its peak development initially, followed by water accumulation in the finer components of vegetation (high-frequency VODs) later in the growth cycle. This behavior is similar to that found in field measurements of C-Band radar over cornfields, which showed that LAI increases sooner and faster than canopy water content (*Wigneron et al.*, 1999).

In contrast, the timing of LVOD relative to both NDVI and high-frequency VODs appears to vary with crop type. In the corn belt of the United States, LVOD reaches its peak growth in synchrony with high-frequency VODs, and both occur after the NDVI. This pattern suggests that water distribution within the plant's components—such as leaves, branches, and stem—remains relatively uniform throughout its development. Additionally, both NDVI and VODs reach their peak approximately 120 days after SM does. This delay between water availability and peak vegetation development indicates a slower rate of vegetative growth following the onset of water availability. In contrast, in regions where wheat is the dominant crop, such as Africa and Australia, VODs and NDVI respond more rapidly to water availability, peaking within 66 days after the SM signal. Unlike areas where corn is the primary crop, the LVOD in these regions closely follows the NDVI and appears before the high-frequency VODs. It appears that vegetation prioritizes photosynthetic activity first, followed by water storage in larger components, and only then stores water in smaller components.

It has been shown (*Nguyen et al.*, 2022) that under water stress conditions, corn exhibits isohydric behavior (constant water potential), maintaining stable leaf water potential (LWP) values. In contrast, winter wheat displays anisohydric characteristics (variable water potential), showing more negative LWP values during water stress episodes. Moreover, a comparative analysis of these two crops (*Wang et al.*, 2012) showed a disparity

in transpiration rates, with corn fields exhibiting consistently higher levels than winter wheat. *Wang et al.* (2012) observed that the transpiration in summer corn accounted for ≈71% of total evapotranspiration, while for winter wheat, it accounted for ≈61%. The stable LWP and higher transpiration rate suggest that corn is actively managing its water use to maintain more consistent water levels throughout its tissues during growth, which may explain the observed synchrony in PMW VODs. Wheat, on the other hand, has been shown to use water more efficiently (*Wang et al.*, 2012) and exhibit distinct physiological strategies, as observed in the adjustment of LWP to water availability by *Nguyen et al.* (2022). The efficient use of available water could explain the faster reaction time to SM pulses. Changes of LWP and smaller transpiration rates may explain why LVOD peaks before high-frequency VODs. Vegetation in these regions seems to prioritize maintaining hydration in essential tissues, particularly larger structural components (indicated by LVOD). Only after, plants allocate water to smaller, more dynamic parts (captured by high-frequency VODs) under fluctuating water availability conditions.

Nonetheless, we also noted that LVOD precedes the NDVI in the USA's wheat-belt and in the Iberian peninsula. This behavior has been observed in other types of vegetation and for other VOD frequencies relative to NDVI. For instance, *Jones et al.* (2011) found that XVOD initiated its growing season before NDVI in vegetation over northern latitudes of North America. They proposed that cold temperatures inhibit vegetation's photosynthetic activity, explaining the temporal offset between VOD and NDVI signals. Their hypothesis suggests that plants may begin water uptake in non-photosynthetic tissue (represented by LVOD or XVOD) before temperatures are warm enough to initiate significant chlorophyll production, thus accounting for the delayed appearance of NDVI relative to VOD. This behavior could apply to perennial crops, where woody structures such as trunks and branches are maintained between growing seasons. However, it is less likely for annual crops, which perish and regrow within the same season and are the focus of our study. Moreover, *Tong et al.* (2019) observed a similar behavior (XVOD preceding NDVI) in Northern Australian shrublands, a region characterized by warm weather. This observation suggests that low temperatures may not be the sole driver behind this behavior. Further research is needed in regions where this particular pattern is observed.

Furthermore, we observed that in the USA's rice-belt, the Argentinian pampas and China's northern rice region, there is a considerable temporal offset in the seasonal signal of VODs and NDVI with respect to SM. *Bousquet et al.* (2021) found negative correlations between LVOD and SM in seasonally inundated areas, such as wetlands and floodplains. They suggested that this behavior is more pronounced in grasslands than in forests because low vegetation is often submerged and becomes invisible to sensors, while standing water becomes more prominent in the signal. This could explain the substantial Δ values observed in these regions which are prone to seasonal inundation, including potential areas of paddy cultivation. This is particularly evident in the Mississippi basin, where *Bousquet et al.* (2021) identified a highly dynamic water fraction. Although less prominent, their results also showed that the Argentinian Pampas and northern China exhibited relatively high dynamic water fractions compared to other regions. Similar phase opposition between VOD and VIs, such as the LAI and the Gross Primary Production (GPP), has been documented in tropical woodlands (*Tian et al.*, 2018), evergreen broadleaf forests, regions with frequent inundations (*Jones et al.*, 2011), and over wetlands (*Teubner et al.*, 2018).

Moreover, a link between weather conditions and the response time of the different variables to SM was observed. Regions in Australia, Africa, and North America where the main crop is considered to be wheat exhibited shorter reaction times to SM pulses when located in drier and warmer regions. For instance, in Australia and Africa, VODs and NDVI are more responsive to water availability compared to North America. Specifically, LVOD reacts up to 35 days faster, NDVI up to 60 days faster, and CVOD up to 47 days faster than in North America. In crops near the Sahel region of Africa, which is known for its dry and warm conditions, the fastest reaction of VODs and NDVI to SM was observed ; the lag for ΔLVOD was approximately 19 days, while ΔNDVI was nearly synchronous with SM. This suggests that the water column stored in crops (captured by PWM VODs) in these regions is more sensitivity to water availability, and that this sensitivity seems to be driven by harsher climatic and environmental conditions. This hypothesis is supported by studies like *Huxman et al.* (2004) which showed that photosynthetic activity in herbaceous vegetation in arid and semi-arid ecosystems responds rapidly and significantly to water availability.

More studies are needed to understand the PMW signals over vegetated regions with particular behaviors, like those where LVOD is strongly synchronized with SM but not with other VIs. As shown by *Liu et al.* (2021), classifying vegetation by hydric properties, such as those inferred from VOD, including vegetation water content or LWP, could enhance the parameterization of plant hydraulics in large-scale land surface

models.

9.5.4 Comparing to dynamic time warping

The dynamic time warping analysis was done for a small number of points and it is not possible to define a particular spatial pattern or apparent clusters. The comparison of DTW results with the pulses analysis cannot be done systematically for the same reason. The results are sometimes similar to those of the time pulses analysis. For instance for the points shown in figure 9.4 BELMANIP_00088 belong to cluster 3 (green, Table **??**) for which the LVOD-NDVI lag is around 68 days on average, while the DTW analysis for this point suggest no shift. Regarding the two sites in the orange cluster (cluster 10), they also suggest no shift while on average the orange cluster has a LVOD-NDVI lag of 27 days. In contrast, the shift computed with DTW for the points in clusters 7 (gold) and 2 (cyan) are 10.35 and 26.7 days, respectively. This values are in better agreemnt with those of the pulse analysis which are 15 and 27 days, respectively.

9.6 Conclusion

By integrating data from passive microwave and optical sensors, this study presents a geographical classification of vegetation over croplands based on the seasonal responses of multi-frequency VODs and NDVI to SM pulses. Results from the clustering algorithm align with known agricultural practices and climatic classifications, highlighting the capability to capture significant ecological and agricultural patterns in vegetation. The analysis identified eleven distinct classes, of which six could be related to distinct (climate) regions and crop types.

Globally, we identified a systematic lag between higher frequency VODs and NDVI, ranging from 15 to 49 days. In contrast, the seasonality of the LVOD with respect to high-frequency VODs and NDVI varied with location. In wheat-dominated areas, the vegetation's water column and photosynthetic activity were highly reactive to SM pulses. We observed that the water content in vegetation's larger structural components, as indicated by LVOD, exhibited near synchrony with NDVI and a faster response to water availability compared to that in smaller, more dynamic plant parts, which are better captured by high-frequency VODs. Conversely, in corn-dominated regions, such as the USA's corn-belt, the PMW VODs and optical indices exhibited a more gradual response to water availability. In these areas, the LVOD was in synchrony with the high-frequency VODs and both lagged behind the NDVI. We hypothesize that this multi-frequency analysis reveals the isohydric nature of regions predominantly cultivating corn and the anisohydric characteristics of areas where wheat is the primary crop.

Additionally, the classification effectively isolated regions with particular seasonal behaviors, like the Mississippi River basin, the Argentinian pampas and China's northern rice region, where PMW VODs and NDVI lagged significantly with respect to SM. This behavior seems to be related to seasonal changes in the water fraction. Other locations, presenting particular patterns such as LVOD in synchrony with SM but not with NDVI/CVOD require additional investigation. This underscores the critical importance of studying PWM in vegetated regions to understand how surface conditions affect the signal.

These findings highlight the benefits of integrating passive microwave and optical indices to better understand the complex interactions between vegetation and water availability. However, the spatial resolution of the data, while suitable for large-scale analysis, may overlook finer-scale variations within croplands. The clear patterns observed in certain regions may be attributed to spatial resolution and extensive monocultural practices. In fragmented landscapes like Europe, the clustering exhibited reduced efficacy in identifying distinct patterns. Subsequent research could incorporate field measurements to validate and refine our classifications.

Chapter 10

Deliverables, progress reports, meetings

10.1 Deliverables and reports

Table 10.1 presents the OSMOSE project deliverables, and Table 10.2 gives the progress reports that were prepared regularly.

Table 10.1: OSMOSE deliverables table

Task #	Deliverable	Long Name	Type	Reference
T0	D1	Web Site	web	https://osmose.magellium.com/
T1	D2	Theoretical Analyses Microwave Radiative Transfer	TN 1	TN_OSMOSE_WP1_22Jun23.pdf
T2	D3	ATBD microwaves	TN 2	OSMOSE_D03TN02_ATBD_VODchain_20240109.pdf
T2	D4	Processing Algorithms for the generation of the VOD	Algo.	
T3	D5	Data base for the generation of VOD	Data	
T3	D6	Data Base Manual	TN 3	OSMOSE-NT-025-MAG-D6 OSMOSE_D06TN03_ManualDataBase_20240118
T4	D7	Output Database VOD	Data	
T4	D8	VOD data set documentation and analysis report	TN 4	OSMOSE_D08TN04_VOD_Analysis_20240402.pdf (ref. SO-TN-CB-GS-0118)
T5	D9	AGB data documentation and analyses report	TN 5	OSMOSE_WP5_report.pdf (ref. SO-TN-CB-GS-0117)
T5	D10	AGB estimates from VOD and TB	Data	
T6	D11	Optical data	TN 6	Review_VegTraits_Estimation.pdf
T7	D12	Pre processing ATBD Optical	TN 7	OSMOSE_ATBD_VWC_Optical_V3.0 OSMOSE-NT-007-MAG-v1.0_D12
T7	D13	ATBD Second Optical Data Cube: - description - validation manual	TN8	OSMOSE-NT-024-MAG-D13.pdf OSMOSE-NT-006-MAG-D13.pdf OSMOSE-NT-037-MAG.pdf
T7	D14	Pre processing algorithm	Algo.	
T8	D15	Input data base for optical	Data	OSMOSE-NT-008-MAG-v1.0_D15
T8	D16	Update technical note 3	TN 3	OSMOSE-NT-025-MAG-D16_v1.0
T9	D17	Output database optical	Data	
T9	D18	Optical Vegetation Data manual	TN 9	OSMOSE-NT-027-MAG-v1.0_D18
T10	D19	Data analysis report	TN 10	SO-TN-CB-GS-0125

Table 10.1: OSMOSE deliverables table

Task #	Deliverable	Long Name	Type	Reference
T11	D20	Road map	TN 11	SO-TN-CB-GS-0121
	FR	Final report	TN	SO-TN-CB-GS-0119
	SR	Summary report	TN	SO-TN-CB-GS-0120
	FP	Final presentation	TN	

Table 10.2: OSMOSE Progress Report table

Number	Period	Long Name
01	08/09/2022-15/11/2022	OSMOSE-PR-013-MAG-V1.0-ProgressReport_20221118_v1.docx.pdf
02	06/12/2022-24/01/2023	OSMOSE-PR-019-MAG-V1.0-ProgressReport_20230120_v1.docx
03	18/02/2023-20/04/2023	OSMOSE-PR-026-MAG-V1.0-ProgressReport_20230417_v1.pdf
04	06/10/2023-06/12/2023	OSMOSE-PR-038-MAG-V1.0-ProgressReport_20231214_v1.1.pdf
05	10/01/2024-05/03/2024	OSMOSE-PR-045-MAG-V1.0-ProgressReport_20240316_v1.pdf
06	09/04/2024-20/06/2024	OSMOSE-PR-049-MAG-V1.0-ProgressReport_20240618.pdf

10.2 Project meetings

The OSMOSE team met once per month to coordinate the activities and to share information, and progress meeting with ESA were organized regularly (see Table 10.3). These meetings took place by videoconference, expect the MTR which was hold in Toulouse at CESBIO, and the Final Review also in Toulouse at CESBIO.

Table 10.3: OSMOSE meetings and milestones table

Meeting/Milestone	Long Meeting Name	Location	Date
KOM	Kick-off Meeting	Teleconference	09-06-2022
PM1	Progress Meeting 1	Teleconference	08-09-2022
PM2	Progress Meeting 2	Teleconference	06-12-2022
PM3	Progress Meeting 3	Teleconference	13-02-2023
PM4	Progress Meeting 4	Teleconference	06-07-2023
MTR	Mid-Term Review	CESBIO Toulouse	05-10-2023
PM6	Progress Meeting 6	Teleconference	09-01-2024
PM7	Progress Meeting 7	Teleconference	09-06-2024
FR	Final Review	CESBIO, Toulouse	03-10-2024

Chapter 11

Communication, outreach, and promotion

11.1 Website

A dedicated website was set up to share information about the OSMOSE project as well as to make all the project documents available for the team and forESA. The link is the following one:
https://osmose.magellium.com/

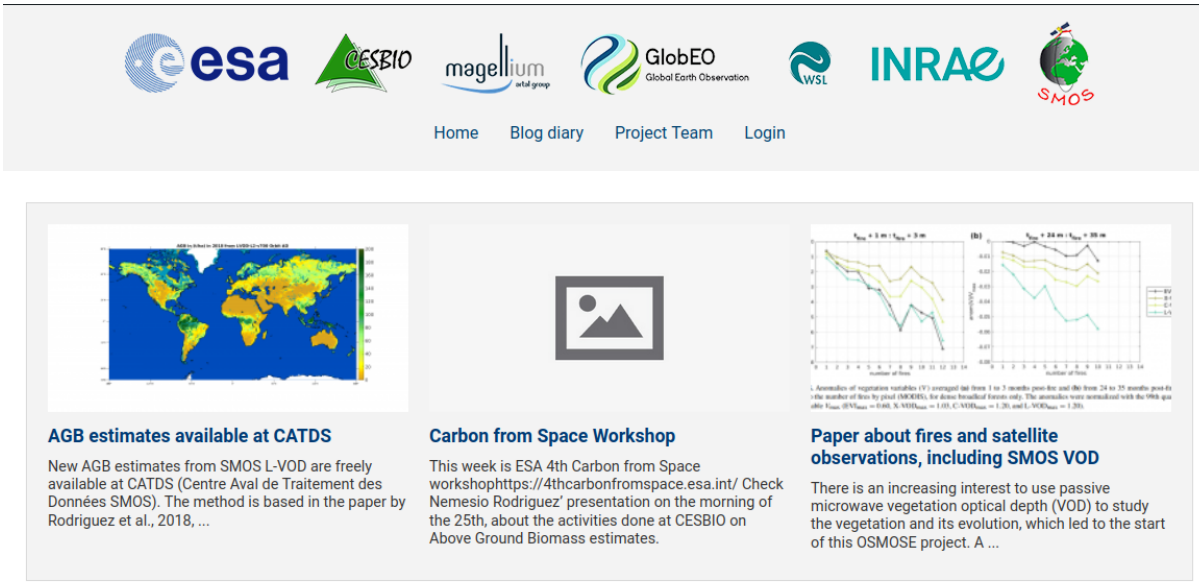


Figure 11.1: Main page of the OSMOSE webpage.

11.2 Outreach to other projects

The outputs of the OSMOSE project are shared with the ESA SMOS ESL group, and in particular within the context of the SMOS 10 years SM project, led by CESBIO, as the change request n^{o}4 ESL (Contract No. 4000130567).

11.3 Publications

The following publications were done in the context of the OSMOSE project:

- Zhou, Y. and Schwank, M. and Kurum, M. and Houtz, D. and Zhao, Q. and Lang, R. and Mialon, A. and Drusch, D., 2024. Microwave Emission Model for Layered Vegetation (MEMLV): An Exemplary Study for Coniferous Forests from P- to Ka- Band. In review TGRS, 2024. doi : https ://doi.org/XXX.
- Salazar-Neira, J. C. and A. Mialon, and P. Richaume, and S. Mermoz, and Y.H. Kerr, and A. Bouvet, and T. Le Toan, and S. Boitard, and N. J. Rodríguez-Fernández. Above-ground biomass estimation based on multi-angular l-band passive microwaves brightness temperatures. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16 :5813–5827, 2023. doi : 10.1109/JSTARS.2023.3285288

- Salazar Neira, J. C. and Mialon, A. and Le Toan, T. and Bouvet, A. and Richaume, P. and Rodríguez-Fernández, N. J., 2024. The response of Microwave Optical Depth and Optical Indices to Water Availability in Croplands, in review Remote Sensing of Environment
- Ma, H. and Weiss, M. and Malik, D. and Berthelot, B. and Yebra, M. and Nolan, R. and Mialon, A. and Zeng, J. and Quan, X. and Tagesson, H. T. and Olioso, A. and Baret, 2024. Satellite vegetation canopy water from Sentinel-2, Landsat8 and MODIS: the principle, algorithm and assessment, submitted to RSE.
- Ma, H., Weiss M., et al., & Baret, F. Blending and comparison of satellite LAI and FAPAR products from multi-sensors using diverse fusion approaches, (to be submitted to RSE)
- Boitard, S. and Mialon, A. and Mermoz, S. and Rodríguez-Fernández, N.J., and Richaume, P. and Salazar Neira, J.C. and Tarot, S. and Yann H. Kerr, 2024. Above ground biomass dataset from SMOS L band vegetation optical depth and reference maps. in review ESSD.

11.4 Conference and workshop contributions

The OSMOSE team participated to various conferences to present the activities, as shown in Table 11.1. The OSMOSE project was regularly presented at the ESA SMOS ESL (Expert Support Laboratory), i.e. PM n^{o}27 in Toulouse-France (05-October-2022), PM n^{o}28 in Roma-Italy (18-January-2023), PM n^{o}29 in Zurich-Switzerland (16-May-2023), PM n^{o}30 in Toulouse-France (03-October-2023), PM n^{o}31 in Sodankyla-Finland (24-January-2024), PM n^{o}32 in Zurich-Switzerland (15-May-2024), PM n^{o}33 in Toulouse-France (09-October-2024).

Conference/Workshop	Location	Date	Website
IGARSS	Athens, Greece	7-12 July 2024	https://www.2024.ieeeigarss.org/index.php
EGU	Vienna, Austria	14-19 April 2024	https://www.egu24.eu/
IGARSS	Pasadena, USA	16-21 July 2023	https://2023.ieeeigarss.org/
ESA PolInSar	Toulouse, France	19-23 June 2023	https://polinsar-biomass2023.esa.int/
ESA LPVE	Frascati ESRIN, Italy	ESA LPVE	https://nikal.eventsair.com/lpve-2023/
EGU 2022	Vienne, Austria	01 May 2023	https://www.egu22.eu/
ESA 4th Carbon from Space	Frascati, ESRIN, Italy	25 Nov 2022	https://4thcarbonfromspace.esa.int/
IGARSS	Kuala Lumpur, Malaysia	17-22 July 2022	https://igarss2022.org/
ESA Living Planet Symposium	Bonn, Germany	23-27 May 2022	https://lps22.esa.int

Table 11.1: Conferences and Workshops

Chapter 12

Conclusions

The ESA PM-VOS project aimed at studying the vegetation layer by remote sensing data from passive microwaves and optical/Infra red observations. It was the continuity of the ESA "SMOS+veg" project (Contract No. 4000137990/22/NL/IA) with the main goals to: i) improve AGB estimates from passive microwave sensors, such as SMOS, AMSR-E, and AMSR-2, and ii) develop harmonized datasets of microwave band-specific VOD to assess evolutions of vegetation parameters (AGB, for instance) over the last 20 years from passive microwave sensors; iii) provide a 20-year data-cube of optical indices (NDVI, LAI,FAPAR); iv) study the complementarity of all these indices to characterize and monitor state parameters of vegetation.
This project lasted two years from June 2022 to October 2024, and provided the following important outputs:

- First, a microwave emission model for layered vegetation (MEMLV) was developed to estimate the effective VOD and scattering albedo of forest canopies from P- to Ka-band (0.4 GHz – 37 GHz). The MEMLV provided a physics-based approach to assess the impact from different parameters, such as frequency, forest structure and vegetation water content, on the retrieved VOD and albedo.
- Second, leaf water content was estimated from optical data
- Third, estimations of AGB were performed from passive microwave observations, including their first assessment based on in situ measurements. AGB were derived from VOD and from a neural network approach.
- Finally, the synergy of all these indices was investigated for croplands. Passive microwave VOD were also used in synergy with vegetation indices from optical/near-infrared data, showing the complementarity of these indices to study the seasonal dynamics of the vegetation.

References

Aires, F., P. Weston, P. de Rosnay, and D. Fairbairn (2021), Statistical approaches to assimilate ascat soil moisture information—i. methodologies and first assessment, *Quarterly Journal of the Royal Meteorological Society*, **147**(736), p. 1823–1852.

Al Bitar, A., A. Mialon, Y. Kerr, F. Cabot, P. Richaume, E. Jacquette, A. Quesney, A. Mahmoodi, S. Tarot, M. Parrens, A. Al-yaari, T. Pellarin, N. Rodriguez-Fernandez, and J.-P. Wigneron (2017a), The globl smos level 3 daily soil moisture and brightness temperature maps, *Earth System Science Data*, **9**, p. 293–315, doi:10.5194/essd-2017-1.

Al Bitar, A., A. Mialon, Y. H. Kerr, F. Cabot, P. Richaume, E. Jacquette, A. Quesney, A. Mahmoodi, S. Tarot, M. Parrens, A. Al-Yaari, T. Pellarin, N. Rodriguez-Fernandez, and J.-P. Wigneron (2017b), The global SMOS Level 3 daily soil moisture and brightness temperature maps, *Earth System Science Data*, **9**(1), p. 293–315, doi:10.5194/essd-9-293-2017.

Andela, N., Y. Liu, A. Van Dijk, R. De Jeu, and T. McVicar (2013), Global changes in dryland vegetation dynamics (1988–2008) assessed by satellite remote sensing: comparing a new passive microwave vegetation density record with reflective greenness data, *Biogeosciences*, **10**(10), p. 6657–6676.

Atzberger, C. (2004), Object-based retrieval of biophysical canopy variables using artificial neural nets and radiative transfer models, *Remote Sensing of Environment*, **93**, p. 53–67.

Baccini, A., S. J. Goetz, W. S. Walker, N. T. Laporte, M. Sun, D. Sulla-Menashe, J. Hackler, P. S. A. Beck, R. Dubayah, M. A. Friedl, S. Samanta, and R. A. Houghton (2012), Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps, *Nature Climate Change*, **2**(3), p. 182–185, doi:10.1038/nclimate1354.

Bacour, C., F. Baret, D. Béal, M. Weiss, and K. Pavageau (2006), Neural network estimation of lai, fapar, fcover and lai×cab, from top of canopy meris reflectance data: Principles and validation, *Remote Sensing of Environment*, **105**, p. 313–325, doi:10.1016/j.rse.2006.07.014.

Baret, F., and M. Weiss (2024), Review of methods to estimate vegetation traits from reflectance observations, *Report*, INRAE.

Baret, F., O. Hagolle, B. Geiger, P. Bicheron, B. Miras, M. Huc, B. Berthelot, F. Nino, M. Weiss, O. Samain, J. L. Roujean, and M. Leroy (2007), Lai, fapar and fcover cyclopes global products derived from vegetation: Part 1: Principles of the algorithm, *Remote Sensing of Environment*, **110**(3), p. 275–286.

Baret, F., M. Weiss, R. Lacaze, F. Camacho, H. Makhmara, P. Pacholcyzk, and B. Smets (2013), Geov1: Lai and fapar essential climate variables and fcover global time series capitalizing over existing products. part1: Principles of development and production, *Remote Sensing of Environment*, **137**(0), p. 299–309, doi:http://dx.doi.org/10.1016/j.rse.2012.12.027.

Barrée, M., A. Mialon, T. Pellarin, M. Parrens, R. Biron, F. Lemaître, S. Gascoin, and Y. Kerr (2021), Soil moisture and vegetation optical depth retrievals over heterogeneous scenes using LEWIS L-band radiometer, *International Journal of Applied Earth Observation and Geoinformation*, **102**, p. 102,424, doi:10.1016/j.jag.2021.102424.

Beck, H. E., T. R. McVicar, N. Vergopolan, A. Berg, N. J. Lutsko, A. Dufour, Z. Zeng, X. Jiang, A. I. J. M. van Dijk, and D. G. Miralles (2023), High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections, *Scientific Data*, **10**(1), p. 724, doi:10.1038/s41597-023-02549-6.

Bousquet, E., A. Mialon, N. Rodriguez-Fernandez, C. Prigent, F. H. Wagner, and Y. H. Kerr (2021), Influence of surface water variations on VOD and biomass estimates from passive microwave sensors, *Remote Sensing of Environment*, **257**, p. 112,345, doi:10.1016/j.rse.2021.112345.

Brandt, M., J.-P. Wigneron, J. Chave, T. Tagesson, J. Penuelas, P. Ciais, K. Rasmussen, F. Tian, C. Mbow, A. Al-Yaari, N. Rodriguez, G. Schurgers, W. Zhang, J. Chang, Y. Kerr, A. Verger, C. Tucker, A. Mialon, and R. Fensholt (2018), Satellite passive microwaves reveal deforestation and drought-induced carbon losses in africa, *Nature Ecology Evolution*, **Vol. 2 Issue 5**, p. 827–835, doi:10.1038/s41559-018-0530-6.

Brocca, L., S. Hasenauer, T. Lacava, F. Melone, T. Moramarco, W. Wagner, W. Dorigo, P. Matgen, J. Martínez-Fernández, P. Llorens, J. Latron, C. Martin, and M. Bittelli (2011), Soil moisture estimation through ascat and amsr-e sensors: An intercomparison and validation study across europe, *Remote Sensing of Environment*, **115**(12), p. 3390–3408, doi:https://doi.org/10.1016/j.rse.2011.08.003.

Brodzik, M. J., and D. G. Long (2018), Nasa measures calibrated passive microwave daily ease-grid 2.0 brightness temperature esdr (cetb) algorithm theoretical basis document, revision 1.0 26 April 2018.

Brodzik, M. J., B. Billingsley, T. Haran, B. Raup, and M. H. Savoie (2012a), EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets, *ISPRS International Journal of Geo-Information*, **1**(1), p. 32–45, doi:10.3390/ijgi1010032, iSSN 2220-9964.

Brodzik, M. J., B. Billingsley, T. Haran, B. Raup, and M. H. Savoie (2012b), EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets, *ISPRS International Journal of Geo-Information*, **1**(1), p. 32–45, doi:10.3390/ijgi1010032.

Brodzik, M. J., B. Billingsley, T. Haran, B. Raup, and M. H. Savoie (2014), Correction: Brodzik, M.J., et al. EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets. ISPRS International J. of Geo-Information 2012, 1, 32–45, *ISPRS International Journal of Geo-Information*, **3**(3), p. 1154–1156, doi:10.3390/ijgi3031154, iSSN 2220-9964.

Brodzik, M. J., D. G. Long, M. A. Hardman, A. Paget, and R. Armstrong (2016), Measures calibrated enhanced-resolution passive microwave daily ease-grid 2.0 brightness temperature esdr, version 1.

Brown, L. A., R. Fernandes, N. Djamai, C. Meier, N. Gobron, H. Morris, F. Canisius, G. Bai, C. Lerebourg, and C. Lanconelli (2021), Validation of baseline and modified sentinel-2 level 2 prototype processor leaf area index retrievals over the united states, *ISPRS Journal of Photogrammetry and Remote Sensing*, **175**, p. 71–87.

Broxton, P. D., X. Zeng, D. Sulla-Menashe, and P. A. Troch (2014a), A global land cover climatology using modis data, *Journal of Applied Meteorology and Climatology*, **53**(6), p. 1593 – 1605, doi:https://doi.org/10.1175/JAMC-D-13-0270.1.

Broxton, P. D., X. Zeng, D. Sulla-Menashe, and P. A. Troch (2014b), A Global Land Cover Climatology Using MODIS Data, *Journal of Applied Meteorology and Climatology*, **53**(6), p. 1593–1605, doi:10.1175/JAMC-D-13-0270.1.

Ceccato, P., S. Flasse, S. Tarantola, S. Jacquemoud, and J.-M. Grégoire (2001), Detecting vegetation leaf water content using reflectance in the optical domain, *Remote Sensing of Environment*, **77**(1), p. 22–33, doi:https://doi.org/10.1016/S0034-4257(01)00191-2.

Chauhan, N. S., D. M. Le Vine, and R. H. Lang (1994), Discrete scatter model for microwave radar and radiometer response to corn: comparison of theory and data, *IEEE Transactions on Geoscience and Remote Sensing*, **32**(2), p. 416–426.

Combal, B., F. Baret, and M. Weiss (2002), Improving canopy variable estimation from remote sensing data by exploiting ancillary information. case study on sugar beet canopies., *Agronomie*, **22**, p. 205–215, in press.

Djamai, N., R. Fernandes, M. Weiss, H. McNairn, and K. Goïta (2019), Validation of the sentinel simplified level 2 product prototype processor (sl2p) for mapping cropland biophysical variables using sentinel-2/msi and landsat-8/oli data, *Remote Sensing of Environment*, **225**, p. 416–430, doi:https://doi.org/10.1016/j.rse.2019.03.020.

Dong, M., S. Liu, R. Jiang, J. Qi, B. de Solan, A. Comar, L. Li, W. Li, Y. Ding, and F. Baret (2024), Comparing and combining data-driven and model-driven approaches to monitor wheat green area index with high spatio-temporal resolution satellites, *Remote Sensing of Environment*, **305**, p. 114,118, doi:https://doi.org/10.1016/j.rse.2024.114118.

Estévez, J., J. Vicent, J. P. Rivera-Caicedo, P. Morcillo-Pallarés, F. Vuolo, N. Sabater, G. Camps-Valls, J. Moreno, and J. Verrelst (2020), Gaussian processes retrieval of lai from sentinel-2 top-of-atmosphere radiance data, *ISPRS Journal of Photogrammetry and Remote Sensing*, **167**, p. 289–304, doi:10.1016/j.isprsjprs.2020.07.004, export Date: 25 January 2021.

Fan, L., J.-P. Wigneron, P. Ciais, J. Chave, M. Brandt, R. Fensholt, S. S. Saatchi, A. Bastos, K. Al-Yaari, A. ad Hufkens, Y. Qin, X. Xiao, C. Chen, R. B. Myneni, R. Fernandez-Moran, A. Mialon, N. Rodriguez-Fernandez, Y. Kerr, F. Tian, and J. Peñuelas (2019), Satellite-observed pantropical carbon dynamics, *Nature plants*, **5**(9), p. 944–951.

FAO, F. (1988), Soil map of the world: revised legend, *World Soil Resources Report*, **60**.

Fernandez-Moran, R., A. Al-Yaari, A. Mialon, A. Mahmoodi, A. Al Bitar, G. De Lannoy, N. Rodriguez-Fernandez, E. Lopez-Baeza, Y. Kerr, and J.-P. Wigneron (2017), SMOS-IC: An Alternative SMOS Soil Moisture and Vegetation Optical Depth Product, *Remote Sensing*, **9**(5), doi:10.3390/rs9050457, article Number 457.

Ferrazzoli, P., and L. Guerriero (1996), Passive microwave remote sensing of forests: A model investigation, *IEEE Transactions on Geoscience and Remote Sensing*, **34**(2), p. 433–443.

Fortuny, J., and A. J. Sieber (1999), Three-dimensional synthetic aperture radar imaging of a fir tree: First results, *IEEE Transactions on Geoscience and Remote Sensing*, **37**(2), p. 1006–1014.

Fung, A. K., and K.-S. Chen (2010), *Microwave scattering and emission models for users*, Artech house.

Fuster, B., J. Sánchez-Zapero, F. Camacho, V. García-Santos, A. Verger, R. Lacaze, M. Weiss, F. Baret, and B. Smets (2020), Quality assessment of proba-v lai, fapar and fcover collection 300 m products of copernicus global land service, *Remote Sensing*, **12**(6), doi:10.3390/rs12061017.

Garcia-Haro, F. J., M. Campos-Taberner, A. Moreno, H. T. Tagesson, F. Camacho, B. Martínez, S. Sánchez, M. Piles, G. Camps-Valls, and M. Yebra (2020), A global canopy water content product from avhrr/metop, *ISPRS Journal of Photogrammetry and Remote Sensing*, **162**, p. 77–93.

Garonna, I., R. De Jong, A. J. De Wit, C. A. Mücher, B. Schmid, and M. E. Schaepman (2014), Strong contribution of autumn phenology to changes in satellite-derived growing season length estimates across Europe (1982–2011), *Global Change Biology*, **20**(11), p. 3457–3470, doi:10.1111/gcb.12625.

Goel, N. (1989), *Inversion of canopy reflectance models for estimation of biophysical parameters from reflectance data*, p. 205–251, asrar, g. ed., Wiley Interscience.

Grant, J. P., J.-P. Wigneron, R. De Jeu, H. Lawrence, A. Mialon, P. Richaume, A. Al Bitar, M. Drusch, M. Van Marle, and Y. Kerr (2016), Comparison of smos and amsr-e vegetation optical depth to four modis-based vegetation indices, *Remote Sensing of Environment*, **172**, p. 87–100, doi:http://dx.doi.org/10.1016/j.rse.2015.10.021.

Griebel, A., M. M. Boer, C. Blackman, B. Choat, D. S. Ellsworth, P. Madden, B. Medlyn, V. Resco de Dios, A. Wujeska-Klause, and M. Yebra (2023), Specific leaf area and vapour pressure deficit control live fuel moisture content, *Functional Ecology*, **37**(3), p. 719–731.

Guan, K., J. Wu, J. S. Kimball, M. C. Anderson, S. Frolking, B. Li, C. R. Hain, and D. B. Lobell (2017), The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields, *Remote Sensing of Environment*, **199**, p. 333–349, doi:10.1016/j.rse.2017.06.043.

Guglielmetti, M., M. Schwank, C. Mätzler, C. Oberdörster, J. Vanderborght, and H. Flühler (2007), Measured microwave radiative transfer properties of a deciduous forest canopy, *Remote Sensing of Environment*, **109**(4), p. 523–532, doi:https://doi.org/10.1016/j.rse.2007.02.003.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, et al. (2013), High-resolution global maps of 21st-century forest cover change, *science*, **342**(6160), p. 850–853.

Hengl, T., J. Sanderman, and L. Parente (2021), Open soil spectral library (training data and calibration models), doi:https://zenodo.org/record/5805138.

Hersbach, H., B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut (2023), Era5 hourly data on single levels from 1940 to present. copernicus climate change service (c3s) climate data store (cds), *Tech. rep.*, ECMWF, doi:10.24381/cds.adbb2d47, accessed on 2022.

Hornbuckle, B. (2023), Vegetation water content... or vegetation canopy water?, in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, p. 3163–3166, doi:10.1109/IGARSS52108.2023.10283364.

Hornbuckle, B. K., A. W. England, R. D. De Roo, M. A. Fischman, and D. L. Boprie (2003), Vegetation canopy anisotropy at 1.4 ghz, *IEEE Transactions on Geoscience and Remote Sensing*, **41**(10), p. 2211–2223.

Hornbuckle, B. K., J. C. Patton, A. VanLoocke, A. E. Suyker, M. C. Roby, V. A. Walker, E. R. Iyer, D. E. Herzmann, and E. A. Endacott (2016), SMOS optical thickness changes in response to the growth and development of crops, crop management, and weather, *Remote Sensing of Environment*, **180**, p. 320–333, doi:10.1016/j.rse.2016.02.043.

Huete, A., K. Didan, T. Miura, E. Rodriguez, X. Gao, and L. Ferreira (2002), Overview of the radiometric and biophysical performance of the MODIS vegetation indices, *Remote Sensing of Environment*, **83**(1-2), p. 195–213, doi:10.1016/S0034-4257(02)00096-2.

Huxman, T. E., K. A. Snyder, D. Tissue, A. J. Leffler, K. Ogle, W. T. Pockman, D. R. Sandquist, D. L. Potts, and S. Schwinning (2004), Precipitation pulses and carbon fluxes in semiarid and arid ecosystems, *Oecologia*, **141**(2), p. 254–268, doi:10.1007/s00442-004-1682-4.

Jackson, T., and T. Schmugge (1991), Vegetation effects on the microwave emission of soils, *Remote Sensing of Environment*, **36**(3), p. 203–212.

Jackson, T., T. Schmugge, and J. Wang (1982), Passive microwave sensing of soil moisture under vegetation canopies, *Water Resour. Res.*, **18**, p. 1137–1142.

Jacquemoud, S., W. Verhoef, F. Baret, C. Bacour, P. J. Zarco-Tejada, G. P. Asner, C. François, and S. L. Ustin (2009), Prospect+sail models: A review of use for vegetation characterization, *Remote Sensing of Environment*, **113**, p. S56–S66, doi:https://doi.org/10.1016/j.rse.2008.01.026.

Jarque, C. M., and A. K. Bera (1987), A test for normality of observations and regression residuals, *International Statistical Review/Revue Internationale de Statistique*, p. 163–172.

Jones, M. O., L. A. Jones, J. S. Kimball, and K. C. McDonald (2011), Satellite passive microwave remote sensing for monitoring global land surface phenology, *Remote Sensing of Environment*, **115**(4), p. 1102–1114, doi:10.1016/j.rse.2010.12.015.

Jones, M. O., J. S. Kimball, L. A. Jones, and K. C. McDonald (2012), Satellite passive microwave detection of North America start of season, *Remote Sensing of Environment*, **123**, p. 324–333, doi:10.1016/j.rse.2012.03.025.

Karkauskaite, P., T. Tagesson, and R. Fensholt (2017), Evaluation of the Plant Phenology Index (PPI), NDVI and EVI for Start-of-Season Trend Analysis of the Northern Hemisphere Boreal Zone, *Remote Sensing*, **9**(5), p. 485, doi:10.3390/rs9050485.

KATTGE, J., S. DÍAZ, S. LAVOREL, I. C. PRENTICE, P. LEADLEY, G. BÖNISCH, E. GARNIER, M. WESTOBY, P. B. REICH, I. J. WRIGHT, J. H. C. CORNELISSEN, C. VIOLLE, S. P. HARRISON, P. M. Van BODEGOM, M. REICHSTEIN, B. J. ENQUIST, N. A. SOUDZILOVSKAIA, D. D. ACKERLY, M. ANAND, O. ATKIN, M. BAHN, T. R. BAKER, D. BALDOCCHI, R. BEKKER, C. C. BLANCO, B. BLONDER, W. J. BOND, R. BRADSTOCK, D. E. BUNKER, F. CASANOVES, J. CAVENDER-BARES, J. Q. CHAMBERS, F. S. CHAPIN III, J. CHAVE, D. COOMES, W. K. CORNWELL, J. M. CRAINE, B. H. DOBRIN, L. DUARTE, W. DURKA, J. ELSER, G. ESSER, M. ESTIARTE, W. F. FAGAN, J. FANG, F. FERNÁNDEZ-MÉNDEZ, A. FIDELIS, B. FINEGAN, O. FLORES, H. FORD, D. FRANK, G. T. FRESCHET, N. M. FYLLAS, R. V. GALLAGHER, W. A. GREEN, A. G. GUTIERREZ, T. HICKLER, S. I. HIGGINS, J. G. HODGSON, A. JALILI, S. JANSEN, C. A. JOLY, A. J. KERKHOFF, D. KIRKUP, K. KITAJIMA, M. KLEYER, S. KLOTZ, J. M. H. KNOPS, K. KRAMER, I. KÜHN, H. KUROKAWA, D. LAUGHLIN, T. D. LEE, M. LEISHMAN, F. LENS, T. LENZ, S. L. LEWIS, J. LLOYD, J. LLUSIÀ, F. LOUAULT, S. MA, M. D. MAHECHA, P. MANNING, T. MASSAD, B. E. MEDLYN, J. MESSIER, A. T. MOLES, S. C. MÜLLER, K. NADROWSKI, S. NAEEM, U. NIINEMETS, S. NÖLLERT, A. NÜSKE, R. OGAYA, J. OLEKSYN, V. G. ONIPCHENKO, Y. ONODA, J. ORDOÑEZ, G. OVERBECK, W. A. OZINGA, S. PATIÑO, S. PAULA, J. G. PAUSAS, J. PEÑUELAS, O. L. PHILLIPS, V. PILLAR, H. POORTER, L. POORTER, P. POSCHLOD, A. PRINZING, R. PROULX, A. RAMMIG, S. REINSCH, B. REU, L. SACK, B. SALGADO-NEGRET, J. SARDANS, S. SHIODERA, B. SHIPLEY, A. SIEFERT, J.-F. SOSINSKI, E. SOUSSANA, E. SWAINE, N. SWENSON, K. THOMPSON, P. THORNTON, M. WALDRAM, E. WEIHER, M. WHITE, S. WHITE, S. J. WRIGHT, B. YGUEL, S. ZAEHLE, A. E. ZANNE, and C. WIRTH (2011), Try – a global database of plant traits, *Global Change Biology*, **17**(9), p. 2905–2935, doi:https://doi.org/10.1111/j.1365-2486.2011.02451.x.

Kattge, J., G. Bönisch, S. Díaz, S. Lavorel, I. C. Prentice, and et al (2020), Try plant trait database – enhanced coverage and open access, *Global Change Biology*, **26**(1), p. 119–188, doi:https://doi.org/10.1111/gcb.14904.

Kerr, Y., P. Richaume, P. Waldteufel, P. Ferrazzoli, J. Wigneron, M. Schwank, and K. Rautiainen (2020), Algorithm theoretical basis document (ATBD) for the smos level 2 soil moisture processor, *Technical Report TN-ESL-SM-GS-0001-4b SM-ESL (CBSA)*, p. 145p.

Kerr, Y. H., P. Waldteufel, J.-P. Wigneron, S. Delwart, F. Cabot, J. Boutin, M.-J. Escorihuela, J. Font, N. Reul, C. Gruhier, et al. (2010), The SMOS mission: New tool for monitoring key elements of the global water cycle, *Proceedings of the IEEE*, **98**(5), p. 666–687.

Kerr, Y. H., P. Waldteufel, P. Richaume, J. P. Wigneron, P. Ferrazzoli, A. Mahmoodi, A. Al Bitar, F. Cabot, C. Gruhier, S. E. Juglea, D. Leroux, A. Mialon, and S. Delwart (2012a), The SMOS Soil Moisture Retrieval Algorithm, *Geoscience and Remote Sensing, IEEE Transactions on*, **50**(5), p. 1384 –1403, doi:10.1109/TGRS.2012.2184548.

Kerr, Y. H., P. Waldteufel, P. Richaume, J. P. Wigneron, P. Ferrazzoli, A. Mahmoodi, A. Al Bitar, F. Cabot, C. Gruhier, S. E. Juglea, D. Leroux, A. Mialon, and S. Delwart (2012b), The SMOS Soil Moisture Retrieval Algorithm, *IEEE Transactions on Geoscience and Remote Sensing*, **50**(5), p. 1384–1403, doi:10.1109/TGRS.2012.2184548.

Kirdiashev, K. P., A. A. Chukhlantsev, and A. M. Shutko (1979), Microwave radiation of the earth's surface in the presence of vegetation cover, *Radiotekhnika i Elektronika*, **24**, p. 256–264.

Knyazikhin, Y., J. V. Martonchik, R. B. Myneni, D. J. Diner, and S. W. Running (1998), Synergistic algorithm for estimating vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from modis and misr data, *J. Geophys. Res.*, **103**(D24), p. 32,257–32,275, doi:10.1029/98jd02462.

Kramer, P. (2012), *Physiology of woody plants*, Elsevier.

Kurum, M. (2013), Quantifying scattering albedo in microwave emission of vegetated terrain, *Remote Sensing of Environment*, **129**, p. 66–74.

Kurum, M., P. E. O'Neill, R. H. Lang, A. T. Joseph, M. H. Cosh, and T. J. Jackson (2012), Effective tree scattering and opacity at L-band, *Remote Sensing of Environment*, **118**, p. 1–9.

Landry, R., R. Fournier, F. Ahern, and R. Lang (1997), Tree vectorization: a methodology to characterize fine tree architecture in support of remote sensing models, *Canadian journal of remote sensing*, **23**(2), p. 91–107.

Lang, R., R. Landry, A. Franchois, G. Nesti, and A. Sieber (1998), Microwave tree scattering experiment: comparison of theory and experiment, in *1998 IEEE International Geoscience and Remote Sensing. Symposium Proceedings*, vol. 5, p. 2384–2386, IEEE.

Lang, R. H., and J. S. Sidhu (1983), Electromagnetic backscattering from a layer of vegetation: A discrete approach, *IEEE Transactions on Geoscience and Remote Sensing*, **GE-21**(1), p. 62–71.

Laurent, V. C., W. Verhoef, A. Damm, M. E. Schaepman, and J. G. Clevers (2013), A bayesian object-based approach for estimating vegetation biophysical and biochemical variables from apex at-sensor radiance data, *Remote Sensing of Environment*, **139**, p. 6–17.

Lauvernet, C., F. Baret, L. Hascoët, S. Buis, and F.-X. Le Dimet (2008), Multitemporal-patch ensemble inversion of coupled surface–atmosphere radiative transfer models for land surface characterization, *Remote Sensing of Environment*, **112**(3), p. 851–861, doi:http://dx.doi.org/10.1016/j.rse.2007.06.027.

Lawrence, H., J.-P. Wigneron, P. Richaume, N. Novello, J. Grant, A. Mialon, A. Al Bitar, O. Merlin, D. Guyon, D. Leroux, S. Bircher, and Y. Kerr (2014a), Comparison between smos vegetation optical depth products and modis vegetation indices over crop zones of the usa, *Remote Sensing of Environment*, **140**, p. 396–406, doi:https://doi.org/10.1016/j.rse.2013.07.021.

Lawrence, H., J.-P. Wigneron, P. Richaume, N. Novello, J. Grant, A. Mialon, A. Al Bitar, O. Merlin, D. Guyon, D. Leroux, S. Bircher, and Y. Kerr (2014b), Comparison between SMOS Vegetation Optical Depth products and MODIS vegetation indices over crop zones of the USA, *Remote Sensing of Environment*, **140**, p. 396–406, doi:10.1016/j.rse.2013.07.021.

Li, C., L. Zhou, and W. Xu (2021), Estimating Aboveground Biomass Using Sentinel-2 MSI Data and Ensemble Algorithms for Grassland in the Shengjin Lake Wetland, China, *Remote Sensing*, **13**(8), p. 1595, doi:10.3390/rs13081595.

Li, X., A. Al-Yaari, M. Schwank, L. Fan, F. Frappart, J. Swenson, and J.-P. Wigneron (2020), Compared performances of SMOS-IC soil moisture and vegetation optical depth retrievals based on Tau-Omega and Two-Stream microwave emission models, *Remote Sensing of Environment*, **236**, p. 111,502.

Li, X., J.-P. Wigneron, L. Fan, F. Frappart, S. H. Yueh, A. Colliander, A. Ebtehaj, L. Gao, R. Fernandez-Moran, X. Liu, M. Wang, H. Ma, C. Moisy, and P. Ciais (2022), A new SMAP soil moisture and vegetation optical depth product (SMAP-IB): Algorithm, assessment and inter-comparison, *Remote Sensing of Environment*, **271**, p. 112,921, doi:https://doi.org/10.1016/j.rse.2022.112921.

Liang, S. (2005), *Quantitative remote sensing of land surfaces*, John Wiley & Sons.

Lilliefors, H. W. (1967), On the kolmogorov-smirnov test for normality with mean and variance unknown, *Journal of the American statistical Association*, **62**(318), p. 399–402.

Liu, X., Y. Chen, Z. Li, and Y. Li (2023), Evaluating the Consistency of Vegetation Phenological Parameters in the Northern Hemisphere from 1982 to 2015, *Remote Sensing*, **15**(10), p. 2559, doi:10.3390/rs15102559.

Liu, Y., R. A. de Jeu, A. I. van Dijk, and M. Owe (2007), Trmm-tmi satellite observed soil moisture and vegetation density (1998–2005) show strong connection with el niño in eastern australia, *Geophysical Research Letters*, **34**(15).

Liu, Y., N. M. Holtzman, and A. G. Konings (2021), Global ecosystem-scale plant hydraulic traits retrieved using model–data fusion, *Hydrology and Earth System Sciences*, **25**(5), p. 2399–2417, doi:10.5194/hess-25-2399-2021.

Liu, Y. Y., R. A. De Jeu, M. F. McCabe, J. P. Evans, and A. I. Van Dijk (2011), Global long-term passive microwave satellite-based retrievals of vegetation optical depth, *Geophysical Research Letters*, **38**(18).

Liu, Y. Y., A. I. Van Dijk, R. A. De Jeu, J. G. Canadell, M. F. McCabe, J. P. Evans, and G. Wang (2015), Recent reversal in loss of global terrestrial biomass, *Nature Climate Change*, **5**(5), p. 470–474.

Loveland, T., and A. Belward (1997), The International Geosphere Biosphere Programme Data and Information System global land cover data set (DISCover), *Acta Astronautica*, **41**(4-10), p. 681–689, doi:10.1016/S0094-5765(98)00050-2.

Loveland, T. R., B. C. Reed, J. F. Brown, D. O. Ohlen, Z. Zhu, L. Yang, and J. W. Merchant (2000), Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data, *International Journal of Remote Sensing*, **21**(6-7), p. 1303–1330, doi:10.1080/014311600210191.

Madelon, R., N. J. Rodríguez-Fernández, R. van der Schalie, T. Scanlon, A. Al Bitar, Y. H. Kerr, R. de Jeu, and W. Dorigo (2022), Toward the removal of model dependency in soil moisture climate data records by using an l-band scaling reference, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **15**, p. 831–848, doi:10.1109/JSTARS.2021.3137008.

Maeda, T., Y. Taniguchi, and K. Imaoka (2016), Gcom-w1 amsr2 level 1r product: Dataset of brightness temperature modified using the antenna pattern matching technique, *IEEE Transactions on Geoscience and Remote Sensing*, **54**(2), p. 770–782, doi:10.1109/TGRS.2015.2465170.

Matzler, C. (1994), Microwave (1-100 GHz) dielectric model of leaves, *IEEE Transactions on Geoscience and Remote Sensing*, **32**(4), p. 947–949.

Mätzler, C. (1998), Improved born approximation for scattering of radiation in a granular medium, *Journal of Applied Physics*, **83**(11), p. 6111–6117.

McMahon, T. A., and R. E. Kronauer (1976), Tree structures: deducing the principle of mechanical design, *Journal of theoretical biology*, **59**(2), p. 443–466.

Meissner, T., F. J. Wentz, and L. Ricciardulli (2014), The emission and scattering of L-band microwave radiation from rough ocean surfaces and wind speed measurements from the uppercaseAquarius sensor, *Journal of Geophysical Research: Oceans*, **119**(9), p. 6499–6522.

Mialon, A., N. J. Rodríguez-Fernández, M. Santoro, S. Saatchi, S. Mermoz, E. Bousquet, and Y. H. Kerr (2020), Evaluation of the sensitivity of smos l-vod to forest above-ground biomass at global scale, *Remote Sensing*, **12**(9), p. 1450.

Mironov, V. L., L. G. Kosolapova, and S. V. Fomin (2009), Physically and mineralogically based spectroscopic dielectric model for moist soils, *IEEE Transactions on Geoscience and Remote Sensing*, **47**(7), p. 2059–2070.

Mo, T., B. Choudhury, T. Schmugge, J. R. Wang, and T. Jackson (1982), A model for microwave emission from vegetation-covered fields, *Journal of Geophysical Research: Oceans*, **87**(C13), p. 11,229–11,237.

Moesinger, L., W. Dorigo, R. De Jeu, R. Van Der Schalie, T. Scanlon, I. Teubner, and M. Forkel (2020a), The global long-term microwave vegetation optical depth climate archive (vodca), *Earth System Science Data*, **12**(1), p. 177–196.

Moesinger, L., W. Dorigo, R. de Jeu, R. van der Schalie, T. Scanlon, I. Teubner, and M. Forkel (2020b), The global long-term microwave Vegetation Optical Depth Climate Archive (VODCA), *Earth System Science Data*, **12**(1), p. 177–196, doi:10.5194/essd-12-177-2020.

Montpetit, B., A. Royer, J.-P. Wigneron, A. Chanzy, and A. Mialon (2015), Evaluation of multi-frequency bare soil microwave reflectivity models, *Remote Sensing of Environment*, **162**, p. 186 – 195, doi:https://doi.org/10.1016/j.rse.2015.02.015.

Myneni, R., V. Gutschick, G. Asrar, and E. Kanemasu (1988), Photon transport in vegetation canopies with anisotropic scattering part ii. discrete-ordinates/exact-kernel technique for one-angle photon transport in slab geometry., *Agricultural and Forest Meteorology*, **42**, p. 17–40.

Nguyen, T. H., M. Langensiepen, T. Gaiser, H. Webber, H. Ahrends, H. Hueging, and F. Ewert (2022), Responses of winter wheat and maize to varying soil moisture: From leaf to canopy, *Agricultural and Forest Meteorology*, **314**, p. 108,803, doi:10.1016/j.agrformet.2021.108803.

Owe, M., R. De Jeu, and T. Holmes (2008), Multisensor historical climatology of satellite-derived global land surface moisture, *Journal of Geophysical Research: Earth Surface*, **113**(F1), p. 2007JF000,769, doi:10.1029/2007JF000769.

Patton, J., and B. Hornbuckle (2013), Initial Validation of SMOS Vegetation Optical Thickness in Iowa, *IEEE Geoscience and Remote Sensing Letters*, **10**(4), p. 647–651, doi:10.1109/LGRS.2012.2216498.

Pellarin, T., J.-P. Wigneron, J.-C. Calvet, M. Berger, H. Douville, P. Ferrazzoli, Y. H. Kerr, E. Lopez-Baeza, J. Pulliainen, L. P. Simmonds, et al. (2003), Two-year global simulation of L-band brightness temperatures over land, *IEEE Transactions on Geoscience and Remote Sensing*, **41**(9), p. 2135–2139.

Ploton, P., F. Mortier, N. Barbier, G. Cornu, M. Réjou-Méchain, V. Rossi, A. Alonso, J.-F. Bastin, N. Bayol, F. Bénédet, et al. (2020a), A map of african humid tropical forest aboveground biomass derived from management inventories, *Scientific Data*, **7**(1), p. 221.

Ploton, P., F. Mortier, M. Réjou-Méchain, N. Barbier, N. Picard, V. Rossi, C. Dormann, G. Cornu, G. Viennois, N. Bayol, A. Lyapustin, S. Gourlet-Fleury, and R. Pélissier (2020b), Spatial validation reveals poor predictive performance of large-scale ecological mapping models, *Nature Communications*, **11**(1), p. 4540, doi:10.1038/s41467-020-18321-y.

Qu, Y., J. Wang, H. Wan, X. Li, and G. Zhou (2008), A bayesian network algorithm for retrieving the characterization of land surface vegetation, *Remote Sensing of Environment*, **112**(3), p. 613–622, doi:http://dx.doi.org/10.1016/j.rse.2007.03.031.

Quan, X., B. He, M. Yebra, C. Yin, Z. Liao, X. Zhang, and X. Li (2017), A radiative transfer model-based method for the estimation of grassland aboveground biomass, *International Journal of Applied Earth Observation and Geoinformation*, **54**, p. 159–168, doi:https://doi.org/10.1016/j.jag.2016.10.002.

Rahmoune, R., P. Ferrazzoli, Y. K. Singh, Y. H. Kerr, P. Richaume, and A. Al Bitar (2014), Smos retrieval results over forests: Comparisons with independent measurements, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **7**(9), p. 3858–3866.

Robitaille, P.-M. (2007), On the Earth microwave background: absorption and scattering by the atmosphere, *Progress in Physics*, **3**(3), p. 3–4.

Rodríguez-Fernández, N. J., A. Mialon, S. Mermoz, A. Bouvet, P. Richaume, A. Al Bitar, A. Al-Yaari, M. Brandt, T. Kaminski, T. Le Toan, Y. H. Kerr, and J.-P. Wigneron (2018), An evaluation of smos l-band vegetation optical depth (l-vod) data sets: high sensitivity of l-vod to above-ground biomass in africa, *Biogeosciences*, **15**(14), p. 4627–4645, doi:10.5194/bg-15-4627-2018.

Roy, D. P., J. Ju, P. Lewis, C. Schaaf, F. Gao, M. Hansen, and E. Lindquist (2008), Multi-temporal modis–landsat data fusion for relative radiometric normalization, gap filling, and prediction of landsat data, *Remote Sensing of Environment*, **112**(6), p. 3112–3130, doi:https://doi.org/10.1016/j.rse.2008.03.009.

Saatchi, S. S., N. L. Harris, S. Brown, M. Lefsky, E. T. A. Mitchard, W. Salas, B. R. Zutta, W. Buermann, S. L. Lewis, S. Hagen, S. Petrova, L. White, M. Silman, and A. Morel (2011), Benchmark map of forest carbon stocks in tropical regions across three continents, *Proceedings of the National Academy of Sciences*, **108**(24), p. 9899–9904, doi:10.1073/pnas.1019576108.

Santi, E., S. Paloscia, P. Pampaloni, and S. Pettinato (2009), Ground-based microwave investigations of forest plots in Italy, *IEEE Transactions on Geoscience and Remote Sensing*, **47**(9), p. 3016–3025, doi:10.1109/TGRS.2009.2021613.

Santoro, M., and O. Cartus (2023), Esa biomass climate change initiative (biomass_cci): Global datasets of forest above-ground biomass for the years 2010, 2017, 2018, 2019 and 2020, v4, *NERC EDS Centre for Environmental Data Analysis*, doi:10.5285/af60720c1e404a9e9d2c145d2b2ead4e.

Santoro, M., R. Lucas, H. Kay, and S. Quegan (2021), CCI BIOMASS PRODUCT USER GUIDE v3.

Schwank, M., and R. Naderpour (2018), Snow density and ground permittivity retrieved from l-band radiometry: Melting effects, *Remote Sensing*, **10**(2), p. 354.

Schwank, M., C. Matzler, M. Guglielmetti, and H. Fluhler (2005), L-band radiometer measurements of soil water under growing clover grass, *IEEE Transactions on Geoscience and Remote Sensing*, **43**(10), p. 2225–2237.

Schwank, M., R. Naderpour, and C. Mätzler (2018), Tau-omega and Two-Stream emission models used for passive L-band retrievals: Application to close-range measurements over a forest, *Remote Sensing*, **10**(12), p. 1868.

Schwank, M., A. Kontu, A. Mialon, R. Naderpour, D. Houtz, J. Lemmetyinen, K. Rautiainen, Q. Li, P. Richaume, Y. Kerr, et al. (2021), Temperature effects on L-band vegetation optical depth of a boreal forest, *Remote Sensing of Environment*, **263**, p. 112,542.

Seker, S., and A. Schneider (1988), Electromagnetic scattering from a dielectric cylinder of finite length, *IEEE Transactions on antennas and propagation*, **36**(2), p. 303–307.

Shabanov, N., D. Huang, W. Yang, B. Tan, Y. Knyazikhin, R. Myneni, D. Ahl, S. Gower, and A. Huete (2005), Analysis and optimization of the modis leaf area index algorithm retrievals over broadleaf forests, *IEEE Transactions on Geoscience and Remote Sensing*, **43**(8), p. 1855–1865.

Skakun, S., J. Wevers, C. Brockmann, G. Doxani, M. Aleksandrov, M. Batič, D. Frantz, F. Gascon, L. Gómez-Chova, O. Hagolle, D. López-Puigdollers, J. Louis, M. Lubej, G. Mateo-García, J. Osman, D. Peressutti, B. Pflug, J. Puc, R. Richter, J.-C. Roger, P. Scaramuzza, E. Vermote, N. Vesel, A. Zupanc, and L. Žust (2022), Cloud mask intercomparison exercise (cmix): An evaluation of cloud masking algorithms for landsat 8 and sentinel-2, *Remote Sensing of Environment*, **274**, p. 112,990, doi:https://doi.org/10.1016/j.rse.2022.112990.

Sulla-Menashe, D., and M. A. Friedl (), User Guide to Collection 6 MODIS Land Cover (MCD12Q1 and MCD12C1) Product.

Tagesson, T., R. Fensholt, I. Guiro, M. O. Rasmussen, S. Huber, C. Mbow, M. Garcia, S. Horion, I. Sandholt, and B. Holm-Rasmussen (2015), Ecosystem properties of semiarid savanna grassland in west africa and its relationship with environmental variability, *Global change biology*, **21**(1), p. 250–264.

Teluguntla, P., P. Thenkabail, J. Xiong, M. Gumma, C. Giri, C. Milesi, M. Ozdogan, R. Congalton, J. Tilton, T. Sankey, R. Massey, A. Phalke, and K. Yadav (2016), NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security Support Analysis Data (GFSAD) Crop Mask 2010 Global 1 km V001, doi:10.5067/MEASURES/GFSAD/GFSAD1KCM.001.

Teubner, I. E., M. Forkel, M. Jung, Y. Y. Liu, D. G. Miralles, R. Parinussa, R. Van Der Schalie, M. Vreugdenhil, C. R. Schwalm, G. Tramontana, G. Camps-Valls, and W. A. Dorigo (2018), Assessing the relationship between microwave vegetation optical depth and gross primary production, *International Journal of Applied Earth Observation and Geoinformation*, **65**, p. 79–91, doi:10.1016/j.jag.2017.10.006.

Tian, F., J.-P. Wigneron, P. Ciais, J. Chave, J. Ogée, J. Peñuelas, A. Ræbild, J.-C. Domec, X. Tong, M. Brandt, A. Mialon, N. Rodriguez-Fernandez, T. Tagesson, A. Al-Yaari, Y. Kerr, C. Chen, R. B. Myneni, W. Zhang, J. Ardö, and R. Fensholt (2018), Coupling of ecosystem-scale plant water storage and leaf phenology observed by satellite, *Nature Ecology & Evolution*, **2**(9), p. 1428–1435, doi:10.1038/s41559-018-0630-3.

Tong, X., F. Tian, M. Brandt, Y. Liu, W. Zhang, and R. Fensholt (2019), Trends of land surface phenology derived from passive microwave and optical remote sensing systems and associated drivers across the dry tropics 1992–2012, *Remote Sensing of Environment*, **232**, p. 111,307, doi:10.1016/j.rse.2019.111307.

Tucker, C. J. (1979), Red and photographic infrared linear combinations for monitoring vegetation, *Remote Sensing of Environment*, **8**(2), p. 127–150, doi:10.1016/0034-4257(79)90013-0.

Ulaby, F. T., and E. A. Wilson (1985), Microwave attenuation properties of vegetation canopies, *IEEE Transactions on Geoscience and Remote Sensing*, **GE-23**(5), p. 746–753.

Van de Griend, A. A., and J.-P. Wigneron (2004), The b-factor as a function of frequency and canopy type at h-polarization, *IEEE Transactions on Geoscience and Remote Sensing*, **42**(4), p. 786–794.

Van Marle, M., G. Van Der Werf, R. De Jeu, and Y. Liu (2016), Annual south american forest loss estimates based on passive microwave remote sensing (1990–2010), *Biogeosciences*, **13**(2), p. 609–624.

Verger, A., F. Camacho, F. García-Haro, and J. Meliá (2009), Prototyping of land-saf leaf area index algorithm with vegetation and modis data over europe, *Remote Sensing of Environment*, **113**(11), p. 2285–2297.

Verger, A., F. Baret, and F. Camacho de Coca (2011), Optimal modalities for radiative transfer-neural network estimation of canopy biophysical characteristics: evaluation over an agricultural area with chris/proba observations, *Remote Sensing of Environment*, **115**, p. 415–426.

Vinué, D., F. Camacho, and B. Fuster (2018), *Validation of Sentinel-2 LAI and FAPAR products derived from SNAP toolbox over a cropland site in Barrax and over an agroforested site in Liria (Spain)*, vol. 248, Fifth recent advances in quantitative remote sensing.

Vittucci, C., P. Ferrazzoli, Y. Kerr, P. Richaume, L. Guerriero, R. Rahmoune, and G. V. Laurin (2016), Smos retrieval over forests: Exploitation of optical depth and tests of soil moisture estimates, *Remote sensing of environment*, **180**, p. 115–127.

Vreugdenhil, M., S. Hahn, T. Melzer, B. BauerMarschallinger, C. Reimer, W. A. Dorigo, and W. Wagner (2017), Assessing Vegetation Dynamics Over Mainland Australia With Metop ASCAT, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **10**(5), p. 2240–2248, doi:10.1109/JSTARS.2016.2618838.

Waldteufel, P., J.-L. Vergely, and C. Cot (2004), A modified cardioid model for processing multiangular radiometric observations, *IEEE Transactions on Geoscience and Remote Sensing*, **42**(5), p. 1059–1063, doi:10.1109/TGRS.2003.821698.

Walthall, C., W. Dulaney, M. Anderson, J. Norman, H. Fang, and S. Liang (2004), A comparison of empirical and neural network approaches for estimating corn and soybean leaf area index from landsat etm+ imagery, *Remote Sensing of Environment*, **92**(4), p. 465–474, doi:http://dx.doi.org/10.1016/j.rse.2004.06.003.

Wang, J., and B. Choudhury (1981), Remote sensing of soil moisture content, over bare field at 1.4 GHz frequency, *Journal of Geophysical Research: Oceans*, **86**(C6), p. 5277–5282.

Wang, J., R. Lopez-Lozano, M. Weiss, S. Buis, W. Li, S. Liu, F. Baret, and J. Zhang (2022), Crop specific inversion of prosail to retrieve green area index (gai) from several decametric satellites using a bayesian framework, *Remote Sensing of Environment*, **278**, p. 113,085.

Wang, M., J.-P. Wigneron, R. Sun, L. Fan, F. Frappart, S. Tao, L. Chai, X. Li, X. Liu, H. Ma, et al. (2021a), A consistent record of vegetation optical depth retrieved from the amsr-e and amsr2 x-band observations, *International Journal of Applied Earth Observation and Geoinformation*, **105**, p. 102,609.

Wang, M., L. Fan, F. Frappart, P. Ciais, R. Sun, Y. Liu, X. Li, X. Liu, C. Moisy, and J.-P. Wigneron (2021b), An alternative amsr2 vegetation optical depth for monitoring vegetation at large scales, *Remote Sensing of Environment*, **263**, p. 112,556.

Wang, M., P. Ciais, R. Fensholt, M. Brandt, S. Tao, W. Li, L. Fan, F. Frappart, R. Sun, X. Li, X. Liu, H. Wang, T. Cui, Z. Xing, Z. Zhao, and J.-P. Wigneron (2024), Satellite observed aboveground carbon dynamics in africa during 2003–2021, *Remote Sensing of Environment*, **301**, p. 113,927, doi:https://doi.org/10.1016/j.rse.2023.113927.

Wang, P., X. Song, D. Han, Y. Zhang, and B. Zhang (2012), Determination of evaporation, transpiration and deep percolation of summer corn and winter wheat after irrigation, *Agricultural Water Management*, **105**, p. 32–37, doi:10.1016/j.agwat.2011.12.024.

Weidong, L., F. Baret, G. Xingfa, T. Qingxi, Z. Lanfen, and Z. Bing (2002), Relating soil surface moisture to reflectance, *Remote Sensing of Environment*, **81**(2-3), p. 238–246, tY - JOUR.

Weiss, M., F. Baret, M. Leroy, O. Hautecœur, C. Bacour, L. Prévot, and N. Bruguier (2002), Validation of neural net techniques to estimate canopy biophysical variables from remote sensing data, *Agronomie*, **22**(6), p. 547–553.

Weiss, M., F. Baret, S. Garrigues, and R. Lacaze (2007), Lai and fapar cyclopes global products derived from vegetation. part 2: validation and comparison with modis collection 4 products, *Remote Sensing of Environment*, **110**(3), p. 317–331, doi:10.1016/j.rse.2007.03.001.

Weiss, M., H. Ma, F. Baret, and S. Jay (2023), Gai, fapar, fvc,and vcw from optical sensors (atbd): Taking sentinel-2 as an example.

White, M. A., K. M. De Beurs, K. Didan, D. W. Inouye, A. D. Richardson, O. P. Jensen, J. O'Keefe, G. Zhang, R. R. Nemani, W. J. D. Van Leeuwen, J. F. Brown, A. De Wit, M. Schaepman, X. Lin, M. Dettinger, A. S. Bailey, J. Kimball, M. D. Schwartz, D. D. Baldocchi, J. T. Lee, and W. K. Lauenroth (2009), Intercomparison, interpretation, and assessment of spring phenology in North America estimated from remote sensing for 1982–2006, *Global Change Biology*, **15**(10), p. 2335–2359, doi:10.1111/j.1365-2486.2009.01910.x.

Wiesmann, A., and C. Mätzler (1999), Microwave emission model of layered snowpacks, *Remote sensing of environment*, **70**(3), p. 307–316.

Wigneron, J.-P., A. Chanzy, J.-C. Calvet, and N. Bruguier (1995), A simple algorithm to retrieve soil moisture and vegetation biomass using passive microwave measurements over crop fields, *Remote Sensing of Environment*, **51**(3), p. 331–341.

Wigneron, J.-P., P. Ferrazzoli, A. Olioso, P. Bertuzzi, and A. Chanzy (1999), A Simple Approach To Monitor Crop Biomass from C-Band Radar Data, *Remote Sensing of Environment*, **69**(2), p. 179–188, doi:10.1016/S0034-4257(99)00011-5.

Wigneron, J.-P., M. Pardé, P. Waldteufel, A. Chanzy, Y. Kerr, S. Schmidl, and N. Skou (2004), Characterizing the dependence of vegetation model parameters on crop structure, incidence angle, and polarization at l-band, *IEEE Transactions on Geoscience and Remote Sensing*, **42**(2), p. 416–425.

Wigneron, J.-P., Y. H. Kerr, P. Waldteufel, K. Saleh, M.-J. Escorihuela, P. Richaume, P. Ferrazzoli, P. de Rosnay, R. Gurney, J.-C. Calvet, J. P. Grant, M. Guglielmetti, B. Hornbuckle, C. Mätzler, T. Pellarin, and M. Schwank (2007a), L-band Microwave Emission of the Biosphere (L-MEB) Model : Description and calibration against experimental data sets over crop fields, *Rem. Sens. Environ.*, **107**, p. 639–655.

Wigneron, J.-P., Y. Kerr, P. Waldteufel, K. Saleh, M.-J. Escorihuela, P. Richaume, P. Ferrazzoli, P. De Rosnay, R. Gurney, J.-C. Calvet, J. Grant, M. Guglielmetti, B. Hornbuckle, C. Mätzler, T. Pellarin, and M. Schwank (2007b), L-band Microwave Emission of the Biosphere (L-MEB) Model: Description and calibration against experimental data sets over crop fields, *Remote Sensing of Environment*, **107**(4), p. 639–655, doi:10.1016/j.rse.2006.10.014.

Wigneron, J.-P., T. Jackson, P. O'Neill, G. de Lannoy, P. de Rosnay, J. Walker, P. Ferrazzoli, V. Mironov, S. Bircher, J. Grant, M. Kurum, M. Schwank, J. Munoz-Sabater, N. Das, A. Royer, A. Al-Yaari, A. A. Bitar, R. Fernandez-Moran, H. Lawrence, A. Mialon, M. Parrens, P. Richaume, S. Delwart, and Y. Kerr (2017), Modelling the passive microwave signature from land surfaces: A review of recent results and application to the L-band SMOS and SMAP soil moisture retrieval algorithms, *Remote Sensing of Environment*, **192**, p. 238 – 262, doi:https://doi.org/10.1016/j.rse.2017.01.024.

Wigneron, J.-P., X. Li, F. Frappart, L. Fan, A. Al-Yaari, G. De Lannoy, X. Liu, M. Wang, E. Le Masson, and C. Moisy (2021), Smos-ic data record of soil moisture and l-vod: Historical development, applications and perspectives, *Remote Sensing of Environment*, **254**, p. 112,238.

Xiao, Z., S. Liang, J. Wang, P. Chen, X. Yin, L. Zhang, and J. Song (2014), Use of general regression neural networks for generating the glass leaf area index product from time-series modis surface reflectance, *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, **52**(1), p. 209–223.

Xu, L., S. S. Saatchi, Y. Yang, Y. Yu, J. Pongratz, A. A. Bloom, K. Bowman, J. Worden, J. Liu, Y. Yin, et al. (2021), Changes in global terrestrial live biomass over the 21st century, *Science Advances*, **7**(27), p. eabe9829.

Yebra, M., P. E. Dennison, E. Chuvieco, D. Riaño, P. Zylstra, E. R. Hunt, F. M. Danson, Y. Qi, and S. Jurdao (2013), A global review of remote sensing of live fuel moisture content for fire danger assessment: Moving towards operational products, *Remote Sensing of Environment*, **136**, p. 455–468, doi:https://doi.org/10.1016/j.rse.2013.05.029.

Zhao, Q. (2013), Methodology of modeling multiple scattering effects in microwave remote sensing of vegetation, Ph.D. thesis, The George Washington University.

Zhou, Y., A. Sharma, M. Kurum, R. H. Lang, P. E. O'Neill, and M. H. Cosh (2020), The backscattering contribution of soybean pods at L-band, *Remote Sensing of Environment*, **248**, p. 111,977.

Zhou, Y., M. Schwank, M. Kurum, and A. Mialon (2023), Modelling scattering albedo of trees from 1 to 37 GHz and its application to VOD retrieval, in *2023 IEEE International Geoscience and Remote Sensing Symposium*, p. 2735–2738, IEEE.

Zhou, Y., D. Houtz, M. Schwank, M. Kurum, Q. Zhao, R. Lang, A. Mialon, and M. Drusch (2024), A microwave emission model for layered vegetation (memlv): A case study for boreal forests from l- to ka-band, *TGRS*.

Zotta, R.-M., L. Moesinger, R. van der Schalie, M. Vreugdenhil, W. Preimesberger, T. Frederikse, R. de Jeu, and W. Dorigo (2024), Vodca v2: Multi-sensor, multi-frequency vegetation optical depth data for long-term canopy dynamics and biomass monitoring, *Earth System Science Data Discussions*, **2024**, p. 1–45, doi:10.5194/essd-2024-35.