Dottorato di Ricerca in / PhD program : Risorse per la nuova P.A.: Persone e Dati

Ciclo / Cycle XXXVIII

Curriculum 2. Dati

Integrating Machine Learning, Remote Sensing, and Eddy Covariance Data to Assess Vegetation Parameters and Carbon Fluxes

Nome / Name Luca Cognome / Surname Tuzzi

Matricola / Registration number 814079

Tutor: prof. Roberto Colombo Co-tutor: Dott. Marta Galvagno

Coordinatore / Coordinator: prof. Alfredo Marra

ANNO ACCADEMICO / ACADEMIC YEAR
2024-2025

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Table of Contents

Abstract

Understanding vegetation structure and carbon fluxes is essential for assessing ecosystem functioning and biogeochemical cycles. Alpine ecosystems pose particular challenges due to extreme environmental conditions and limited data availability. This thesis presents two complementary research lines integrating remote sensing, machine learning, and in situ measurements to improve the estimation of key vegetation parameters and carbon fluxes.

The first line focuses on estimating Leaf Area Index (LAI) and Fractional Cover (FC) using the SBG-TIR mission optical data and a machine learning algorithm trained on synthetic reflectance from the SCOPE radiative transfer model. The algorithm achieved excellent performance on synthetic validation data, with RMSE of 0.046 for FC and 0.053 for LAI. When applied to real observational data, the retrieval yielded RMSE of 0.19 for FC and 1.02 for LAI, consistent with state-of-the-art operational methods. These results confirm that combining physically-based simulations with data-driven learning allows robust and transferable estimation of vegetation structural parameters in spectrally complex and heterogeneous environments.

The second line addresses local carbon flux quantification in alpine ecosystems, focusing on the Torgnon site (Aosta Valley). A local optimization of the FLUXCOM framework, trained on fewer than 20 alpine sites with environmental conditions similar to Torgnon, outperformed the global FLUXCOM model in reproducing gross primary productivity (GPP). This demonstrates that regionally tuned models better capture underrepresented ecosystems, improving the accuracy of regional carbon budget estimates.

Overall, this thesis shows that physically-informed machine learning approaches can provide reproducible, interpretable, and scalable estimates of vegetation structure and carbon dynamics. The methodologies developed bridge the gap between theoretical modeling and operational applications, enhancing our understanding of alpine ecosystem functioning and supporting informed environmental monitoring and management.

Chapter 1

Introduction

Understanding ecological and biogeochemical processes on Earth requires accurate and consistent data on vegetation and carbon fluxes. The estimation of key canopy parameters is essential to describe vegetation structure and to model processes such as the modulation of surface albedo and emissivity, as well as evapotranspiration, which is closely linked to carbon exchanges between the biosphere and the atmosphere (Bonham, 2013; Chen & Black, 1992). Carbon and energy fluxes can be measured at the local scale by eddy covariance towers, while integration with satellite observations allows upscaling of these fluxes at regional-to-global scales. This integration allows precise quantification of the role of vegetation in regional carbon budgets and supports informed scientific and policy decisions. Remote sensing methods, combined with data-driven techniques such as machine learning, are powerful tools to estimate these parameters, particularly for complex and underrepresented ecosystems, such as alpine environments, where accurate assessments are especially challenging (Jung et al., 2020). Machine learning algorithms allow flexible linking of target variables to available measurements without the need to define explicit functional relationships. These models establish numerical relationships between inputs and outputs, iteratively optimizing internal parameters based on the data, offering greater adaptability and robustness compared to traditional methods (Verrelst et al., 2015).

Within this framework, this doctoral research pursues two main objectives. The first is to develop a data-driven approach to estimate structural vegetation parameters. The second is to optimize an existing method for quantifying carbon exchanges in alpine ecosystems, using machine learning models to integrate satellite and ground-based data in a consistent and interpretable framework. This approach enables reproducible, accurate, and interpretable estimates, representing an innovative yet already stable method suitable for complex operational applications.

To achieve the first objective, the Surface Biology and Geology Thermal InfraRed (SBG-TIR) mission, developed in collaboration between the Italian Space Agency (ASI)

and NASA/JPL, provides essential data for characterizing vegetation and surface processes. The mission will acquire spectral data across both the optical (400-2500 nm) and thermal infrared domains (3-5 µm and 8-12 µm) both at 60 m spatial resolution. In particular, the Visible InfraRed Earth Observation camera (VIREO) constitutes the optical instrument. It is a two-band scanner operating in the Visible and Near Infrared (VNIR) regions, with channels centered at 655nm (VNIR0, red region) and 835 nm (VNIR1, near infrared region) both featuring a 60 m ground sample distance. Additionally, a panchromatic band (PAN) centred at 750 nm provides enhanced spatial detail at 30 m resolution.

Beyond its instrumental design, the mission enables the observation of several geophysical variables, including surface temperature, evapotranspiration, snow properties, and volcanic activity. In this work we exploit the optical component of the SBG-TIR mission to develop an algorithm capable of accurately estimating Fractional Cover (FC) and Leaf Area Index (LAI), two key parameters for a wide range of terrestrial applications and for evapotranspiration modelling. FC indicates the fraction of ground covered by green vegetation, quantifying canopy spatial extent (Bonham, 2013), whereas LAI represents the one-sided leaf area per unit ground area and is closely linked to transpiration as an indicator of active leaf biomass (Chen & Black, 1992). The machine learning model was trained using synthetic data generated from the Soil Canopy Observation, Photochemistry, and Energy fluxes model (SCOPE; Van der Tol et al., 2009), a radiative transfer model simulating the spectral reflectance of vegetation and soil. The simulated reflectance was resampled to match the characteristics of the VIREO instrument on the SBG-TIR mission, creating a training dataset consistent with the target values of FC and LAI. Using synthetic data combines the theoretical rigor of physical models with the predictive capability of machine learning, enhancing the reliability and transferability of estimates under operational conditions. The algorithm was subsequently applied to real observational data, not included in the training dataset, to validate its operational performance and assess its potential applications within the mission. Technical details on model construction, feature selection, and quantitative results are reported in Chapter 2.

Concurrently, the second line of research focuses on quantifying carbon exchanges at the local scale, a critical aspect for understanding biogeochemical cycles and evaluating the impacts of climate change. Among established methods, the FLUXCOM (Jung et al., 2020 or Nelson et al., 2024) framework utilizes data from over 300 eddy covariance stations, combining in situ measurements, meteorological information, and satellite observations to estimate carbon fluxes. This approach allows extending estimates to unobserved sites, producing globally consistent spatial maps of carbon fluxes derived from model-based extrapolations (Nelson et al., 2024). The original contribution of this work lies in the local optimization of the model for the Alpine region, with particular focus on the Torgnon site in the Aosta Valley, characterized by extreme ecological conditions poorly represented in global datasets, such as approximately 200 days of snow cover per year. The analysis evaluated whether more accurate values for gross primary productivity (GPP) could be achieved by applying a localized model built using less than 20 experimental sites out of the 300 available in the global dataset, with environmental characteristics similar to the reference site rather than a global one. This approach has important implications for assessing carbon budgets in alpine ecosystems and for understanding the dynamics of local biogeochemical cycles. Methodological details and quantitative results are presented in Chapter 3 of the thesis.

In summary, this thesis documents two main lines of research developed during the doctoral program: Chapter 2 presents the development and validation of an algorithm for estimating Fractional Cover (FC) and Leaf Area Index (LAI) within the context of the SBG-TIR mission, while Chapter 3 describes the local optimization of the FLUXCOM framework for carbon fluxes in alpine ecosystems.

References

Bonham, C. D., 2013. Measurements for Terrestrial Vegetation, 2nd ed. Hoboken, NJ: Wiley.

Chen, J.-M., Black, T. A., 1992. Defining leaf area index for non-flat leaves. Plant Cell and Environment, vol. 15, no. 4, pp. 421–429. https://doi.org/10.1111/j.1365-3040.1992.tb00992.x

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., Gans, F., Goll, D. S., Haverd, V., Köhler, P., Ichii, K., Jain, A. K., Liu, J., Lombardozzi, D., Nabel, J. E. M. S., Nelson, J. A., O'Sullivan, M., Pallandt, M., Papale, D., Peters, W., Pongratz, J., Rödenbeck, C., Sitch, S., Tramontana, G., Walker, A., Weber, U., e Reichstein, M., 2020. Scaling carbon fluxes from eddy covariance sites to the globe: synthesis and evaluation of the FLUXCOM approach. Biogeosciences, vol. 17, pp. 1343–1365. https://doi.org/10.5194/bg-17-1343-2020

Nelson, J. A., e Walther, S., et al., 2024. X-BASE: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, FLUXCOM-X. Biogeosciences, 21, 5079–5115. https:// doi.org/10.5194/bg-21-5079-2024

Van der Tol, C., Verhoef, W., Timmermans, J., Verhoef, A., Su, Z., 2009. An integrated model of soil-canopy spectral radiances, photosynthesis, fluorescence, temperature and energy balance. Biogeosciences, vol. 6. https://doi.org/10.5194/bg-6-3109-2009

Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J. P., Veroustraete, F., Clevers, J. G. P. W., Moreno, J., 2015. Jochem Verrelst, Gustau Camps-Valls, Jordi Muñoz-Marí, Juan Pablo Rivera, Frank Veroustraete, Jan G.P.W. Clevers, José Moreno. ISPRS

Journal of Photogrammetry and Remote Sensing, vol. 108, pp. 273–290. https://doi.org/10.1016/j.isprsjprs.2015.05.005

Chapter 2

Estimation of Leaf Area Index and Vegetation Fractional Cover in SBG-TIR configuration using SCOPE simulated data and Sentinel-2 images.

Luca Tuzzi[1], Sara Venafra[2], Roberto Colombo[1]

[1]Department Earth and Environmental Sciences Department, University of Milano-Bicocca, Milano (Italy)

[2]Italian Space Agency, Roma (Italy)

Keywords: Machine learning, Remote sensing, Sentinel2, Multispectral imaging, Radiative Transfer Model, Data analysis, SBG-TIR

Highlights

• Accurate retrieval of LAI and FC using Gaussian Process Regression from SCOPE synthetic data in SBG-TIR configuration.

• RED and NIR channels alone provide high predictive accuracy for vegetation metrics. Panchromatic channel and NIRv enhances accuracy in mixed-pixel scenario.

• Application to Sentinel-2 data indicate the transferability to other multispectral sensors.

2.1. ABSTRACT

In this paper, we present a study conducted within the framework of the forthcoming joint NASA/ASI Surface Biology and Geology Thermal Infrared (SBG-TIR) mission, which will operate in a sun-synchronous polar orbit collecting data at global scale. The mission will acquire spectral data in the thermal infrared domains (3–5 µm and 8–12 µm) and in the Visible and Near-Infrared (VNIR) with two spectral bands centered at 655 nm (VNIR0) and 835 nm (VNIR1), along with a panchromatic channel spanning 580–920 nm. All spectral bands will be collected at 60 m spatial resolution. In this context, accurate retrieval of Fractional Vegetation Cover (FC) and Leaf Area Index (LAI) is of particular relevance, as it enables synergistic use of VNIR and TIR observations to support vegetation monitoring and surface energy flux estimation.

To address this objective, several machine learning algorithms were evaluated under different configurations for the retrieval of FC and LAI. Model training was performed using synthetic datasets generated with the Soil-Canopy Observation, Photochemistry and Energy Fluxes (SCOPE) radiative transfer model, which simulates reflectance and radiance from soil, leaf, and canopy components across the optical spectrum (Van der Tol et al., 2009; Yang et al., 2020). Calibration and validation were conducted using independent synthetic datasets to ensure generalizability and robustness. Input features included VNIR spectral bands, the panchromatic channel, vegetation indices, and variables describing illumination and observation geometry.

Model performance was assessed on independent test data, with uncertainty quantification included. The optimal configuration achieved RMSE values of 0.046 for FC and 0.053 m^2/ m^2 for LAI using a six-channel input set. These results are consistent with previous studies (e.g., García-Haro et al., 2018; Weiss et al., 2020), supporting the validity of the proposed approach. The trained models were subsequently applied to Sentinel-2 imagery to evaluate performance under real-world conditions. Validation was performed using the GBOV (Ground-Based Observations for Validation) dataset (Bai et

al., 2019) and standard Sentinel-2 biophysical retrievals. The results demonstrated strong statistical agreement with the Biophysical Processor implemented in the ESA Sentinel Application Platform (SNAP) toolbox (Weiss and Baret, 2016), confirming the robustness of the proposed framework for operational estimation and mapping of FC and LAI in the context of SBG-TIR space mission.

2.2. INTRODUCTION

The SBG (Surface Biology and Geology) Thermal InfraRed (TIR) mission is a cooperation of the Italian Space Agency (ASI) with NASA/JPL for the development of an earth observing satellite for the measurement and detection of land surface temperature and emissivity, evapotranspiration, snow properties, soil moisture, minerals, wildfires and volcanoes. The instrumental payload of the SBG-TIR satellite is composed by two instruments on the same rotating mirror. The thermal instrument, the Observing Thermal Emission Radiometer (OTTER) consists of a TIR multispectral scanner with six spectral bands operating between 8 and 12.5 μm and two mid-infrared bands at 4 μm and 4.8 μm, with a 60 m ground sample distance. The optical instrument, Visible InfraRed Earth Observation camera (VIREO) consists of a Visible and Near Infrared (VNIR) two bands scanner at 655 (VNIR0 in the red spectral region) and 835 nm (VNIR1, in the Near Infrared spectral region) with a 60 m ground sample distance, and a panchromatic band (PAN) centred at 750 nm with 30 m spatial resolution obtained from the native 60 m resolution through super-resolution. The full width at half maximum of the multispectral camera is 80 nm for VNIR bands and 300 nm for PAN channel. The instruments field of view is ± 34.4° and the overall swath is 935 km with a revisit time less than 3 days. SBG-TIR has the possibility to collect images day and night and the local time descending node for the daily overpass is at 12:30.

One of the most important products of the mission is the real evapotranspiration (ET) computed at high spatial resolution. In this context, the exploitation of the data coming from the VIREO camera can allow the estimation of some vegetation parameters that represent key input for ET modelling. In particular Leaf Area Index (LAI, m^2/m^2) and

the Fraction of vegetation Cover (FC, %) are crucial for a broad variety of land applications and for the modelling of ET.

The FC corresponds to the fraction of ground covered by green vegetation (Bonham, 2013) and quantifies the spatial extent of the vegetation as projected at nadir. LAI indicates the one-sided leaf area per unit area of ground (Chen and Black, 1992) and it is closely related to the transpiration process since it drives the actual green biomass. Both parameters are widely used to modulate surface albedo and surface emissivity and for mixed pixels it is possible to define a weighted function of the mixture components (e.g., vegetation and soil). The link of this vegetation structural parameters with thermal applications is well demonstrated (e.g. Lin et al., 2020, Yin et al., 2020) and this link open synergies between VNIR and TIR cameras during the SBG-TIR space mission.

Indeed, the direct retrieval of FC and LAI from the VIREO camera is very important in the context of SBG-TIR mission, since it ensures co-registration with land surface temperature maps and eliminates geolocation errors. It also reduces reliance on external data sources, minimizing the impact of cloud cover associated with the assimilation of data coming from other space missions. Moreover, it enables consistent retrievals across the sensor's wide swath and facilitates synergistic use of VNIR and TIR observations throughout the mission. For these reasons, the development of an ad-hoc algorithm to retrieve FC and LAI within the SBG-TIR mission is desirable.

In recent years, various approaches based on satellite data have been developed to estimate LAI and FC from multispectral data. Although the SBG-TIR offer spectral bands in the red and near-infrared regions only, these are particularly effective for estimating biophysical parameters such as LAI and FC since they contain most of the spectral information sensitive to structural variations in the canopy. These two bands form the foundation of widely used vegetation indices, notably the Normalized Difference Vegetation Index (NDVI), which exploits the contrast between strong chlorophyll absorption in the red and high reflectance in the NIR due to leaf internal structure. Red and NIR are strongly correlated with vegetation density and condition and can serve as a proxy for estimating structural vegetation parameters. Even in the absence

of broader spectral coverage, the availability of just red and NIR bands allows for reliable estimations, especially when supported by auxiliary data or local calibration. This approach is particularly advantageous for satellite missions with multispectral payload capacity, such as SBG-TIR, where the VIREO instrumental simplicity translates into higher revisit frequency and broader spatial coverage.

Retrieval methods based on the inversion of radiative transfer models (RTMs) have been used to generate operational biophysical products from Earth observation data. For instance, the CYCLOPES products (Carbon cYcle and Change in Land Observational Products from an Ensemble of Satellites) (Baret et al., 2007) are derived from VEGETATION satellite data through inversion of the PROSAIL model. The MODIS (MODerate Resolution Imaging Spectroradiometer) LAI and Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) products are based on a 3D RTM defined for eight biomes (Myneni et al., 2002; Knyazikhin et al., 1998). The GEOV1 products from Copernicus GEOland2 (Baret et al., 2013) are generated by fusing and scaling MODIS and CYCLOPES products using data from SPOT/VEGETATION and PROBA-V.

However, selecting the most appropriate algorithm requires careful consideration of the reliability of the estimated parameters and their associated uncertainties, considering expected accuracy, robustness, and timeliness for operational production (Garcia-Haro, 2018). These criteria tend to favour hybrid approaches based on machine learning techniques, which are computationally efficient, adhere to the physical principles embedded in RTMs, and are generally less prone to the convergence issues that affect physical inversion methods (Verrelst et al., 2012, 2015; Camps-Valls et al., 2016), often caused by the non-linearity of the inverse problem in remote sensing (Camps-Valls et al., 2017).

Hybrid methods combine the generalization capabilities of physical models with the accuracy and efficiency of non-parametric machine learning techniques (Verger et al., 2008; Verrelst et al., 2012; Houborg & McCabe, 2018, Garcia-Haro, 2018). Among these, Neural Networks (NNs) have been implemented in operational processing chains

to retrieve global biophysical parameters by inverting the PROSAIL model (Bacour et al., 2006; Baret et al., 2007). More recently, kernel-based algorithms have been introduced for classification and regression tasks in remote sensing (Pérez-Suay et al., 2017). Support Vector Regression (SVR) has been applied to retrieve LAI, FVC, and evapotranspiration (Yang et al., 2006; Durbha et al., 2007), while Gaussian Process Regression (GPR) has demonstrated superior performance in LAI retrieval (Lázaro-Gredilla et al., 2014; Campos-Taberner et al., 2016). GPR is particularly effective in handling heterogeneous and noisy data and provides confidence intervals for its predictions. However, machine learning approaches have not yet been implemented in operational retrieval chains for global-scale vegetation parameter estimation.

In this context, we conducted a study to evaluate the suitability of the VIREO camera to retrieve FC and LAI. The Soil, Canopy Observation, Photochemistry and Energy Fluxes (SCOPE, Van der Tol et al., 2009) radiative transfer model has been used to simulates soil, leaf and canopy reflectance. The simulations have been conducted in a wide range of viewing geometries and biophysical vegetation parameters. Then, different Machine Learning (ML) techniques have been tested on simulated data, and the best performance model was applied to Sentinel-2 imagery to validate model transferability and the performance under real cases. This ensures that the algorithms trained on synthetic simulations are robust under realistic observation conditions, capturing real-world variability and sensor noise. Separate single-output models were developed for each target variable (i.e., LAI and FC), allowing independent optimization while maintaining conceptual consistency. Controlled noise was introduced during training to better represent input variability and potentially reduce retrieval errors, reflecting a strategy increasingly recognized in state-of-the-art remote sensing frameworks.

Overall, the proposed design provides a physically based, data-driven retrieval framework that bridges the gap between radiative transfer modelling and machine learning. It ensures interpretability and operational flexibility, while providing actionable biophysical information to support applications ranging from agricultural management and forest monitoring to climate change studies. These capabilities enhance the operational value of SBG-TIR observations for both the scientific community and

decision makers. The methodology is fully transferable, allowing adaptation to other multispectral sensors, and lays a solid foundation for future large-scale and physically-based retrievals of vegetation biophysical parameters.

2.3 OVERALL METHODOLOGY

The overall procedure to retrieve LAI and vegetation FC in SBG-TIR configuration is outlined in Figure 1. Initially, the SCOPE radiative transfer model (Van Der Tol et al., 2009; Yang et al., 2020) was employed to simulate spectral signatures of soil, leaves, and canopy. SCOPE simulations have been conducted over a wide range of viewing geometries, leaf and canopy vegetation parameters, soil optical and physical properties and thermal variables.

By varying the initial conditions, a spectral library was constructed by pairing the simulated reflectance and thermal signals with the corresponding biophysical parameters. These spectra were then processed by using the Instrument Spectral Response Function (ISRF) and then employed in the training phase of a machine learning algorithm aiming at solving the inversion problem, i.e. associating the biophysical variables to the given spectral information. Separate models were trained for each target parameter and subsequently tested on simulated data to identify the best-performing one in terms of accuracy and robustness, based on the synthetic dataset. The final models were then evaluated on Sentinel-2 images and the GBOV (Ground-Based Observations for Validation) was used to validate the estimates.

2.3.1 Simulation pipeline

2.3.1.1 Model description and parameterization

SCOPE is a canopy-scale model that couples optical radiative transfer theory with models of photosynthetic activity and thermal emission. It is designed to simulate spectral reflectance, solar-induced chlorophyll fluorescence (SIF), and thermal emission in the visible and near-infrared (VNIR), and Thermal InfraRed (TIR) domains. The model integrates the PROSPECT leaf model (Jacquemoud & Baret, 1990) and the SAIL

canopy model (Verhoef, 1984), extending them with modules that account for key biochemical and physiological processes including stomatal regulation, leaf temperature dynamics, carbon assimilation, and xanthophyll cycle activity (Van der Tol et al., 2009; Yang et al., 2021). Within the SCOPE model, the leaf and canopy are represented as a one-dimensional (1D) turbid medium, where the electromagnetic radiative transfer equations are solved using a multi-stream approximation. Basically, the 1D models assume that the canopy varies only with height above the ground surface and is horizontally homogeneous.

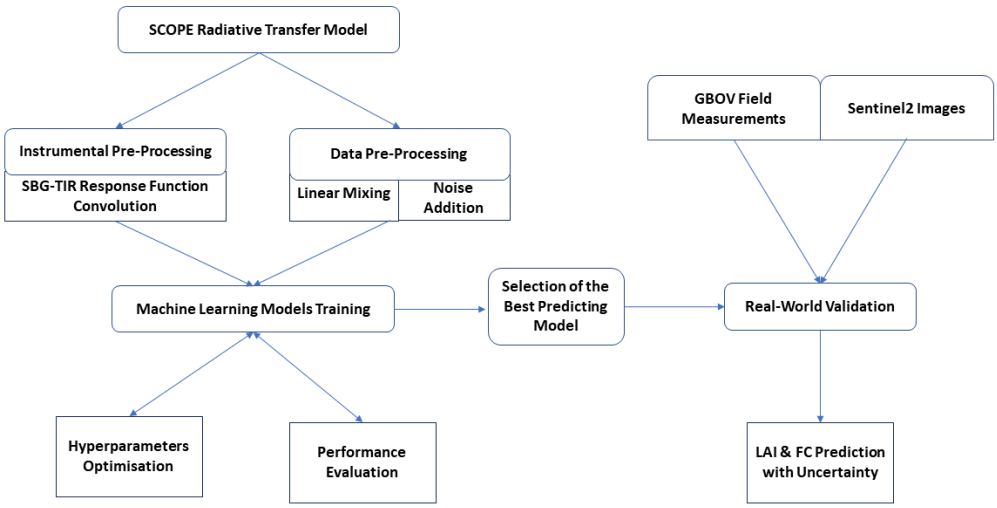

Figure 1: Flowchart summarizing the main steps for FC and LAI retrieval. Details are provided in sections 2.1, 2.2 and 2.3.

For each simulation run, a random sampling procedure was applied, where the input parameters were drawn from their respective probability distributions as defined in accordance with previous studies (e.g., Lauvernet et al., 2008; Claverie et al., 2013). SCOPE input parameters used for the simulations are listed in Table 1. Parameter domains were established based on empirical knowledge and designed to adequately capture the heterogeneity of real-world scenes and to uniformly sample the hypercube

of plausible experimental configurations (Baret et al., 2007, Verreslt et al., 2015). The Gaussian distribution was used most frequently, as it preserves the mean and standard deviation during the aggregation of variables describing both bare soil and vegetated surfaces (Frank, 2009). In some cases, this distribution was truncated to ensure physical plausibility. The uniform distribution was used only when each value within the domain had an equal likelihood of occurrence.

Some parameters exhibit strong internal correlation. Specifically, LAI and chlorophyll content (Cab) were sampled according to a joint probability structure, dividing the parameter space into four regions for each variable (Cab: 0-2, 5-20, 20-50, 50-80 $\mu g\, cm^{-2}$; LAI: 0-0.2, 0.5-2, 2-5, 5-8), which were sampled with a uniform probability density function. For each sampled LAI region, Cab values were drawn from the corresponding subset ensuring realistic combinations, while avoiding implausible combinations such as high LAI paired with very low Cab. For the remaining parameters, mutual correlations are often species or context dependent and therefore difficult to generalize; imposing joint probability structures could introduce artificial biases into the parameter space. Instead, additional plausibility constraints were applied to the leaf water content (Cw) and dry matter content (Cdm), to ensure realistic leaf composition: the water fraction of total leaf mass was constrained between 40% and 95%, reflecting typical values for green foliage (Lauvernet et al., 2008). Leaf orientation within the canopy was described using the Leaf Inclination Distribution Functions (LIDFa and LIDFb). LIDFa represents the mean leaf inclination, indicating the predominant leaf angle, while LIDFb quantifies the variability around this mean orientation. Both parameters are constrained by geometric considerations to ensure that their sum does not exceed 1 (Verhoef, 1984), thus maintaining a physically realistic canopy structure. For soil parameters, no explicit constraints were applied. The ranges used were based on Verhoef et al. (2018), simulating the effects of soil parameters on soil reflectance by modifying the Bach and Mauser (1994) model using a statistical-based approach. For example, in the case of soil moisture content (SMC), the

model directly generates realistic soil reflectance spectra for topsoil moisture contents up to 55% volumetric.

Geometric solar parameters (viewing and illumination angles) were defined according to the mission-specific observation geometry. Azimuth angles were treated as free parameters, although only the Relative Azimuth Angle (RAA) significantly influences radiative transfer model outputs under the 1-D assumption (Verhoef, 1984). Geometrical configurations are crucial due to the large off-nadir geometry of the space mission. Hence, we explicitly include Solar Zenith Angle (SZA), View Zenith Angle (VZA, constrained by the satellite's 935 km swath width), RAA, derived from the solar and sensor azimuth angles).

Running SCOPE involves numerically solving radiative transfer equations across its modules, using the input parameters listed in Table 1. Soil spectra were generated using the Brightness-Shape-Moisture (BSM) module, while vegetation leaf spectra were computed via PROSAIL. These outputs were combined into a spectral library consisting of vegetation soil reflectance pairs associated with their respective input parameter sets. This allowed for the creation of a synthetic dataset representing the diversity of scenes expected during the satellite mission (Verhoef, 1984). SCOPE outputs are provided with high spectral resolution of 1 nm in the VNIR-Short Wave Infrared Region (SWIR) range (400-2400 nm), 0.1 µm in the TIR range (2.5-15 µm), 1 µm in the far TIR (15-50 µm). Reflectance in the VNIR-SWIR was computed as the ratio of reflected to incoming radiation (both direct and diffuse components).

Parameter	PDF	Min	Max	Mean	Std	Definition
N	truncated gaussian	1.2	2.2	1.5	0.3	Leaf mesophyll structure parameter
LAI	uniform	0.001	8			Leaf Area Index (m^2/m^2)
Cab		0.01	80			Chlorophyll a and b content ($\mu g/cm^2$)
Cca	truncated gaussian	0	30	10	5	Carotenoid content ($\mu g/cm^2$)
Cdm	joint TND	0.003	0.021	0.005	0.005	Dry matter content (g/cm^2)
Cw		0.005	0.035	0.02	0.006	leaf water equivalent thickness (cm)
LIDFa	uniform	-1	1			Leaf Inclination Distribution Function
LIDFb		-1	1			
SMC	truncated gaussian	5	55	25	12.5	Soil moisture content in the root zone (%)
BSMBrightness	truncated gaussian	0.01	0.9	0.5	0.25	BSM model parameter for soil brightness
BSMlat	truncated gaussian	20	40	25	12.5	BSM model parameter 'lat'
BSMlon	truncated gaussian	45	65	50	10	BSM model parameter 'long'
tts (SZA)	uniform	0	60			Solar zenith angle (deg)
tto (VZA)	uniform	0	36			View zenith angle (deg)
psi (RAA)	uniform	0	180			Relative azimuth angle (deg)
Ta	uniform	-5	45			Air temperature (°C)
rs_thermal	uniform	0	0.1			broadband soil reflectance in the thermal range
rho_thermal	uniform	0	0.1			Broadband thermal reflectance
tau_thermal	uniform	0	0.1			Broadband thermal transmissivity
rss	uniform	1	3000			Soil surface evaporation resistance (s/m)
vCover	uniform	0	1			pixel fraction covered by vegetation

Table 1: Input parameters used for the simulations. Each parameter was sampled from its Probability Density Function (PDF). Normal distributions are defined by their mean and standard deviation (Std), truncated normal distributions additionally include minimum (Min) and maximum (Max) bounds. Uniform distributions are specified only by their minimum and maximum values.

The number of simulated spectra (Ns) varied between 500 and 12,000 to assess model performance and potential overfitting. In these simulations, the only source of variability was the random sampling of input parameters. Each simulation batch was initialized with a different predefined random seed (θr) to evaluate model robustness, avoid sampling bias and ensure reproducibility. The spectral library can thus be denoted as: $L(Ns, \theta r)$. Random numbers were generated using MATLAB's built-in Mersenne Twister algorithm with varying seeds. This strategy ensured an even sampling on a logarithmic scale, allowing assessment of model sensitivity to the choice of library, and evaluation of the optimal training size by balancing computational cost and parameter space coverage while mitigating overfitting.

2.3.1.2 Fractional cover modelling

The SCOPE code was extended to compute vegetation Fractional Cover. FC is related to the concept of gap fraction, which describes the probability of radiation not intercepting the canopy and reaching the ground (Zheng and Moskal, 2009). Mathematically the gap fraction is expressed as: $P(\theta) = \exp[-LAI \cdot \frac{G(\theta)}{cos(\theta)}]$. Here LAI is the one-sided leaf area per ground area, θ is the solar zenith angle (SZA). G(θ) represents the mean projection of leaf normal vector along the solar direction and accounts for the leaf inclination distribution function within the canopy. Within SCOPE the canopy is treated as a 1D turbid medium, implying a random distribution of leaves within the canopy volume (Van der Tol et al., 2009). Clumping effects are therefore not explicitly modelled in the computation of G(θ); instead the Poisson distribution is assumed (Nilson, 1971). G(θ) is computed following Nilson (1971), Campbell (1986), and Verhoef (1984). It is obtained by integrating the absolute projection of leaf normal along the solar direction over all possible leaf inclinations (φ) and azimuths (ψ), weighted by the parametric Leaf Inclination Distribution Function (LIDF). In simple terms, this means calculating how much the leaves "face" the sun on average, accounting for their typical orientation and variability. Equation 1 provides the numerical values of the Gamma function (Γ), used to represent the weighted averaging, evaluated for the LIDF parameters a and b used in the SCOPE simulation:

$$G(\theta) = cos(\theta)\ \frac{\Gamma\left(\frac{b+2}{2}\right)\Gamma\left(\frac{a+b+3}{2}\right)}{\Gamma\left(\frac{b+1}{2}\right)\Gamma\left(\frac{a+b+4}{2}\right)} \qquad (1)$$

Usually we refer to $k = \frac{G(\theta)}{cos(\theta)}$ as the extinction coefficient in the Lambert-Beer law for transmittance: $T = exp(-k \cdot LAI)$ leading to the common formulation for fractional cover: $FC = 1 - exp(-k \cdot LAI)$. This approach has been well established for FC estimation and adopted in several previous studies (Ding et al., 2016; García-Haro et al.,

2018; De Grave et al., 2020). By implementing this extension, Fractional Cover becomes a direct output of SCOPE, providing paired values of LAI and FC for each simulation.

2.3.1.3 Linear unmixing

To handle spatial heterogeneity a linear mixing model was exploited starting from the simulated data described in Section 2.1.2 (Figure 2). Under this assumption, each simulated pixel can be composed of a fraction of bare soil and a fraction of vegetation. The mixed reflectance spectrum of the SCOPE simulated data was therefore computed as in Equation 2:

$$R = vCover \cdot Rveg \ + (1 - vCover) \cdot Rsoil \qquad (2)$$

where Rveg is the reflectance from vegetation, Rsoil is the reflectance from soil and vCover is the fraction of the pixel covered by vegetation. Unlike FC, which is derived from physical laws vCover is an input parameter representing landscape-level aggregation and used to scale biophysical parameters (e.g., LAI). Vegetation parameters are linearly adjusted under the assumption that bare soil has LAI = 0, thus the effective LAI of the pixel becomes: $LAI\ mixed = LAI\ veg * vCover$. After spectrum mixing FC is recalculated from the adjusted LAI and resulting spectral data. Additionally, soil reflectance was weighted by a brightness parameter (Baret et al., 2007; Claverie et al., 2013).

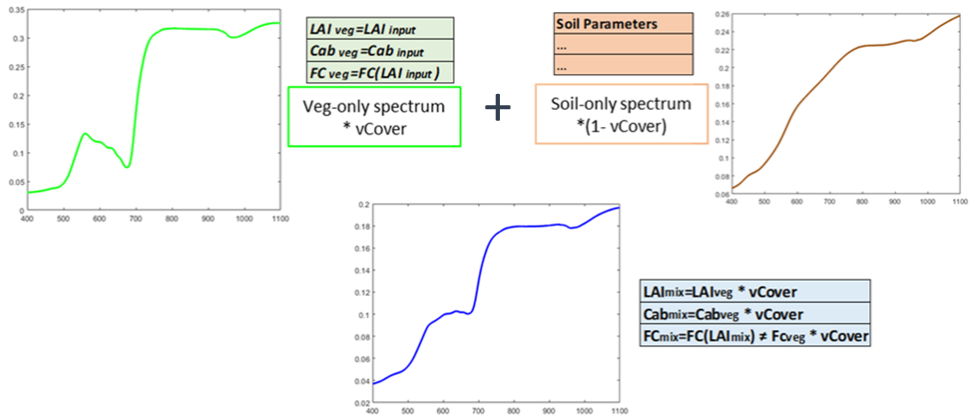

Figure 2: Linear unmixing model scheme. To better represent the spectrum of a real pixel, a vegetation spectrum was combined with a soil spectrum through a weighted average.

The output of this analysis was therefore the generation of simulated spectra under heterogeneous conditions, thereby bringing the simulations closer to scenarios that are more representative of real-world cases.

2.3.1.4 Noise implementation

Radiative transfer models provide an idealized Top-Of-Canopy (TOC) signal. However, real sensor observations are affected by uncertainties arising from different sources including: physical variability (e.g., clouds, adjacency effects), sensor limitations (e.g., calibration errors), scene heterogeneity at sub-pixel scale, atmospheric, geometric, and radiometric corrections. To introduce realism and avoid overfitting Gaussian white noise with variable intensity was applied to the synthetic reflectance (Brede et al., 2020), acting as a regularization mechanism by relaxing the strict input - output mapping during training. In line with Locherer et al. (2015), we implemented a wavelength-dependent multiplicative Gaussian noise model $\epsilon(\lambda) \sim N(0, \sigma)$, where σ denotes the standard deviation controlling its magnitude. The noise model is defined as in Equation 3:

$$R(\lambda) = R\ SCOPE\ (\lambda) \cdot [1 + \epsilon(\lambda)] = R\ SCOPE\ (\lambda) + R\ noise\ (\lambda) \quad (3)$$

where $R\ SCOPE\ (\lambda)$ is the reflectance spectrum obtained from SCOPE simulations and $R\ noise\ (\lambda)$ represents the noise component in the resulting noisy reflectance spectrum $R(\lambda)$. This noise model approximates realistic sensing conditions, since relative spectral shape is more important than absolute reflectance in retrieval tasks. The Signal to Noise Ratio (SNR) is defined as (Schowengerdt, 2007): $SNR = \frac{Power(R\ SCOPE)}{Power(R\ noise)} = \frac{1}{\sigma^2}$

For real-word multispectral instrument (MSI) SNR values vary across spectral bands. For instance, in Sentinel-2 the SNR in the visible range spans from 70 to 175. (ESA, 2015; Drusch et al., 2012). In this study, three noise levels were considered to assess model performance under varying observational uncertainties condition: i) no noise ($SNR = \infty$), ii) low to medium noise (SNR = 100) typical of operational satellite data and iii) high noise (SNR = 10), simulating severely degraded atmospheric or instrumental conditions. Statistically, these noise levels correspond to reflectance perturbations with 68% probability of falling within ± [10%, 33%] of the original signal.

2.3.1.5 Input data, dataset resampling and spectral indices computation

As previously described, the SBG-TIR mission includes multiple onboard sensors operating across different spectral regions of the VNIR-TIR spectral range. Regarding the VIREO camera, the Spectral Instrument Response Functions of the VNIR0, VNIR1 and PAN are shown in Figure 3. For vegetation applications, VNIR0 in the red and VNIR1 in the near infrared, are sensible to chlorophyll absorption and canopy scattering, respectively, while the inclusion of the PAN channel can account for the vegetation red-edge spectral region and for the bare soil contribution that typically characterizes mixed pixels in real scenes. The reflectance values in these three optical channels were obtained by applying the discrete convolution between the SCOPE-simulated reflectance spectra ($R(\lambda)$) and the ISRF of each channel (Figure 3). For a given channel C, the convolution is computed as: $C = \sum R(\lambda) \cdot ISRF(\lambda)$.

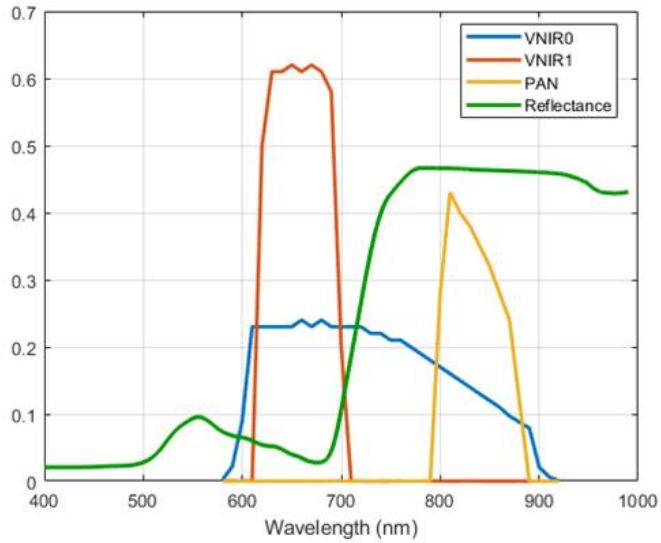

Figure 3: SBG-TIR Spectral Instrument Response Functions (ISRF) of the in the optical domain, showing three main channels: VNIR0 (RED region), VNIR1 (Near-Infrared, NIR), and the panchromatic (PAN) band. The green line represents an example of a simulated vegetation spectrum.

Apart from the three VIREO channels and the solar Illumination and sensor viewing angle, different input data were tested in the models selected for the estimation of LAI and FC. For example, different spectral indices were computed and among them, we tested the near-infrared reflectance of vegetation index (NIRv), defined as the product of near-infrared reflectance and NDVI. This index is commonly used to approximate the fraction of pixel reflectance attributable to vegetation within a mixed pixel (Badgley et al., 2017). Moreover, NIRv has been shown to better capture variations influenced by diurnal or weekly changes in canopy structure and function (Chen et al., 2023). Finally, since daily temporal resolution was not considered in the simulation framework, NIRv may provide a useful proxy to reduce the influence of intra-day variability when the temporal resolution is limited to weekly or longer, as in the experimental validation dataset used. For these reasons, the NIRv spectral index was selected as new input data in the modelling analysis furtherly described.

Basically, the following six configurations (based on different input data), were considered to test the selected six Machine Learning models described in Section 2.2. The input data configurations considered in this study are: i) VNIR0, VNIR1, SZA, VZA, RAA; ii) VNIR0, VNIR1, SZA, VZA, RAA, NIRv; iii) VNIR0, VNIR1, SZA, VZA, RAA, PAN, NIRv; iv) VNIR0, VNIR1, SZA, VZA, RAA, PAN; v) VNIR0, VNIR1, SZA, VZA, RAA; vi) VNIR0, VNIR1.

2.3.2 Machine Learning models, training, implementation, validation and uncertainty

2.3.2.1 Model description

To explore a wide range of learning paradigms and assess their predictive capabilities, we tested six different supervised regression models based on Machine Learning: i) decision tree regression (DT), ii-iii) ensemble methods (Random Forest, RF, and Gradient Boosted Ensemble specifically Least-Squares Boosting, LSB), iv) kernel-based regression (Gaussian Process Regression, GPR), v) Support Vector Machine (SVM), (vi) Neural Network approaches (NN). This methodological diversity allows for a comprehensive evaluation of model performance across different regression strategies and provides insights into their generalization abilities in the context of vegetation biophysical parameter retrieval. Each model was optimized via Bayesian optimization over a problem-specific hyperparameter space as further described.

Decision trees are interpretable, non-parametric models that recursively partition the input space to maximize the homogeneity within resulting regions. Their main advantage lies in their ability to capture complex, nonlinear relationships and interactions among variables, without requiring prior feature scaling or transformation so performing well when the data presents well-defined patterns, even if nonlinear. However, individual trees are typically prone to high variance and overfitting. In our implementation we

controlled tree complexity through the minimum number of samples per leaf and the maximum depth (or number of splits).

Random forests are ensemble methods that mitigate the overfitting tendency of single trees by aggregating the predictions of multiple trees trained on bootstrapped data and random feature subsets. They are relatively robust to noise and exhibit consistent performance across different datasets. However, their ensemble nature can limit extrapolation beyond the domain of the training data. We optimized the number of trees, the minimum leaf size, and the number of randomly selected features at each split.

Boosted ensembles, such as Least-Squares Boosting, build a sequence of models in which each successive learner attempts to correct the residuals of its predecessor. This often leads to superior predictive performance but also introduces greater sensitivity to hyperparameters. Specifically, the learning rate, number of boosting iterations, and individual tree complexity must be carefully tuned.

Support Vector Regression (SVR) is particularly well suited for high-dimensional spaces and sparse datasets. Using kernel functions, SVRs can approximate complex, nonlinear relationships while maintaining strong regularization properties. They are relatively robust to overfitting, especially with limited training data. However, performance is highly dependent on the correct choice of kernel function, kernel scale (γ), the box constraint (C), and the ε-insensitive margin.

Gaussian Process Regression offers a fully probabilistic, non-parametric approach to regression, modelling the predictive map as a distribution over possible functions, with correlations between outputs defined through a covariance (kernel) function. Rather than learning explicit parameters, GPR defines a prior over functions that is updated with training data to produce a posterior that provide both mean predictions and predictive uncertainty. This makes GPR particularly well suited for small or noisy datasets, where quantifying confidence is important. The choice of kernel governs the smoothness and

flexibility of the inferred function corresponding to different regularity classes Ck in functional space. The noise level affects the model's trust in the training data. Despite its advantages the main limitation of GPR is its computational cost which scales cubically with the number of training samples due to the inversion of large covariance matrices. In this work, we optimized the kernel type, the basis function, and the noise level to enhance both accuracy and uncertainty estimation.

Neural networks are powerful tools for learning complex and hierarchical nonlinear relationships from data. We implemented a shallow feedforward neural network with a single hidden layer, inspired by applications in vegetation biophysics (Baret et al., 2007). NN offer great flexibility but require careful selection of architecture and hyperparameters, such as the number of neurons, activation function, and regularization strength.

The performance of the different Machine Learning models was evaluated to identify the best input variable configuration for retrieving FC and LAI. Additionally, we assessed the effect of the number of spectra (Ns) used by considering a set of 13 different values within the range 500-10,000, chosen to evenly sample the range and capture both low and high sample size scenarios. Each spectrum was generated from a distinct configuration of the SCOPE input parameter set, obtained by randomly sampling the multivariate probability density function associated with their distribution. The random seed defining the number generation algorithm was fixed prior to sampling to ensure reproducibility. To evaluate the stability of the pipeline with respect to dataset variability, the experiment was repeated 11 times for each Ns, using a different fixed random seed in each repetition. In total, 143 libraries were generated, comprising approximately 800,000 spectra.

2.3.2.2 Training and testing of machine learning models

Each of the models presented in the previous section is a supervised learning model, meaning that the learning process depends on the target labels. During training, the model optimizes internal parameters to map input variables to their corresponding labels with minimal error. For each target variable (LAI and FC), separate models were trained. For this training phase, within a given spectral library $L(Ns, \theta r)$, 80% of the spectra were randomly selected for training, while the remaining 20% were reserved for testing. During testing we computed statistical metrics to assess model accuracy and explanatory power. Specifically, we used the Root Mean Square Error (RMSE), which quantifies the average magnitude of prediction errors in the same units as the target variable, and the coefficient of determination (R^2), which measures the proportion of variance in the target variable explained by the model. These are defined as: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2}$, $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{\sum_{i=1}^{N}(y_i - \mu)^2}$, where y_i is the observed value (i.e., the label provided with the training data), $\hat{y}_i$ is the model prediction, μ is the observations mean computed over the N test data.

Model performance also depends on hyperparameters which are structural characteristics of each model that should be appropriately selected. These may include the number of trees and maximum tree depth for decision tree-based methods, the learning rate and regularization terms for ensemble or kernel-based methods, the number of neurons and layers for neural networks, and other algorithm-specific parameters. To identify the optimal hyperparameter configuration, a Bayesian optimization strategy was adopted. Rather than exhaustively exploring all combinations in a predefined grid, this method treats the chosen objective function (cross-validation error) as a random function, updating its posterior iteratively based on trial outcomes. For each iteration the resulting posterior distribution is passed to an acquisition function that proposes the next configuration to evaluate. In our case the Expected Improvement Plus acquisition function was used which includes a penalization mechanism to avoid convergence to local minima. Each iteration of the hyperparameter tuning process involves training and validating the model. To achieve this, the training set is further divided using k-fold

cross-validation with randomly generated folds. This technique consists of partitioning the training data into k subsets and iteratively training the model on k-1 folds while validating on the remaining one. The average performance metric across all folds is used to update the objective function and guide the selection of the best hyperparameter configuration, reducing the risk of overfitting to a specific fold. Once the optimal configuration is found, the model is then evaluated on the test set to assess predictive performance.

2.3.2.3 Model performance assessment and robustness analysis

Models robustness was evaluated with respect to variability in both spectral simulations and randomness inherent in the training process. Therefore, each model was trained and tested on every different spectral library $L(Ns, \theta r)$. In addition, the ML algorithms themselves are influenced by internal sources of randomness (denoted by φr), which affect i) the splitting of the dataset into training and test sets, ii) the k-fold division and finally iii) the optimization procedures (both parameter and hyperparameter learning). While the final selected model uses a fixed value for Ns, θr and φr, the variability of this parameters was explored to assess model stability. It is worth noting that θr are inherently stochastic and cannot be set a priori, as doing so would interfere with the generation of random numbers.

An iterative procedure was followed to train and test the six models and considering the six input data configurations. These ranged from minimal setups using only two visible-range features (VNIR0 and VNIR1), to richer configurations progressively including angular information (SZA, VZA, RAA), the PAN channel and the NIRv vegetation index. All features were independently normalized to the [0,1] interval using min-max scaling, i.e., subtracting the minimum and dividing by the range of each feature. Prior analysis (data not shown) confirmed that model performance was not sensitive to the specific normalization method employed. For consistency and to isolate model performance from biases associated with absolute magnitudes, the target variables were

mapped to the [0,1] interval. For FC, this scaling is intrinsic by definition, whereas for LAI normalization was performed using the extreme values observed from the training set.

To evaluate the impact of feature dimensionality, a Principal Component Analysis (PCA) was implemented and tested as a dimensionality reduction method. However, due to the relatively small number of input variables, PCA did not yield significant improvements and was ultimately not adopted in the final models (data not shown), also to preserve model explainability.

2.3.2.4 Best configuration selection

To identify the optimal value of Ns, the RMSE and R^2 were analysed as a function of Ns to determine the convergence point, beyond which observed differences are attributed to stochastic fluctuations rather than genuine performance improvements. This test was repeated across varying initial conditions, i.e., for different values of the randomness parameter θr (as described in Section 2.1.1), producing a distribution of results rather than a single metric for each Ns. This approach provides statistical robustness for the assessment of model performance, ensuring that the identified optimal Ns is not an artefact of specific random initializations but reflects the true convergence behaviour of the algorithm.

For each value of Ns, this procedure yields an independent set of performance metrics corresponding to that specific number of spectra, thus forming a statistically distributed sample. To identify the smallest Ns beyond which convergence is achieved (i.e., further increases of Ns do not yield statistically significant improvements), a group comparison approach was applied, where each group corresponds to the distribution of results obtained for a given Ns. The comparisons were performed using the Kruskal–Wallis (KW) rank-sum test, suitable for independent groups with potentially non-Gaussian distributions. A p-value > 0.05 indicates no statistically significant differences between

groups, and the optimal Ns is selected as the smallest value, favouring computational efficiency. Conversely, a p-value < 0.05 implies that at least one group differs, necessitating a Dunn post-hoc test. This test uses the ranks from the KW analysis and applies a correction for multiple comparisons. By setting the significance level for each individual comparison to 0.05 divided by the number of comparisons, the overall Type I error is maintained at approximately 5%, ensuring statistically reliable results. The threshold of 0.05 was chosen as a conventional criterion for statistical significance, providing a practical balance between detecting meaningful differences and control false positives.

This same rank-based logic was extended to evaluate the compatibility between distributions, allowing assessment of whether different feature configurations yield statistically equivalent performance. For the selection of the best variable set and model, training and testing were repeated for each configuration across the different randomness parameter values. This procedure produces a statistical sample of results per configuration, which can be visualized using boxplots showing the median (central value), interquartile range (IQR, 25th–75th percentile), and potential outliers, providing a robust characterization of variability. Using the median rather than the mean ensures that extreme values have a limited influence on the summary statistics.

2.3.2.5 Uncertainty estimation

The explicit quantification of uncertainty has become a central requirement in remote sensing retrievals (Tran et al., 2023), as it enables a more reliable integration of biophysical parameters into data assimilation models, land-surface schemes, and climate analyses. The trained models provide forward estimates of the target variables, which can be treated as realizations of random variables and thus described by a probability density function (PDF). Modern retrieval frameworks are expected not only to provide a single best estimate but also to characterize the confidence of their predictions, thereby enhancing interpretability and operational usability.

For some models, prediction uncertainty is inherently provided by the model. For example, in the GPR, uncertainty is typically expressed as the standard deviation of the posterior distribution. In this case, this intrinsically estimate naturally captures both types of uncertainty considered in our analysis using synthetic data. Aleatoric uncertainty is associated with the intrinsic variability in the data. In the context of synthetic dataset, it includes instrumental noise, the variability obtained by mixing vegetation and soil spectra, and the natural variability of the simulated system not captured by the model (i.e. the random sampling of the probability distribution for each SCOPE biophysical input parameter). In contrast epistemic uncertainty originates from limitations of the model, such as structural assumption, incomplete coverage of the input parameter space and also sensitivity to initialization. It reflects the confidence in the predicted values given the model formulation and the available training data. Together, aleatoric and epistemic uncertainties provide a comprehensive characterization of the reliability of model predictions (Kendall & Gal, 2017).

For the other models, uncertainty was estimated through external procedures. Here, we exploited the bootstrap approach, which consists of repeatedly training the model on different bootstrap samples. Each bootstrap sample, is generated by randomly resampling the original dataset with replacement, so that some observation appears multiple times while others are omitted, producing a sample of the same size as the original dataset. The final prediction was then computed as the mean of the individual model outputs, while the associated uncertainty is estimated as the standard deviation across these outputs. This ensemble-based method captures only the epistemic uncertainty. To also account for the aleatoric component, we augmented the bootstrap-derived uncertainty with the empirical variance of the residual error observed on the training data. Let ŷ1 denote the prediction of the b-th bootstrap model for a given input, and let μ be the mean of the prediction across all B bootstrap models. The total uncertainty for the bootstrap prediction ($\sigma_{tot}{}^2$) is given by the sum of the epistemic and aleatoric components and expressed in Equation 4:

$$\sigma_{tot}{}^2 = \sigma_{ep}{}^2 + \sigma_{al}{}^2 = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{y}_i - \mu)^2 + \frac{1}{N-1}\sum_{i=1}^{N} r_i{}^2 \qquad (4)$$

where N is the number of data points in the original (non-bootstrapped) training set used to train a single model for computing the residuals r_i. The first term ${\sigma_{ep}}^2$ represents the epistemic uncertainty, estimated from the variability of the bootstrap predictions, while the second term ${\sigma_{al}}^2$ represents the aleatoric uncertainty, estimated from the residuals of the model trained on the full dataset.

Evaluating the calibration of predicted uncertainties is essential to determine whether the model's estimated confidence accurately reflects the true prediction errors. In other words, a well-calibrated uncertainty indicates that the predicted standard deviation corresponds to the observed variability between model outputs and reference values. To assess this calibration, we computed a normalized residual (t-statistic) for each prediction, defined as the residual divided by the estimated standard deviation. The resulting t-values were analysed by plotting histograms and fitting several candidate distributions, namely Gaussian, t-Student, and Laplace distributions. A standard deviation of approximately 1 in the fitted t-Student distribution indicates that the estimated uncertainties are well calibrated, providing a stable and reliable measure of predictive error. Deviations from this value highlight potential under- or overestimation of uncertainty. In summary, while the intrinsic GPR uncertainty corresponds to the standard deviation of a Gaussian process, the bootstrap-derived uncertainty does not have a known probability density function a priori. However, if the number of bootstrap repetitions is sufficiently large, the Central Limit Theorem ensures that the distribution of predictions approaches a normal distribution. This justifies the use of the t-test, to evaluate prediction accuracy and compute confidence intervals.

statistics.

2.3.3 Application to real data

2.3.3.1 Sentinel-2 images and GBOV dataset

Once the best-performing model was selected based on the simulated data, it was subsequently applied to Sentinel-2 images and the estimates validated using the Copernicus GBOV service.

Given that the SBG-TIR mission is still in the preparatory phase we selected a satellite mission offering spectral characteristics that match as close as possible to those used during training. The Sentinel-2 mission, developed and operated by the European Space Agency (ESA) within the Copernicus Programme, provides globally consistent multispectral reflectance data observations, making it particularly suitable for emulating its observational configuration of SBG. In particular, the VNIR0 and VNIR1 bands of the VIREO camera were spectrally resampled with Sentinel-2 bands B4 (RED, 665 nm) and B8 (NIR, 842 nm), respectively, with a spatial resolution of 10m. The panchromatic band which plays a key role in our model, due to its coverage of the red-edge region, was reconstructed by combining multiple S2 bands within the relevant spectral domain. To derive the synthetic panchromatic band, we adopted a linear approximation method. Each Sentinel-2 band was represented by a Gaussian response function using the central wavelength and full-width at half-maximum (FWHM). A discrete convolution between the Sentinel-2 response functions and the VIREO PAN spectral ISRF was performed to compute the weighted contribution of each band. The resulting synthetic PAN band was generated at a spatial resolution of 20 m, corresponding to the lowest native resolution of the input Sentinel-2 bands.

GBOV aims to develop and distribute robust in-situ datasets for systematic and quantitative validation of Earth Observation (EO) land products. The dataset contains in-situ reference measurements of vegetation biophysical parameters including LAI and FC data, with associated uncertainty. The selected dataset used in this study includes approximately 5,000 measurements collected from 20 experimental sites across Europe, North America, and Australia, covering a wide range of soil types, vegetation covers, and different climatic zones. Each GBOV record is accompanied by precise spatial and temporal metadata, enabling direct spatial-temporal matching with satellite observations.

For each ground measurement we extracted the corresponding Sentinel-2 (S2) images (from both platforms A and B) that included the point of interest and acquired within a ±3 days window around the GBOV measurement date. Both single-pixel extraction and neighborhood-based approaches (3×3 and 5×5-pixel buffers) were evaluated to mitigate the impact of sensor noise and sub-pixel heterogeneity.

A brief characterization of the experimental dataset is presented in Figure 4. The first panel shows the geographical distribution of the experimental sites, located across three different continents and covering multiple land cover classes, each represented with varying frequency. The two target variables, LAI and FC, are both well-defined within their theoretical ranges and each measurement is represented with the corresponding uncertainty value, whose distributions are illustrated in the histograms in Figure 4. In relative terms, uncertainty to measurement ratio is below 20% respectively for 90% of FC measurements and for 99% of LAI measurements.

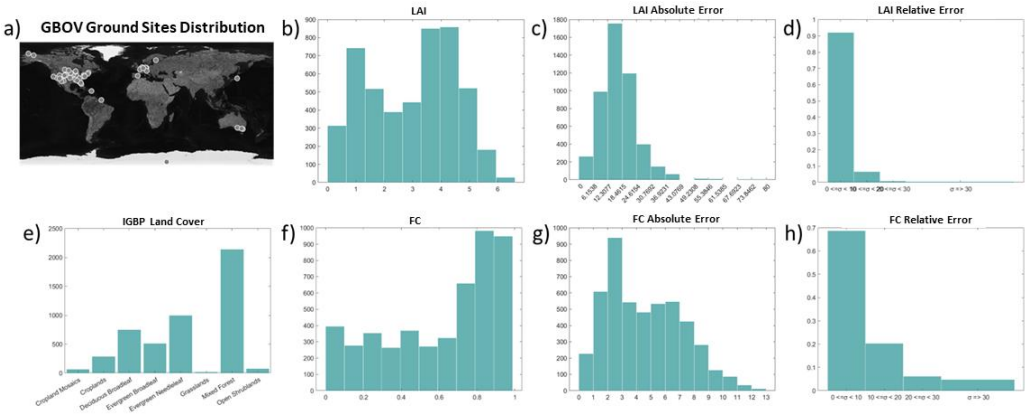

Figure 4: The first panel (a) reports the geographical distribution of the experimental sites together with the associated land cover classes (e). Adjacent histograms illustrate the distributions of the two target variables, LAI (b) and FC (f), along with their corresponding uncertainty (c, g) and relative error distributions (d, h).

We used 248 GBOV ground observation extracted from 98 different Sentinel-2 images, collected over six years and covering five distinct and representative land cover types.

This setup allows for a robust and representative evaluation of model performance under realistic observational conditions.

In summary, for each GBOV sample, a feature vector was constructed comprising three spectral bands (B4, B8, PAN), the NIRv vegetation index, and three solar and viewing variables (view zenith angle, sun zenith angle, relative azimuth). The dataset was then used as input to the trained models in predictive mode generating LAI and FC under realistic observation geometries. The predictions were subsequently compared with the GBOV ground-truth and performance metrics including RMSE and R^2 were computed.

2.3.3.2 Comparison with traditional approach

In addition to our ML-based inversion, we applied the well-established benchmark model, the Sentinel-2 Biophysical Processor from ESA's SNAP toolbox (Weiss and Baret, 2016). This processor relies on a neural network trained on PROSAIL radiative transfer simulations and requires the full set of Sentinel-2 spectral bands as input. Each biophysical variable (LAI, FC, etc.) is predicted by the same neural architecture but optimized with variable-specific weights.

Regarding the inputs, Sentinel-2 provides spectral bands at native spatial resolutions of 10, 20, or 60 meters. Two versions of the processor are available: one exploiting only the 10 m bands and another operating at 20 m resolution, which also includes resampled version of some 10 m bands. For our comparison, we selected the 20 m version, as it provides richer spectral information and includes the bands used to simulate the panchromatic channel in our dataset.

The comparison between our approach and the SNAP-based estimates was performed over the same spatial and temporal domains and conditions, ensuring a consistent benchmarking framework. Model performance was evaluated as prediction accuracy, by using the same error metrics (RMSE and R^2) used during the best model selection phase but computed between model predictions and GBOV ground-truth labels.

In addition, the analysis was conducted for different land-cover types to assess model performance across diverse vegetation classes. Several regions of interest (ROIs) of varying sizes were tested to examine the influence of spatial representativeness on prediction accuracy.

2.4. RESULTS AND DISCUSSION

2.4.1 Simulated dataset

Each SCOPE simulation required a set of input parameters sampled from marginal or joint distributions, as described in Table 1. Figure 5, show an example of the SCOPE simulated reflectance spectra for varying SZA and VZA for different combinations of LAI and chlorophyll content and the joint probability distribution between LAI and Cab, indicating the realistic combination of the vegetation parameters avoiding implausible combinations.

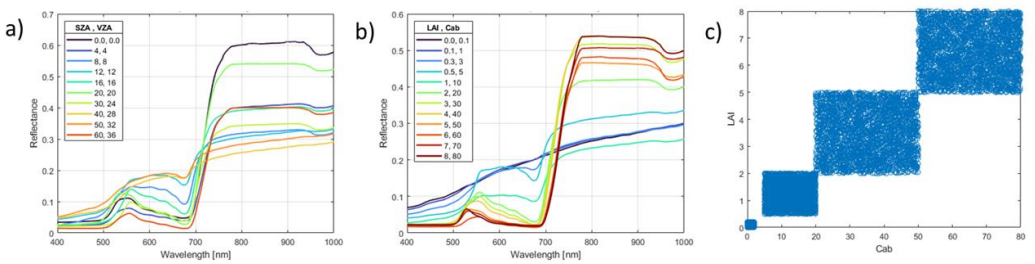

Figure 5. a): reflectance spectra for varying Solar Zenith Angle (SZA) and View Zenith Angle (VZA). b): spectra for different combinations of LAI and chlorophyll content (Cab). c): distribution of Cab and LAI parameter pairs.

Figure 5a shows the depth of the characteristic chlorophyll absorption feature near 670 nm varies with geometry, reflecting the combined effects of leaf orientation, shadowing, and multiple scattering in the canopy and no strictly monotonic trend is observed. The

spectral red-edge slope also varies with geometry without following a clearly monotonic pattern. The effect of progressively increasing both LAI and Cab content is shown in the b panel. At low Cab and LAI values, spectra exhibit soil-like reflectance with minimal absorption features. As Cab and LAI increase, the chlorophyll absorption at 670 nm becomes deeper, followed by a sharp rise in reflectance at the red-edge region (~700 nm). In the intermediate range (LAI: 1-4; Cab: 10-40), the behaviour becomes more vegetation-like, with reflectance spectra showing a progressive increase in red-edge amplitude and a slight blueshift of the green reflectance peak (~520 nm) toward shorter wavelengths, consistent with reported behaviour in the literature (e.g., Garcia Martinez et al., 2025). For Cab values above approximately 40 μg cm^{-2}, absorption in the red region approach saturation, and further increases primarily enhance red-edge reflectance.

Alternative methodologies employ three-dimensional (3D) radiative transfer models (RTM), such as DART (Discrete Anisotropic Radiative Transfer; Gastellu-Etchegorry et al., 2010), which are capable of representing complex canopy architectures not considered in the present study. However, the increased structural realism of 3D models comes at the cost of significantly higher computational demands and the need for detailed ground-based measurements of both optical and structural vegetation properties (Meroni et al., 2004). Furthermore, under conditions of horizontal homogeneity, the outputs of 3D models tend to converge toward those produced by one-dimensional (1D) models (Banskota et al., 2015). For these reasons, a 1D radiative transfer approximation was adopted in this study, offering computational efficiency and greater control over input parameters and scene configuration, as defined by the radiative transfer equations. Additionally, surface elements were assumed to behave as Lambertian reflectors, implying reflectance independence from viewing geometry. This assumption allows for the exclusion of explicit Bidirectional Reflectance Distribution Function (BRDF) effects, in accordance with the SCOPE 1D radiative transfer formalism. While this simplification limits the representation of angular reflectance variability, specifically by precluding the explicit simulation of leaf-level anisotropy, clumping effects, and hotspot phenomena, it is widely adopted in canopy-scale RTMs, as it enables computationally efficient model

inversion and robust retrieval of biophysical parameters (Verhoef, 1984; Jacquemoud et al., 2009).

When the unmixing model is applied to the previously simulated spectra, the resulting mixed spectra is shown in Figure 6. When vCover is minimum (Figure 6a) the resulting spectrum closely resembles a bare soil spectrum. For intermediate values (Figure 6b) both vegetation and soil components shape the spectral signature, producing a mixture of features such as partial absorption in the chlorophyll bands (~670 nm) and a subtle red-edge slope (~700 nm). Finally, when vCover is high (Figure 6c), the vegetation component dominates, and the spectrum exhibits the characteristic chlorophyll absorption and prominent red-edge transition associated with a fully vegetated canopy.

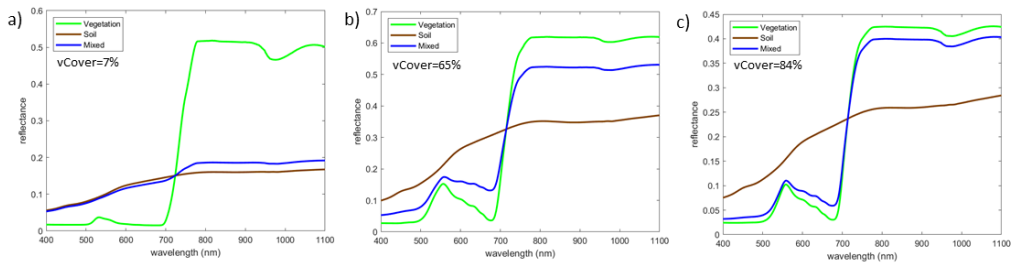

Figure 6. Simulated mixed reflectance spectra obtained using a linear mixing model of vegetation and soil components. The parameter vCover represents the fractional contribution of vegetation in the mixture. a): the resulting spectrum is dominated by soil reflectance. b): a mixed spectrum in which both vegetation and soil contribute significantly. c): the spectrum is nearly fully vegetated.

Overall, these spectra show a smooth transition from soil-dominated to vegetation-dominated reflectance, capturing the variability that can occur in a single pixel in real landscapes. In this context, the intermediate cases are particularly important, as they represent the most common conditions in natural environments and provide valuable examples for training machine learning algorithms. The variability in the underlying soil

spectra generated here using the BSM model, adds further realism to the mixed spectra, especially at low vCover values. At high vCover, or for high Cab and LAI the saturation of the red-edge and absorption features became evident, highlight non-linearity that must be considered when retrieving FC and LAI from reflectance data. Overall, these simulations provide a comprehensive spectral dataset that reflects both the vegetation fraction and soil heterogeneity.

The multiplicative noise applied to the simulated spectra to account for measurement variability produced the final spectra showed in Figure 7, that have been resampled in SBG configuration and used as input for the inversion.

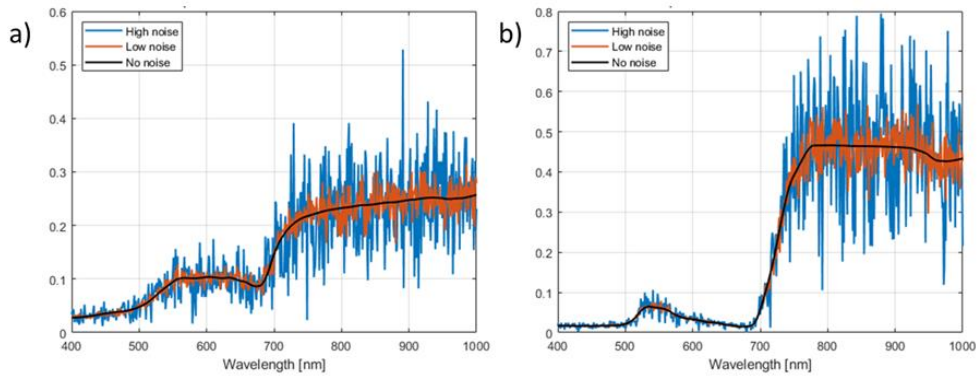

Figure 7. Simulated canopy reflectance spectra for two distinct base reflectance cases, each with three noise levels: Low noise corresponds to a signal-to-noise ratio (SNR) of 100. High noise corresponds to SNR = 10.

From Figure 7, it is evident that low noise (SNR = 100) slightly perturbs the spectra without altering their overall shape. Variations are mainly visible in regions of high reflectance, such as the red-edge, where multiplicative noise produces modest oscillations around the underlying trend. High noise (SNR = 10) induces larger fluctuations; however, the overall spectral pattern remains conserved. Importantly, these high-frequency fluctuations are subsequently mitigated by the spectral convolution

applied to derive sensor-level bands (see Section 2.3.1.5), which acts as a smoothing operation by averaging the signal within each band. As a result, local noise contributions are reduced while the main spectral features are preserved. Consequently, even under high-noise conditions, trends such as the increase in reflectance in the red-edge region with higher chlorophyll content remain discernible.

Adding noise ensures that the simulated dataset better represents realistic measurement variability, improving the generalization of retrieval algorithms trained on these spectra while preserving the integrity of the spectral information. In this context, noise accounts for intrinsic variability in the observed signal, arising both from instrumental limitations and from natural fluctuations in the interaction between radiation and vegetation elements, which cannot be fully captured by deterministic radiative transfer models (Van Der Tol et al., 2009).

2.4.2 Machine Learning Inversion

In this section, we present the results of the machine learning model performance analyses. As described in Section 2.1.1, a total of 143 distinct spectral libraries were available for training, differing in random seed initialization (θr, 11 values) and the number of spectra (Ns, 13 values). Each spectrum was pre-processed and associated with seven predictive variables (VNIR0, VNIR1, SZA, VZA, RAA, PAN, NIRv) and two target variables (FC and LAI), which constitute the input-output pairs for the models. For each training run, one of the two targets was selected, and a single spectral library was used as input.

Several factors were systematically varied and optimized during training. Specifically, six different machine learning models were tested, along with six different combinations of predictive variables and three levels of added noise. The stability of each model was also evaluated with respect to the random seed to account for the intrinsic stochasticity of the training process.

2.4.2.1 Sample Size Sensitivity Analysis

The first analysis aimed to investigate how the number of samples in the training dataset affects retrieval accuracy across all methods. The initial dataset consists of Ns spectra, of which 80% are used for training and the remaining 20% for testing model performance. Ns values ranged from 500 to 3,000 with steps of 500, and from 4,000 to 10,000 with steps of 1,000. As presented in the method section, for each Ns, six different models were trained and evaluated by computing RMSE and R^2. Only the full set of variables was considered in this analysis, and no additional noise was added to the training spectra. Figure 8 shows the plots of the evaluation metrics as a function of dataset size.

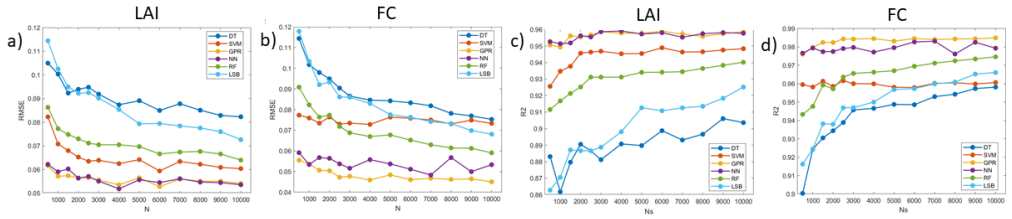

Figure 8: Performance metrics as a function of the number of spectra (Ns). a and b panels shows RMSE for LAI and FC respectively, while c and d show R^2 for LAI and FC, respectively. Results are shown for all selected methods, without added noise to the reflectance. Each point represents the median of results obtained from models trained on 11 different training set.

Each point corresponds to the median value obtained by varying the random seed θr described in section 2.2.3 and associated with randomness in dataset generation, which effectively corresponds to a dataset of Ns different spectra. The use of the median ensures robustness against outliers.

The results indicate an almost constant trend of the metrics with increasing Ns, particularly for the two best-performing methods (GPR and NN). For all methods, performance is lower for Ns below 2,000. Smaller then Ns = 500 datasets were excluded due to poor generalization performance. Overall, a rapid convergence of metrics is observed as Ns increases.

The Kruskal-Wallis rank-sum test was applied to determine the optimal Ns on the GPR and NN models. This non-parametric approach is appropriate because 11 independent measurements per performance metric were available for each Ns, and no assumptions were made regarding the underlying distribution of data. The KW test returned a p-value < 0.05 indicating significant differences among the groups, and subsequent Dunn post-hoc tests revealed that the first convergence point of Ns varies depending on both the model type and the evaluation metric, ranging between 1,000 and 6,000. To ensure consistency across models and to adopt a conservative approach that avoids potential bias from data selection, Ns = 6,000 was chosen as a conservative training size for all subsequent analyses.

For numerical evaluation, the interquartile range (IQR) was associated with each metric. Reported values are consistently small, indicating stable predictions that are largely independent of the specific library spectra. In the noise-free case, the numerical metrics are as follows: for FC, RMSE = 0.046 (IQR = 0.002) and R^2 = 0.98 (IQR = 0.001); for LAI, the best performance is RMSE = 0.053 (IQR = 0.005) and R^2 = 0.95 (IQR = 0.001), both obtained with the GPR model.

For SNR = 100, results are nearly identical, with GPR again providing the best performance for FC, while LAI exhibits a slightly larger IQR (0.005 for both RMSE and R^2). Under high noise conditions (SNR = 10), FC results are RMSE = 0.056 (IQR = 0.003) and R^2 = 0.977 (IQR = 0.002), while for LAI, the best RMSE = 0.063 (IQR = 0.006) and R^2 = 0.943 (IQR = 0.006).

Overall, while GPR provides slightly better median performance, the differences with NN are minimal and statistically comparable. Similarly, different feature configurations produce results close to those reported here.

2.4.2.2 Variable selection and preliminary input data definition

To further assess the stability of the results, boxplots of RMSE and R^2 for both GPR and NN performance were generated for different feature subsets (Figure 9). As in the previous section, the boxplots were obtained by varying the random seed associated with the training set generation. For this analysis, we restricted the comparison to the two best-performing models, GPR and NN. From Figure 9, it is evident that some feature configurations provide better statistical performance than others. Importantly, this conclusion does not depend on the metric considered (RMSE or R2), but the best feature selection strongly depends on the specific target variable.

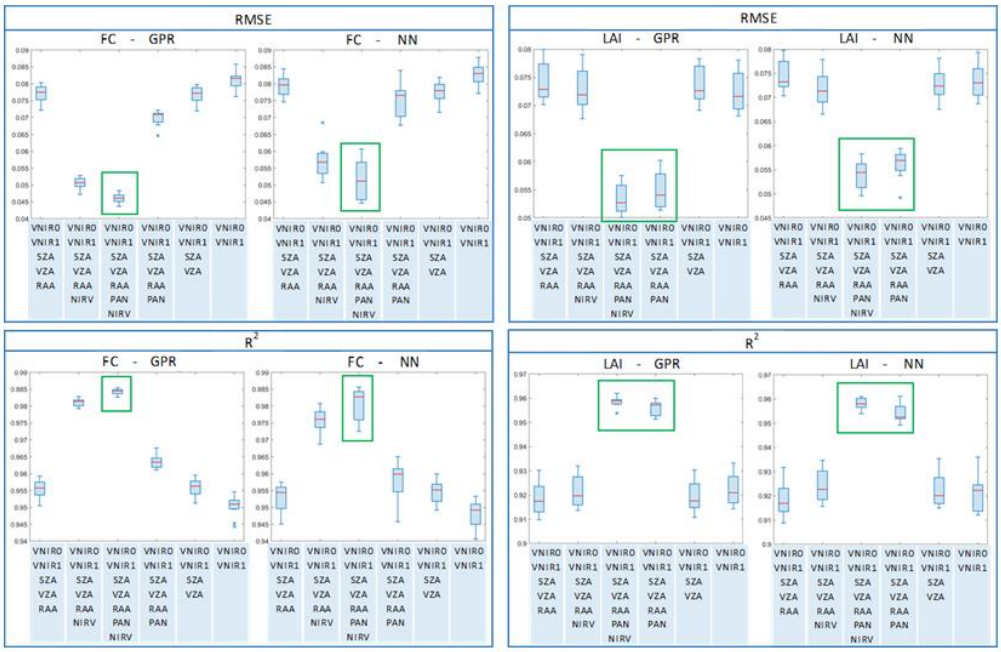

Figure 9: Boxplots of performance metrics summarizing results from 11 different training sets, used to evaluate the impact of training set variability on model stability. The x-axis shows different feature subsets, starting with the baseline configuration that includes the two spectral variables (VNIR0 and VNIR1) plus three geometric variables (VZA, SZA, RAA), shown as the first configuration in each plot. The variables are then

added or removed to assess feature selection effects. Four pairs of plots are presented, each pair comparing GPR (left) and NN (right) models. The top pairs show RMSE, while the bottom pairs show R^2; FC is displayed on the two left pairs, LAI on the right ones. The best-performing configuration from a statistical perspective are highlighted in green.

In both cases, results obtained with the baseline configuration (RED, NIR + SZA, VZA, RAA) already provides satisfactory results. Adding further variables improves performance with the effect becoming particularly evident when introducing the PAN band for LAI estimation and the NIRv index for FC. The best overall performance is obtained when both variables are used together, confirming their complementary contribution to model accuracy.

When variables are removed no significant degradation is observed when excluding the mutual viewing geometry (RAA), while zenith angles (SZA and VZA) leads to a slight decrease in performance. This can be explained by the fact that the spectral variables (RED and NIR) already contain most of the information related to the biophysical parameters, although geometric variables (particularly the zenith angles), still play a supportive role. Their inclusion remains advisable, especially when dealing with real imagery, where illumination and observation conditions vary spatially (e.g. Verrelst et al., 2008; Petri et al., 2019).

From a physical perspective, the structural properties of vegetation are well represented by the red and near-infrared regions of the spectrum. However, the inclusion of soil reflectance through the mixed-pixel approach is required to simulate the spatial variability within real pixels. This introduces additional spectral variability which can complicate the retrieval process. Introducing a variable such as NIRv, although partially redundant from a spectral point of view, improves model performance for FC. This improvement can be attributed to NIRv ability to isolate the signal of photosynthetically active vegetation (Badgley et al., 2017) by partially removing background and

illumination effects (Gitelson et al., 2018), allowing the model to better discriminate vegetation from soil contributions.
Conversely, for LAI estimation, the PAN band plays a key role. By covering a broad spectral region including the red-edge, it captures subtle variations in canopy chlorophyll and structure, enhancing sensitivity to canopy density and layering. This improves LAI retrieval, which is intrinsically linked to these structural properties (Delegido et al., 2011).

The difference between FC and LAI in their sensitivity to PAN can also be explained by the way FC is derived. Since FC is derived as $FC=1-e^{-k.LAI}$, it is effectively a proxy of LAI and saturates for dense canopies (Carlson & Ripley, 1997). As a result, while FC tends to saturate and becomes less sensitive to canopy structure, LAI remains responsive to subtle variations in canopy layering making PAN derived structural information, critical for accurately modelling LAI variations in more heterogeneous canopy structures.

To select the optimal features configuration, the statistical KW test was performed between the two set configurations that yielded the best performance in predicting LAI. The results indicate that, choosing one subset over the other leads to statistically equivalent outcomes. For FC however, a difference is observed: the improvement from baseline + NIRv to baseline + NIRv + PAN is statistically significant.

We selected this latter configuration for both variables, as it provides the best performance for FC and is among the top two configurations for LAI. Moreover, when comparing median values, it still yields the best results, even though, from a statistical standpoint, the difference cannot be asserted within a type I error threshold (i.e., with less than a 5% probability of being wrong). It is also worth to remark that all the variables are available from the mission dataset.

This remains true across all SNR values and for both models. Some numerical results obtained with the full features configuration, are reported in Table 2 for both the noise-free case ($SNR\ Inf\ = \infty$), and the high-noise case (SNR 10). In the latter case (not shown as boxplots), the two methods perform equivalently. From the perspective of model performance, GPR appears more stable than the NN algorithm in terms of IQR when predicting FC, whereas no significant differences are observed in the IQR for LAI prediction.

2.4.2.3 Stability analysis and best model selection

Two models, GPR and NN, were selected for further analysis because they exhibited comparable performance. To assess the stability of the results, the training and testing procedure was repeated while keeping the dataset fixed. During these repetitions the parameter φr (described in Section 2.2.3), was varied; this parameter controls the randomness in the model training process, in contrast to θr, which governs the randomness in dataset construction and is kept constant in this analysis.

SNR Inf	RMSE %	DT	GPR	LSB	NN	RF	SVM
	FC	8.33	4.61	7.63	5.11	6.52	7.58
	LAI	8.49	5.27	7.94	5.44	6.66	5.93
	R2	DT	GPR	LSB	NN	RF	SVM
	FC	0.949	0.985	0.957	0.983	0.969	0.958
	LAI	0.899	0.959	0.911	0.958	0.934	0.949

IQR %	DT	GPR	LSB	NN	RF	SVM
FC	0.17	0.20	0.22	1.12	0.28	0.34
LAI	0.48	0.46	0.27	0.49	0.42	0.58
IQR	DT	GPR	LSB	NN	RF	SVM
FC	0.004	0.001	0.004	0.008	0.002	0.004
LAI	0.005	0.002	0.006	0.004	0.006	0.004

SNR 10	RMSE %	DT	GPR	LSB	NN	RF	SVM
	FC	8.82	5.58	8.17	5.58	6.86	7.91
	LAI	8.81	6.35	8.69	6.34	6.86	7.05
	R2	DT	GPR	LSB	NN	RF	SVM
	FC	0.949	0.985	0.957	0.983	0.969	0.958
	LAI	0.899	0.959	0.911	0.958	0.934	0.949

IQR %	DT	GPR	LSB	NN	RF	SVM
FC	0.25	0.27	0.70	0.44	0.28	0.29
LAI	0.64	0.49	0.51	0.54	0.52	0.63
IQR	DT	GPR	LSB	NN	RF	SVM
FC	0.004	0.001	0.004	0.008	0.002	0.004
LAI	0.005	0.002	0.006	0.004	0.006	0.004

Table 2: Test results for both RMSE and R² under two different SNR conditions. For each model the table reports the best median performance along with the interquartile range, corresponding to the best subset of features configuration. RMSE values are expressed as absolute percentages to enhance readability.

The boxplots obtained for different feature configurations are shown in Figure 7. The plots report the outcomes for both GPR and NN models under the previously selected feature configuration. GPR exhibits marked stability, whereas NN performance is strongly affected by randomness, resulting in less consistent outcomes that are highly dependent on the initial conditions. This behaviour holds across different metrics and also when spectral noise is introduced (SNR = 10, not shown). Therefore, GPR was selected as the best-performing model, as it provides more stable results. Importantly, GPR not only delivers more stable predictions but also provides intrinsic estimates of uncertainty, which are more robust than the approach used for NN (see next Section 2.4.2.4.) This combination of stability and uncertainty quantification represents a key methodological advantage, further justifying the selection of GPR as the preferred modeling approach.

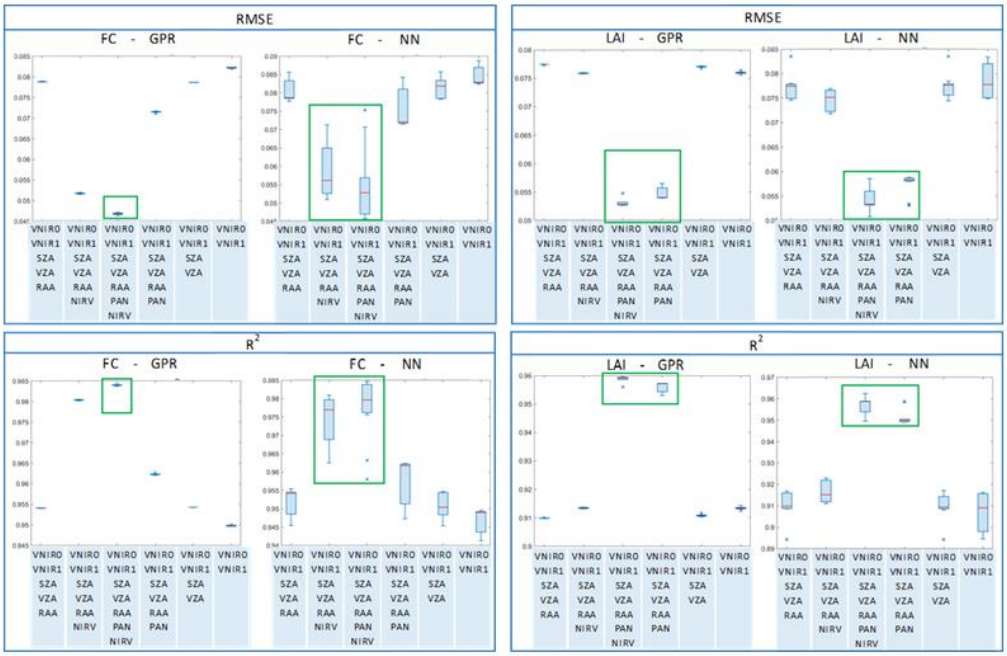

Figure 10: Boxplots of performance metrics obtained by training the selected model 11 times with different initial random conditions, used to evaluate the stability of the results with respect to the training procedure. As in figure 9, the x-axis shows different feature subsets, starting with the baseline configuration that includes the two spectral variables (VNIR0 and VNIR1) plus three geometric variables (VZA, SZA, RAA), shown as the first configuration in each plot. The variables are then added or removed to assess feature selection effects. Four pairs of plots are presented, comparing GPR (left) and NN (right) models. The top pairs show RMSE, while the bottom pairs show R^2; FC is displayed on the left, LAI on the right. The best-performing configurations from a statistical perspective are highlighted in green.

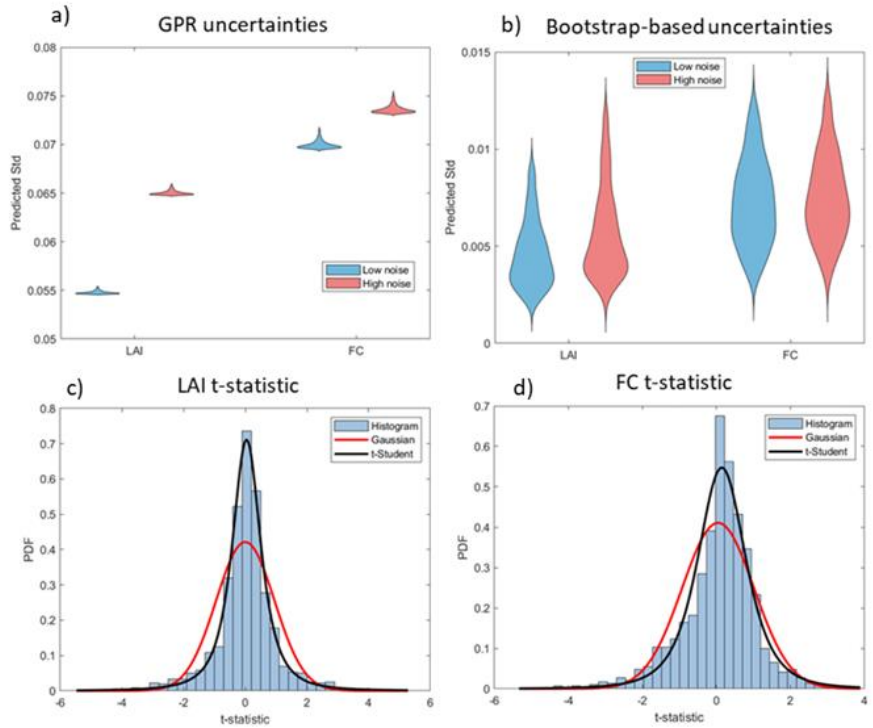

Figure 11: Uncertainty related plots. The two top panels (a &b) show violin plots of the estimated uncertainties over the test set: GPR-based on the left(a), bootstrap-based on the right(b). For both LAI and FC, the distributions are stable around the mode (with respect to labels normalized between 0 and 1), and results are reported for two different SNR conditions (low, SNR=100 and high, SNR=10). The bottom panels show histograms

of the t-statistics computed from the GPR-based residuals of the test set: c) LAI on the left, d) FC on the right. Gaussian and t-student best fitting distributions are shown.

2.4.2.4 Uncertainty analysis

To quantify prediction uncertainty, we evaluated two different approaches for the selected GPR model, as described in section 2.2.4: bootstrap derived uncertainty and the intrinsic uncertainty. Figure 11a shows the violin plot of the uncertainty distributions obtained using the intrinsic GPR method. These plots combine a boxplot with a density estimation, providing a clear visualization of both the central tendency and the overall spread of the data. These distributions show a well-defined mode around a stable value (the horizontal line), with only a few outliers. For FC, the uncertainty is slightly higher than for LAI, and in both cases, it increases when using the high-noise dataset. Figure 11b shows that the uncertainty distributions obtained from bootstrap based approach appear less concentrated around the mode and exhibit longer tails. Noise tends to further stretch the tails, although without producing significant visible effects. As expected, the absolute magnitude of uncertainty from the bootstrap is much smaller than the intrinsic estimates provided by GPR. This reflects the fact that GPR is highly stable, leading the bootstrap procedure to underestimate the effective confidence interval. Recent studies have also confirmed GPR's intrinsic uncertainty as the benchmark method for this type of analysis (Garcia-Soria et al., 2024).

Figure 11c and 11d show the distributions of the t-statistic. Both histograms were fitted to several candidate distributions, and in both cases, they were found to be well described by rescaled Student's t distributions, with p-values of 0.56 and 0.47 for LAI and FC, respectively. When bootstrap-based uncertainties are used, the resulting t-values are excessively large compared to those obtained with GPR, confirming that bootstrap variance underestimates the effective uncertainty of the model. In contrast, the intrinsic GPR uncertainties yield residuals normalized with standard deviation close to 1,

providing a more conservative and slightly overestimated measure of uncertainty, consistent with the residual spread.

The analysis of the mean values highlights different behaviours for the two target variables. For LAI, the mean of the t-distribution deviates from zero by less than 5%, indicating that uncertainties are well calibrated relative to the residuals. For FC, residuals themselves are unbiased and centred on zero, but the normalized residuals (t-values) show a shifted mean. This suggests that while the predictions are unbiased, the estimated uncertainties are not perfectly calibrated, leading to systematic deviations in the normalized error distribution. From a numerical perspective, the t-statistic can be viewed as a rescaling of the residuals by an uncertainty value that remains approximately constant across predictions. Therefore, the bias observed in the FC t-distribution arises from a predominance of slightly overestimated predictions, as also evident in the scatterplot (Figure 11d). The longer left tail corresponds to a small number of underestimated predictions. Nevertheless, the variance of the FC t-distribution is closer to 1 than that of LAI, which can be interpreted as a smaller degree of uncertainty overestimation.

In summary the uncertainties estimated by GPR provide a reliable and conservative measure of predictive error, with good calibration for LAI and a slight overestimation for FC. Bootstrap-based estimates tend to underestimate the model uncertainty, therefore the intrinsic GPR uncertainties will be used in the final product.

2.4.2.5 Hyperparameters optimization

The optimal hyperparameter configuration identified through the iterative Bayesian cross-validation strategy are reported in Table 3. For NN, the selected hyperparameters varied slightly across repetition, due to the non-convexity of the hyperparameters space and high sensitivity to initial conditions. In contrast GPR showed highly consistent hyperparameter selection, as the marginal likelihood surface is relatively smooth. This resulted in minimal variability in model performance across repetitions and noise levels.

Overall, the hyperparameter optimization ensured that each model was evaluated under its best configuration, allowing a fair comparison of predictive accuracy and stability across both models and feature subsets. Additionally, the Bayesian optimization approach mitigates the risk of overfitting by efficiently exploring the hyperparameter space, leading to robust model performance across varying data subsets and noise conditions. The observed differences in stability between NN and GPR highlight the importance of model-specific hyperparameter tuning, especially for models sensitive to initial conditions.

Algorithm	Hyperparameter	Range Values	Best - LAI - No Noise	Best - FC - No Noise	Best - LAI - High Noise	Best - FC - High Noise
Tree	Min Leaf Size	[1, 50]	44	6	25	6
	Max Depth	[10, 500]	234	414	72	438
SVM	Box Constraint - C	[1e-3, 1e3]	0.0012596	0.0012596	0.0012596	0.0012596
	Kernel Function	{gaussian, linear, polynomial}	linear	linear	linear	linear
	Kernel Scale	[1e-3, 1e2]	0.0012173	0.0012173	0.0012173	0.0012173
	Epsilon	[1e-3, 1]	0.014517	0.014517	0.014517	0.014517
GPR	Basis Function	{constant, linear, quadratic}	constant	constant	constant	constant
	Kernel Function	{squaredexp, mat3/2, mat5/2}	squaredexp	squaredexp	squaredexp	squaredexp
	Sigma	[0.0030142, 1]	0.1876	0.1831	0.1876	0.063197
NN	Number of Neurons	[5, 100]	66	39	29	69
	Number of Layers	[1,3]	1	1	1	1
	Lambda	[1e-5, 1e-1]	2.12E-05	2.13E-05	2.15E-05	1.23E-05
	ActivationFunction	{relu, tanh, sigmoid}	relu	relu	relu	relu
Ensemble_RF	Number of Trees	[10, 500]	485	18	464	17
	Min Leaf Size	[1, 50]	1	6	1	1
	Subsample fraction	[1, 7]	5	5	4	5
Ensemble_LSBoost	Number of Trees	[50, 300]	124	124	124	124
	Min Leaf Size	[1, 50]	27	27	27	27
	Learning rate	[1e-3, 1]	0.19523	0.19523	0.19523	0.19523
	NumVariablesToSample	[1, 7]	6	6	6	6

Table 3: Hyperparameter configurations yielding the best performance for each model. Two SNR condition, namely no noise and high noise (i.e. SNR = 10), are reported. Values were obtained through Bayesian hyperparameter optimization, with k-fold cross-validation used to iteratively update the objective function.

2.4.2.6 Best-case results and benchmarking

Before being applied to real satellite data, the best configurations identified were first evaluated on the independent test set to assess its generalization performance. The best configuration as described in the previous sections, corresponds to the GPR model with the hyperparameters listed in Table 3. The training, validation, and testing phases were

performed using a library of Ns = 6,000 simulated spectra generated with the SCOPE model. All available variables were retained after the feature selection process. Consequently, in addition to the baseline feature set (RED, NIR, SZA, VZA, RAA), the PAN and NIRv indices were also included. When the model operates in predictive mode, input data must provide all these variables. Two independent algorithms were trained, one for each target variable (LAI and FC).

As shown in Section 3.2.1, results for the low-noise and no-noise cases were not significantly different. Therefore, the low-noise setup was selected for application to real data, since the inclusion of moderate noise slightly increased variability in the training data and improved the model's ability to represent the irreducible (aleatoric) uncertainty, thereby reducing overfitting and enhancing robustness.

The two random parameters respectively controlling the initialization of the simulated library generation and the model training, were randomly drawn once and then fixed. Otherwise, different random seeds could lead to variations in the learned parameters that would compromise the statistical validity and generalization capability of the model. Consequently, the resulting models represent those with the highest probability of achieving optimal performance metrics, as discussed in Sections 3.2.2 and 3.2.3, rather than the absolute best values obtained by post hoc selection.

Figure 12 shows the results obtained for the four selected models. In the low-noise scenario (SNR = 100), the test-phase performance metrics were an RMSE of 0.052 and R^2 = 0.95 for LAI, and an RMSE of 0.046 and R^2 = 0.98 for FC. Under high-noise conditions (SNR = 10), the corresponding results increased slightly, with an RMSE of 0.063 and R2 = 0.93 for LAI, and an RMSE of 0.054 and R^2 = 0.98 for FC.

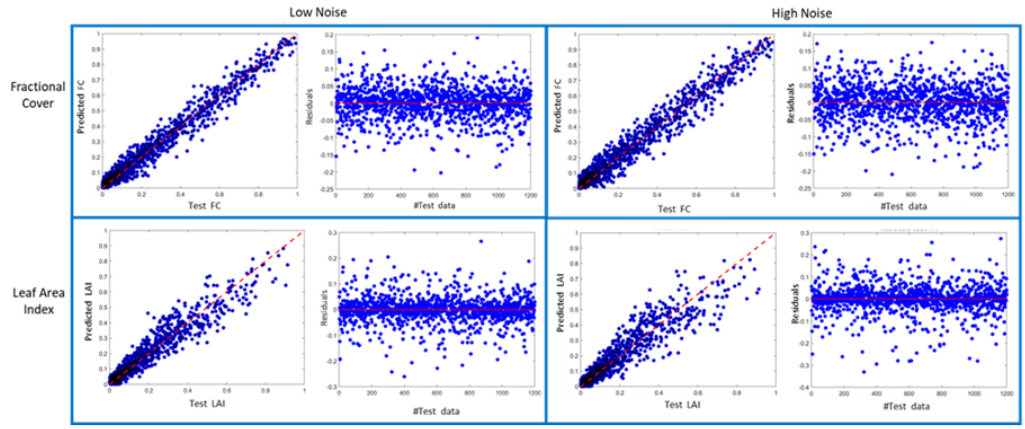

Figure 12: Application of the best model to the unseen test set. Scatter plots and residual are presented for FC (top) and LAI (bottom), with low noise and high noise conditions shown on the left and right side, respectively.

These results indicate that FC predictions are slightly more accurate than those for LAI, and that while high noise slightly affects performance, very high R^2 values are maintained for both variables. This behaviour is consistent with results reported in previous studies. The obtained validation metrics are in line with those in the literature. It should be noted that LAI values were normalized to the [0,1] range, as described in Section 2, to isolate model performance from biases associated with absolute LAI magnitudes. When applied to real data, predictions are rescaled using the inverse of the normalization relation. For comparison with other works, RMSE values were expressed as percentages. In some cases, where other studies did not follow this normalization procedure, their results have been rescaled using our relation and the minimum and maximum values reported in the corresponding references to enable a clear and consistent comparison across methods.

In Weiss et al. (2020), the model was trained on PROSAIL-simulated data using Sentinel-2A/B spectral response functions. Their best results were achieved with a neural network (one per target variable) using 8 spectral bands and 3 geometric variables as inputs. Averaging the Sentinel-2A/B results, $R^2 = 0.98$ for FC (identical to ours), and R^2

= 0.82 for LAI, slightly lower than our model. Their RMSEs are comparable: RMSE(FC) = 0.041 and RMSE(LAI) = 0.060 (after normalization to [0,1] with reported LAI max = 15).

Similarly, in García-Haro et al. (2018), the model was trained on PROSAIL-simulated data using the spectral response functions of the AVHRR optical channels. Their best model, a multi-output GPR, achieved R^2>0.88 across variables, in agreement with our results. Their reported RMSE(FC) = 0.043, slightly better than ours, while RMSE(LAI) = 0.081, slightly worse (already normalized).

Overall, performances are comparable across models, with very similar accuracy levels. The slightly lower LAI performance in García-Haro et al. (2018) may be attributed to the spectral configuration. Our setup includes one band in the red, one in the NIR, and one covering both plus the red-edge region, whereas García-Haro's configuration uses three different channels: one in the red, one spanning red-edge to NIR, and one in the SWIR (1.6 µm). The absence of separate bands capturing the rapid increase in vegetation reflectance in the red-edge and the subsequent plateau may explain the small accuracy gap. Conversely our results (see Figure 9), suggest that similar accuracy can be achieved using only two visible channels, as using just the RED and NIR bands increased FC RMSE by only about two percentage points. Supporting this observation, Weiss et al. (2020) do not experience this limitation because their setup includes a larger number of spectral bands. These findings indicate that most of the relevant information for estimating LAI and FC is captured by the red and NIR bands, with the red-edge band providing a modest improvement in accuracy.

2.4.3 Model application to Sentinel-2 data

As a result of the training and evaluation on synthetic datasets (Section 3.2), the best-performing and most robust models were selected for application to real observational data. For this purpose, a Gaussian Process Regression (GPR) model with the seven selected input variables (RED, NIR, SZA, VZA, RAA, PAN, NIRv) was adopted. Two separate single-output GPR models were trained, one for LAI and one for FC. Each

model was implemented under two noise conditions, low and high, reflecting the variability introduced during synthetic training, yielding a total of four models to be applied and assessed on real-world measurements.

To ensure meaningful results, input parameters were consistent with those defined during model simulation training. The use of Sentinel-2 imagery provided the bands reconfigured according to the SBG-TIR setup, as described in Section 2.3.2. Basically, Sentinel-2 resampled spectral bands, Sentinel-2 NIRv and SZA, VZA, RAA corresponding to each acquisition and computed for each pixel represented the input data for the application to estimate LAI and FC. Results were then compared with GBOV field measurements.

2.4.3.1 Validation on the GBOV dataset

The scatterplots and residual plots between Sentinel-2 modelled LAI and FC and GBOV field measurements are presented in Figure 13. In these scatterplots, the y-axis shows the model predictions while the x-axis represents GBOV field measurements. Overall, a clear decrease in performance is observed compared to the synthetic data, with considerable dispersion, as expected due to the complexity of real-world conditions.

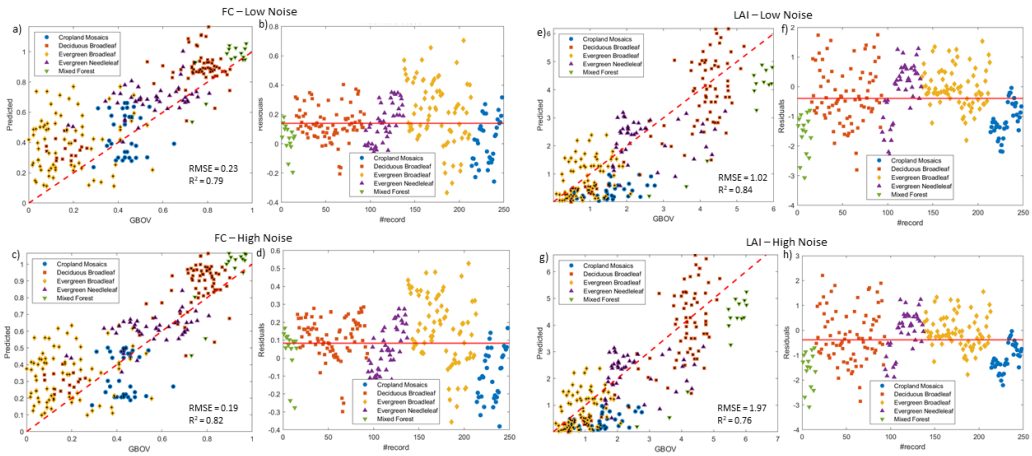

Figure 13: Sentinel-2 data analysis for FC (left) and LAI (right). The scatterplots for low noise (a, e) and high noise (c, g) cases are shown, while Panels (b, f) and (d, h) display the corresponding residual plots. Different colours represent the investigated land cover.

Adding noise to the training data had contrasting effects on model performance. For FC the inclusion of noise slightly improved predictions, reducing RMSE from 0.23 to 0.19 (Figure 13a, c). In contrast, LAI performance was substantially degraded, with RMSE increasing from 1.02 to 1.97 (Figure 13e, g). This difference can be attributed to the different sensitivity of each target variable to noise in the spectral domain. LAI as a structural variable, depends on complex canopy-radiation interactions and is therefore more prone to degradation when additional uncertainty is introduced, while FC being more directly related to broadband reflectance, may benefit from the regularizing effect of moderate noise (e.g. Garrigues et al., 2008, Weiss et al., 2004).

Additionally, inspection of the hyperparameters shown in Table 3 further illustrates this behaviour. In response to the added noise, the optimized GPR intrinsically σ_n remained largely unaffected for LAI, indicating that the model could not adjust to compensate for the perturbations. In contrast for FC σ_n exhibited a noticeable adjustment. This contrasting behaviour can be explained by the nature of the perturbations: the added noise simulates small fluctuations in the local parameter space explored during training, allowing the model to experience realistic variability. The GPR model is intrinsically robust to such variations, as it explicitly accounts for Gaussian noise through the relation $y_i = f(x_i) + \varepsilon_i$, where $\epsilon(\lambda) \sim N(0, {\sigma_n}^2)$ with the noise variance ${\sigma_n}^2$ optimized as a hyperparameter.

This result highlights the intrinsic difficulty of reproducing real-world data with synthetic simulations, especially when validation relies on spectral features that are slightly biased and not perfectly aligned with SBG characteristics. Moreover, the need to account for both SBG and Sentinel-2 instruments configuration within the training

space further limited performance. Additional sources of uncertainty may arise from the atmospheric correction process, which can introduce biases or errors in the retrieved reflectance spectra. To mitigate these limitations, approaches such as active learning (where the model iteratively selects the most informative real observations to refine the training set) or coupling synthetic data generation with atmospheric radiative transfer simulation (e.g. using MODTRAN) could be explored, potentially enhancing both predictive accuracy and robustness.

Residual plots show the deviation between predicted and observed values on the y-axis. No pronounced trends were observed and bias levels were consistently low. Land cover-stratified analysis reveals clear patterns, with different classes clustering within specific value ranges. These distributions align with the seasonal timing of field data collection, which was primarily conducted during spring and summer. Systematic overestimation of FC is observed for evergreen broadleaf forests, and slight overestimation for deciduous broadleaf and mixed forest classes. These biases are mitigated by adding noise to the training data, reducing the mean residual bias (from 0.14 to 0.08) and also decreases the number of predictions yielding non-physical values (i.e., exceeding 1). For LAI estimation, cropland and mixed forest tend to be underestimated, whereas deciduous broadleaf exhibits a high degree of variability.

We also tested regions of interest (ROI) of 3×3 and 5×5 size to evaluate whether spatial representativeness could affect the results, without finding significant differences in model performance (data not shown). No significant differences were observed, which is consistent and expected given that GBOV reference measurements (Bai et al., 2019) are point-based and Sentinel-2 imagery has a 20 m spatial resolution conditions.

2.4.3.2 Comparison with SNAP toolbox

ML results were finally compared with those obtained using the LAI/FC/FAPAR algorithm included in the SNAP toolbox (Weiss et al, 2020). Since the SNAP algorithm

operates on full Sentinel-2 images, the LAI and FC values were extracted specifically for the same pixels and regions of interest (ROIs) used to validate our models. This ensured a direct and consistent comparison between the GPR predictions and the SNAP outputs. The SNAP processor consistently underestimates LAI values across all land-cover classes, whereas our proposed approach demonstrates improved performance on this dataset (Figure 14c, d). Regarding Fractional Cover (FC), both methods exhibit class-dependent clustering effects (Figure 14a, b). Evergreen broadleaf forests are generally overestimated, with SNAP showing a lower bias compared to our model. In contrast, mixed forest and deciduous broadleaf classes are slightly underestimated by SNAP, whereas our approach yields a mild overestimation. Cropland areas also tend to be underestimated by SNAP. Despite these differences, the overall bias remains limited.

Comparisons were conducted at the pixel level. Increasing the ROI size did not result in any significant improvement in model performance or correlation metrics, suggesting that spatial aggregation does not enhance the reliability of the retrievals in this context.

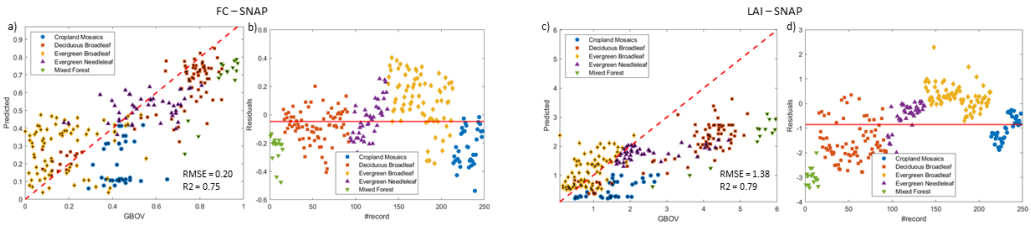

Figure 14. Real-data testing results for the SNAP Biophysical Processor. Scatterplots and residual plots are shown for LAI (a, b) and FC (c, d), based on Sentinel-2 pixels corresponding to GBOV reference points.

To further quantify model performance, Figure 15 presents a statistical comparison in terms of RMSE and R^2 for both the proposed model, evaluated under low and high noise configurations, and the SNAP Biophysical Processor. For LAI retrieval, the low-noise configuration of our model yields the highest accuracy, outperforming both the SNAP processor and the high-noise variant. Consequently, this configuration is recommended for LAI estimation within the SBG-TIR framework. In the case of FC, the inclusion of

high noise in the training data produces results statistically comparable to those obtained with SNAP, indicating that this configuration is more suitable for FC retrieval in the same context.

Beyond the accuracy metrics, the uncertainty distributions, shown in Figure 15c, align well with theoretical expectations. When noise is introduced during training, the resulting uncertainty distribution becomes more concentrated, characterized by a higher mode and shorter tails. This behaviour suggests that the model more effectively captures the intrinsic and aleatoric variability of the data, while simultaneously reducing spurious fluctuations associated with epistemic uncertainty. The addition of noise acts as a regularization mechanism, contributing to a more realistic and stable characterization of uncertainty.

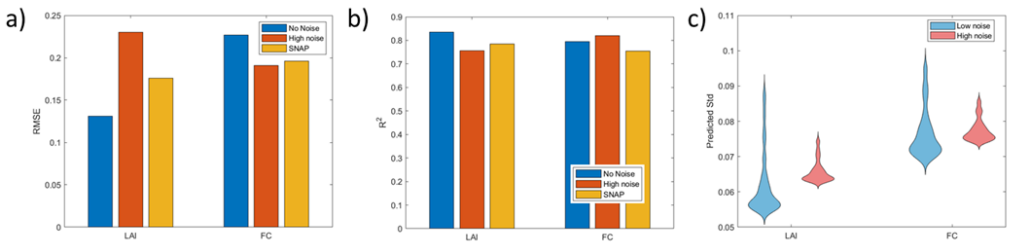

Figure 15: Performance comparison on real data. a) bar plots of RMSE for LAI and FC predictions validated against the GBOV dataset, under three conditions (no noise, SNR = 10, SNAP biophysical processor). b): same bar plots for R^2. c): violin plot of the estimated uncertainties.

A further comparison between the two methods can be made in terms of their practical applicability and operational differences. While the two approaches share similar methodological foundations, some practical differences exist. The SNAP processor requires the full Sentinel-2 scene and a dedicated software environment, while our ML method can be applied directly to specific pixels of interest, offering greater flexibility and computational efficiency. Moreover, our ML-based inversion provides pixel-level uncertainty estimates, which are not available in SNAP. Conversely, the SNAP processor includes data quality flags based on likelihood analyses both within and across

spectral bands, identifying pixels whose spectral signatures fall outside the training hypercube domain.

To summarize, it can be asserted that the overall proposed approach lies in the integration of a physically based synthetic training framework with a dedicated end-to-end machine learning retrieval pipeline, specifically tailored to the spectral configuration of the SBG-TIR mission. A key feature of the proposed methodology is its explicit treatment of mixed-pixel conditions, wherein sub-pixel mixtures of soil and vegetation exert a significant influence on reflectance properties. To address this complexity, the framework incorporates the panchromatic channel, which captures broadband visible reflectance and enhances sensitivity to both soil and canopy brightness. This integration improves the model's robustness to spatial heterogeneity and supports more accurate retrievals of biophysical parameters with both simulated and Sentinel-2 data. Furthermore, Fractional Vegetation Cover is here explicitly derived from canopy structural parameters simulated by the SCOPE radiative transfer model, ensuring a physically consistent definition that is directly linked to vegetation architecture rather than relying on empirical spectral proxies. This structural grounding enhances the interpretability and generalizability of the retrievals across different ecosystems and sensor platforms, as demonstrated by using Sentinel-2 data resampled in SBG-TIR configuration.

2.5 CONCLUSIONS

In this study, we present a machine learning framework for the predictive retrieval of Leaf Area Index and Fractional Vegetation Cover in the context of the upcoming SBG-TIR mission. Multiple models were trained and validated using both synthetic and real datasets. The optimal configurations were selected based on their consistent performance across varying initial conditions, including randomized parameters for synthetic library generation and training initialization. This strategy ensures robustness to training variability and enhances model stability with respect to data and parameter fluctuations.

Two independent models were developed, one for each target variable. Gaussian Process Regression emerged as the most accurate approach for both LAI and FC. Specifically, the LAI model performed best under low-noise training conditions, while the FC model benefited from high-noise training, indicating that noise injection can act as an effective regularization strategy.

The selected models were applied to Sentinel-2 imagery, demonstrating predictive behaviour consistent with theoretical expectations and comparable to existing benchmarks. All seven input variables (RED, NIR, PAN, SZA, VZA, RAA, and NIRv) were utilized; however, sensitivity analysis revealed that the RED and NIR channels alone provided high predictive accuracy, with marginal gains from additional variables. Mixed-pixel scenarios were essential for achieving reliable results on real-world data. In this context, the panchromatic channel proved particularly valuable, offering an integrative measure of the visible spectrum, enhancing sensitivity to soil reflectance, and improving the retrieval of biophysical parameters.

A distinctive methodological contribution of this work is the explicit derivation of FC from leaf angle distribution, yielding results that are both biophysically meaningful and numerically consistent. The framework is fully transferable to other sensors or missions with similar spectral VIREO characteristics and can be further extended by incorporating additional input features or accounting for soil variability.

Overall, the primary innovation of this study lies in the tailored application of the framework to the SBG-TIR mission, considering its specific spectral configuration and operational constraints. While individual components such as noise injection and mixed-pixel handling have been explored in previous research, this work integrates them into a coherent and operationally viable machine learning pipeline. The resulting framework enables accurate and robust retrievals of LAI and FC and represents a transferable solution ready for deployment in SBG-TIR and future multispectral missions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit author statement

L. Tuzzi: Conceptualization, Methodology, Investigation, Writing – original draft.
R. Colombo: Conceptualization, Methodology, Review & editing, Funding acquisition.
S. Venafra: Review & editing, THERESA Project management.

Acknowledgements:

This work has been supported by the THERESA project (THErmal infRarEd SBG Algorithms) funded by the Italian Space Agency (ASI).

2.6 References

Bach, H.; Mauser, W. Modelling and Model Verification of the Spectral Reflectance of Soils under Varying Moisture Conditions. In Proceedings of the Proceedings of IGARSS '94 - 1994 IEEE International Geoscience and Remote Sensing Symposium; IEEE: Pasadena, CA, USA, 1994; Vol. 4, pp. 2354–2356.

Bacour, C.; Bréon, F.-M.; Maignan, F. Normalization of the Directional Effects in NOAA–AVHRR Reflectance Measurements for an Improved Monitoring of Vegetation Cycles. Remote Sensing of Environment 2006, 102, 402–413, https://doi.org/10.1016/j.rse.2006.03.006

Badgley, G.; Field, C.B.; Berry, J.A. Canopy Near-Infrared Reflectance and Terrestrial Photosynthesis. Sci. Adv. 2017, 3, e1602244, https://doi.org/10.1126/sciadv.1602244.

Bai, G.; Gobron, N.; Dash, J.; Brown, L.; Meier, C.; Lerebourg, C.; Ronco, E.; Lamquin,

N.; Bruniquel, V.; Clerici, M. GBOV (Ground-Based Observation for Validation): A Copernicus Service for Validation of Vegetation Land Products. In Proceedings of the IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium; IEEE: Yokohama, Japan, July 2019; pp. 4592–4594.

Banskota, A.; Serbin, S.P.; Wynne, R.H.; Thomas, V.A.; Falkowski, M.J.; Kayastha, N.; Gastellu-Etchegorry, J.-P.; Townsend, P.A. An LUT-Based Inversion of DART Model to Estimate Forest LAI from Hyperspectral Data. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 2015, 8, 3147–3160, https://doi.org/10.1109/JSTARS.2015.2401515.

Baret, F.; Hagolle, O.; Geiger, B.; Bicheron, P.; Miras, B.; Huc, M.; Berthelot, B.; Niño, F.; Weiss, M.; Samain, O.; et al. LAI, fAPAR and fCover CYCLOPES Global Products Derived from VEGETATION. Remote Sensing of Environment 2007, 110, 275–286, https://doi.org/10.1016/j.rse.2007.02.018.

Baret, F.; Weiss, M.; Lacaze, R.; Camacho, F.; Makhmara, H.; Pacholcyzk, P.; Smets, B. 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 2013, 137, 299–309, https://doi.org/10.1016/j.rse.2012.12.027.

Bonham, C.D. Measurements for Terrestrial Vegetation; John Wiley & Sons, Ltd.: Chichester, UK, 2013. https://doi.org/10.1002/9781118534540

Brede, B.; Verrelst, J.; Gastellu-Etchegorry, J.-P.; Clevers, J.G.P.W.; Goudzwaard, L.; Den Ouden, J.; Verbesselt, J.; Herold, M. Assessment of Workflow Feature Selection on Forest LAI Prediction with Sentinel-2A MSI, Landsat 7 ETM+ and Landsat 8 OLI. Remote Sensing 2020, 12, 915, https://doi.org/10.3390/rs12060915.

Campbell, G.S. Extinction Coefficients for Radiation in Plant Canopies Calculated Using an Ellipsoidal Inclination Angle Distribution. Agricultural and Forest Meteorology 1986, 36, 317–321, https://doi.org/10.1016/0168-1923(86)90010-9.

Campos-Taberner, M.; García-Haro, F.J.; Camps-Valls, G.; Grau-Muedra, G.; Nutini, F.; Crema, A.; Boschetti, M. Multitemporal and Multiresolution Leaf Area Index Retrieval for Operational Local Rice Crop Monitoring. Remote Sensing of Environment 2016, 187, 102–118, https://doi.org/10.1016/j.rse.2016.10.009.

Camps-Valls, G.; Svendsen, D.H.; Martino, L.; Muñoz-Marí, J.; Laparra, V.; Campos-Taberner, M.; Luengo, D. Physics-Aware Gaussian Processes for Earth Observation. In Image Analysis. SCIA 2017; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; pp. 205–217. doi.org/10.1007/978-3-319-59129-2_18

Camps-Valls, G.; Verrelst, J.; Munoz-Mari, J.; Laparra, V.; Mateo-Jimenez, F.; Gomez-Dans, J. A Survey on Gaussian Processes for Earth-Observation Data Analysis: A Comprehensive Investigation. IEEE Geosci. Remote Sens. Mag. 2016, 4, 58–78, https://doi.org/10.1109/MGRS.2015.2510084.

Carlson, T.N.; Ripley, D.A. On the Relation between NDVI, Fractional Vegetation Cover, and Leaf Area Index. Remote Sensing of Environment 1997, 62, 241–252, https://doi.org/10.1016/S0034-4257(97)00104-1.

Chambrelan, A.; S2 MPC team. Sentinel-2 Level-1 Algorithm Theoretical Bases Document (ATBD), Issue 1.1; European Space Agency: Paris, France, 2023. Available online: https://sentiwiki.copernicus.eu/__attachments/1692737/S2-PDGS-MPC-ATBD-L1%20-%20Sentinel-2%20Level%201%20Algorithm%20Theoretical%20Bases%20Document%20202

3%20-%201.1.pdf?inst-v=19224347-b78a-4e98-8878-0bb2ef5b4589 (accessed on 19 December 2025).

Chen, J.M.; Black, T.A. Defining Leaf Area Index for Non-Flat Leaves. Plant Cell Environ. 1992, 15, 421–429. doi.org/10.1111/j.1365-3040.1992.tb00992.x

Chen, S.; Zhao, W.; Zhang, R.; Sun, X.; Zhou, Y.; Liu, L. Higher Sensitivity of NIRv,Rad in Detecting Net Primary Productivity of C4 Than That of C3: Evidence from Ground Measurements of Wheat and Maize. Remote Sensing 2023, 15, 1133, https://doi.org/10.3390/rs15041133.

Claverie, M.; Vermote, E.F.; Weiss, M.; Baret, F.; Hagolle, O.; Demarez, V. Validation of Coarse Spatial Resolution LAI and FAPAR Time Series over Cropland in Southwest France. Remote Sensing of Environment 2013, 139, 216–230, https://doi.org/10.1016/j.rse.2013.07.027.

De Grave, C.; Verrelst, J.; Morcillo-Pallarés, P.; Pipia, L.; Rivera-Caicedo, J.P.; Amin, E.; Belda, S.; Moreno, J. Quantifying Vegetation Biophysical Variables from the Sentinel-3/FLEX Tandem Mission: Evaluation of the Synergy of OLCI and FLORIS Data Sources. Remote Sensing of Environment 2020, 251, 112101, https://doi.org/10.1016/j.rse.2020.112101.

Delegido, J.; Verrelst, J.; Alonso, L.; Moreno, J. Evaluation of Sentinel-2 Red-Edge Bands for Empirical Estimation of Green LAI and Chlorophyll Content. Sensors 2011, 11, 7063–7081, https://doi.org/10.3390/s110707063.

Ding, Y.; Zheng, X.; Jiang, T. Comparison of Fractional Vegetation Cover Estimating Methods Using In-Situ Measurements and the PROSAIL Model. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS); IEEE: Beijing, China, July 2016; pp. 4351–4354.

Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment 2012, 120, 25–36, https://doi.org/10.1016/j.rse.2011.11.026.

Durbha, S.S.; King, R.L.; Younan, N.H. Support Vector Machines Regression for Retrieval of Leaf Area Index from Multiangle Imaging Spectroradiometer. Remote Sensing of Environment 2007, 107, 348–361, https://doi.org/10.1016/j.rse.2006.09.031.

Frank, S.A. The Common Patterns of Nature. J of Evolutionary Biology 2009, 22, 1563–1585, https://doi.org/10.1111/j.1420-9101.2009.01775.x.

García-Haro, F.J.; Campos-Taberner, M.; Muñoz-Marí, J.; Laparra, V.; Camacho, F.; Sánchez-Zapero, J.; Camps-Valls, G. Derivation of Global Vegetation Biophysical Parameters from EUMETSAT Polar System. ISPRS Journal of Photogrammetry and Remote Sensing 2018, 139, 57–74, https://doi.org/10.1016/j.isprsjprs.2018.03.005.

García-Martínez, C.; Moncholi-Estornell, A.; Pescador-Dionisio, S.; Cendrero-Mateo, M.P.; Rodrigo, M.J.; Moreno, J.; Van Wittenberghe, S. Canopy Imaging Spectroscopy Reveals a Stepwise Redshifted Energy Redistribution in the Antenna under Drought Stress. Sci Rep 2025, 15, 17241, https://doi.org/10.1038/s41598-025-01940-0.

García-Soria, J.L.; Morata, M.; Berger, K.; Pascual-Venteo, A.B.; Rivera-Caicedo, J.P.; Verrelst, J. Evaluating Epistemic Uncertainty Estimation Strategies in Vegetation Trait Retrieval Using Hybrid Models and Imaging Spectroscopy Data. Remote Sensing of Environment 2024, 310, 114228,

https://doi.org/10.1016/j.rse.2024.114228.

Garrigues, S.; Lacaze, R.; Baret, F.; Morisette, J.T.; Weiss, M.; Nickeson, J.E.; Fernandes, R.; Plummer, S.; Shabanov, N.V.; Myneni, R.B.; et al. Validation and Intercomparison of Global Leaf Area Index Products Derived from Remote Sensing Data. J. Geophys. Res. 2008, 113, 2007JG000635, https://doi.org/10.1029/2007JG000635.

Gastellu-Etchegorry, J.P.; Grau, E.; Lauret, N. DART: A 3D Model for Remote Sensing Images and Radiative Budget of Earth Surfaces. In Modeling and Simulation in Engineering; Alexandru, C., Ed.; InTech, 2012 ISBN 978-953-51-0012-6.

Gitelson, A.A.; Arkebauer, T.J.; Suyker, A.E. Convergence of Daily Light Use Efficiency in Irrigated and Rainfed C3 and C4 Crops. Remote Sensing of Environment 2018, 217, 30–37, https://doi.org/10.1016/j.rse.2018.08.007.

Gitelson, A.A.; Gritz †, Y.; Merzlyak, M.N. Relationships between Leaf Chlorophyll Content and Spectral Reflectance and Algorithms for Non-Destructive Chlorophyll Assessment in Higher Plant Leaves. Journal of Plant Physiology 2003, 160, 271–282, https://doi.org/10.1078/0176-1617-00887.

Houborg, R.; McCabe, M.F. A Hybrid Training Approach for Leaf Area Index Estimation via Cubist and Random Forests Machine-Learning. ISPRS Journal of Photogrammetry and Remote Sensing 2018, 135, 173–188, https://doi.org/10.1016/j.isprsjprs.2017.10.004.

Jacquemoud, S.; Baret, F. PROSPECT: A Model of Leaf Optical Properties Spectra. Remote Sensing of Environment 1990, 34, 75–91, https://doi.org/10.1016/0034-4257(90)90100-Z.

Jacquemoud, S.; Verhoef, W.; Baret, F.; Bacour, C.; Zarco-Tejada, P.J.; Asner, G.P.; François, C.; Ustin, S.L. PROSPECT+SAIL Models: A Review of Use for Vegetation Characterization. Remote Sensing of Environment 2009, 113, S56–S66, https://doi.org/10.1016/j.rse.2008.01.026.

Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? 2017.

Knyazikhin, Y.; Martonchik, J.V.; Diner, D.J.; Myneni, R.B.; Verstraete, M.; Pinty, B.; Gobron, N. Estimation of Vegetation Canopy Leaf Area Index and Fraction of Absorbed Photosynthetically Active Radiation from Atmosphere-corrected MISR Data. J. Geophys. Res. 1998, 103, 32239–32256, https://doi.org/10.1029/98JD02461.

Lauvernet, C.; Baret, F.; Hascoët, L.; Buis, S.; Le Dimet, F.-X. Multitemporal-Patch Ensemble Inversion of Coupled Surface–Atmosphere Radiative Transfer Models for Land Surface Characterization. Remote Sensing of Environment 2008, 112, 851–861, https://doi.org/10.1016/j.rse.2007.06.027.

Lazaro-Gredilla, M.; Van Vaerenbergh, S. A Gaussian Process Model for Data Association and a Semidefinite Programming Solution. IEEE Trans. Neural Netw. Learning Syst. 2014, 25, 1967–1979, https://doi.org/10.1109/TNNLS.2014.2300701.

Lin, X.; Wen, J.; Liu, Q.; You, D.; Wu, S.; Hao, D.; Xiao, Q.; Zhang, Z.; Zhang, Z. Spatiotemporal Variability of Land Surface Albedo over the Tibet Plateau from 2001 to 2019. Remote Sensing 2020, 12, 1188, https://doi.org/10.3390/rs12071188.

Locherer, M.; Hank, T.; Danner, M.; Mauser, W. Retrieval of Seasonal Leaf Area Index

from Simulated EnMAP Data through Optimized LUT-Based Inversion of the PROSAIL Model. Remote Sensing 2015, 7, 10321–10346, https://doi.org/10.3390/rs70810321.

Meroni, M.; Colombo, R.; Panigada, C. Inversion of a Radiative Transfer Model with Hyperspectral Observations for LAI Mapping in Poplar Plantations. Remote Sensing of Environment 2004, 92, 195–206, https://doi.org/10.1016/j.rse.2004.06.005.

Myneni, R.B.; Hoffman, S.; Knyazikhin, Y.; Privette, J.L.; Glassy, J.; Tian, Y.; Wang, Y.; Song, X.; Zhang, Y.; Smith, G.R.; et al. Global Products of Vegetation Leaf Area and Fraction Absorbed PAR from Year One of MODIS Data. Remote Sensing of Environment 2002, 83, 214–231, https://doi.org/10.1016/S0034-4257(02)00074-3.

Nilson, T. A Theoretical Analysis of the Frequency of Gaps in Plant Stands. Agricultural Meteorology 1971, 8, 25–38, https://doi.org/10.1016/0002-1571(71)90092-6.

Pérez-Suay, A.; Amorós-López, J.; Gómez-Chova, L.; Laparra, V.; Muñoz-Marí, J.; Camps-Valls, G. Randomized Kernels for Large Scale Earth Observation Applications. Remote Sensing of Environment 2017, 202, 54–63, https://doi.org/10.1016/j.rse.2017.02.009.

Petri, C.A.; Galvão, L.S. Sensitivity of Seven MODIS Vegetation Indices to BRDF Effects during the Amazonian Dry Season. Remote Sensing 2019, 11, 1650, https://doi.org/10.3390/rs11141650.

Schowengerdt, R.A. Remote Sensing: Models and Methods for Image Processing, 3rd ed.; Academic Press: Burlington, MA, USA, 2007.

Tran, B.N.; Van Der Kwast, J.; Seyoum, S.; Uijlenhoet, R.; Jewitt, G.; Mul, M. Uncertainty Assessment of Satellite Remote-Sensing-Based Evapotranspiration Estimates: A Systematic Review of Methods and Gaps. Hydrol. Earth Syst. Sci. 2023, 27, 4505–4528, https://doi.org/10.5194/hess-27-4505-2023.

Van Der Tol, C.; Verhoef, W.; Timmermans, J.; Verhoef, A.; Su, Z. An Integrated Model of Soil-Canopy Spectral Radiances, Photosynthesis, Fluorescence, Temperature and Energy Balance. Biogeosciences 2009, 6, 3109–3129, https://doi.org/10.5194/bg-6-3109-2009.

Verger, A.; Baret, F.; Weiss, M. Performances of Neural Networks for Deriving LAI Estimates from Existing CYCLOPES and MODIS Products. Remote Sensing of Environment 2008, 112, 2789–2803, https://doi.org/10.1016/j.rse.2008.01.006.

Verhoef, W. Light Scattering by Leaf Layers with Application to Canopy Reflectance Modeling: The SAIL Model. Remote Sensing of Environment 1984, 16, 125–141, https://doi.org/10.1016/0034-4257(84)90057-9.

Verhoef, W.; Van Der Tol, C.; Middleton, E.M. Hyperspectral Radiative Transfer Modeling to Explore the Combined Retrieval of Biophysical Parameters and Canopy Fluorescence from FLEX – Sentinel-3 Tandem Mission Multi-Sensor Data. Remote Sensing of Environment 2018, 204, 942–963, https://doi.org/10.1016/j.rse.2017.08.006.

Verrelst, J.; Camps-Valls, G.; Muñoz-Marí, J.; Rivera, J.P.; Veroustraete, F.; Clevers, J.G.P.W.; Moreno, J. Optical Remote Sensing and the Retrieval of Terrestrial Vegetation Bio-Geophysical Properties – A Review. ISPRS Journal of Photogrammetry and Remote Sensing 2015, 108, 273–290, https://doi.org/10.1016/j.isprsjprs.2015.05.005.

Verrelst, J.; Muñoz, J.; Alonso, L.; Delegido, J.; Rivera, J.P.; Camps-Valls, G.; Moreno, J. Machine Learning Regression Algorithms for Biophysical Parameter Retrieval: Opportunities for Sentinel-2 and -3. Remote Sensing of Environment 2012, 118, 127–139, https://doi.org/10.1016/j.rse.2011.11.002.

Verrelst, J.; Schaepman, M.E.; Koetz, B.; Kneubühler, M. Angular Sensitivity Analysis of Vegetation Indices Derived from CHRIS/PROBA Data. Remote Sensing of Environment 2008, 112, 2341–2353, https://doi.org/10.1016/j.rse.2007.11.001.

Weiss, M.; Baret, F. ATBD_S2ToolBox_L2B_V1.1; INRAE, Montpellier, France, 2016. Available online: https://step.esa.int/docs/extra/ATBD_S2ToolBox_L2B_V1.1.pdf (accessed on 19 December 2025).

Weiss, M.; Baret, F.; Jay, S. S2ToolBox Level 2 Products LAI, FAPAR, FCOVER, 2020.

Weiss, M.; Baret, F.; Smith, G.J.; Jonckheere, I.; Coppin, P. Review of Methods for in Situ Leaf Area Index (LAI) Determination. Agricultural and Forest Meteorology 2004, 121, 37–53, https://doi.org/10.1016/j.agrformet.2003.08.001.

Yang, F.; White, M.A.; Michaelis, A.R.; Ichii, K.; Hashimoto, H.; Votava, P.; Zhu, A.-X.; Nemani, R.R. Prediction of Continental-Scale Evapotranspiration by Combining MODIS and AmeriFlux Data Through Support Vector Machine. IEEE Trans. Geosci. Remote Sensing 2006, 44, 3452–3461, https://doi.org/10.1109/TGRS.2006.876297.

Yang, P.; Prikaziuk, E.; Verhoef, W.; Van Der Tol, C. SCOPE 2.0: A Model to Simulate Vegetated Land Surface Fluxes Andsatellite Signals 2020.

Yin, C.L.; Meng, F.; Yu, Q.R. Calculation of Land Surface Emissivity and Retrieval of Land Surface Temperature Based on a Spectral Mixing Model. Infrared Physics & Technology 2020, 108, 103333, https://doi.org/10.1016/j.infrared.2020.103333.

Zheng, G.; Moskal, L.M. Retrieving Leaf Area Index (LAI) Using Remote Sensing: Theories, Methods and Sensors. Sensors 2009, 9, 2719–2745, https://doi.org/10.3390/s90402719.

Chapter 3

Enhancing quantification of local carbon sinks at regional scale through Eddy Covariance CO2 Flux and Machine Learning.

Luca Tuzzi [1], Marta Galvagno[2], Gianluca Filippa[2], Jacob A Nelson[3]

1 Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milano (Italy)

2 Environmental Protection Agency of Aosta Valley, Climate Change Dept. - ARPA VdA – Aosta (Italy)

3 Max Planck Institute for Biogeochemistry, Biogeochemical Integration, Jena (Germany)

Keywords: Eddy Covariance, Machine learning, Remote sensing, GPP, CO2 Flux, Data analysis, FLUXCOM

3.1 ABSTRACT

Under the Paris Agreement, countries are encouraged to preserve and enhance existing carbon sinks. Europe, in particular, has committed to achieving climate neutrality-attaining a balance between anthropogenic emissions from sources and removals by sinks-by 2050.Achieving these ambitious goals requires accurate and credible estimation of CO_2 fluxes which in turn requires, as a first step, a reliable quantification of gross photosynthetic fluxes. However, discrepancies between observations and global models hinder the tracking of collective progress towards climate neutrality. Improving transparency and data comparability is therefore crucial to better align local mitigation strategies with global pathways. In particular, effective climate mitigation policies increasingly depend on local-level actions where detailed data on CO_2 removals from forests and other land uses are traditionally lacking. Addressing the uncertainty in land-sector mitigation potential and enhancing the availability and comparability of data are critical for achieving climate goals by cities and regions. Different models, including process-based and data-driven approaches, exist to estimate land carbon fluxes, but their application and accuracy often vary significantly depending on the scale and quality of input data.

In this study, we tested a data-driven method based on eddy covariance (EC) data to quantify the current role of the regional carbon sink of the Aosta Valley Region (Italy) through the integration of various approaches. Our model relies on FLUXCOM-X framework specifically trained to achieve robust results at the regional scale. An XGBoost (eXtreme Gradient Boosting) model was developed using global hourly meteorological data from sites across the global eddy covariance networks paired with remote sensing data from MODIS (Moderate-Resolution Imaging Spectroradiometer). The algorithm was optimized through feature selection analysis and best training subset selection, identifying the ensemble of experimental sites that provided the most accurate predictions while avoiding overfitting. The optimal training subset was obtained via partitioning the full range of sites into subsets based on key characteristics (Plant Functional Type, geographical zone, biogeographical region, elevation). This approach ensured the biophysical comparability of the sites with the target region (Aosta Valley)

particularly with the EC site of Torgnon (IT-Tor) while maintaining a balance between generalizability and specificity. Model evaluation focused on how the model performed on the local eddy covariance measurements.

Consistent with this framework, model performance was evaluated in terms of daily gross primary productivity (GPP) estimates. The best estimates at the local scale for Torgnon were obtained by using the Alps subset with RMSD = 1.69 μmol CO_2 m^{-2} s^{-1}, NSE = 0.72 e R^2 = 0.85. Only in 2 out of 15 alpine sites did the local model outperform the global one; otherwise, the complete model performed better. In a spatialization perspective, when using a reduced set of variables (meteorological, flux, and satellite predictors), the mountain subset achieved the best results with RMSD = 1.76 μmol CO_2 m^{-2} s^{-1}, NSE = 0.70 e R^2 = 0.84. Again, only in 2 cases out of 20 sites did the reduced model perform better, while in 3 cases the performances were comparable. The selected model reproduced well the daily and seasonal dynamics of GPP at the site, particularly during the snow-covered period, while overestimating productivity in spring and autumn. The use of the reduced predictor set mitigated this overestimation, especially in autumn. Persistent snow cover (≈ 200 days per year) likely contributes to this bias, as such conditions are underrepresented in the global training database. Long-term analysis of daily estimates (2008-2021) revealed stable GPP estimation, suggesting limited changes in the carbon exchange regime of this alpine grassland site that is Torgnon. The methodology demonstrated potential for replication in other local context or regions, providing a flexible framework for assessing local carbon budgets and supporting climate-smart management strategies.

3.2. INTRODUCTION

Climate change is one of the most pressing challenges of our time, driven primarily by the increase in atmospheric CO_2 concentrations resulting from human activities. Ecosystems and forests in particular, acts as the main carbon sinks through photosynthesis, playing a crucial role in mitigating anthropogenic emissions (Migliavacca et al., 2025). Several international agreements are active in this regard, including the Paris Agreement, which aims to limit the increase in global average

temperature, as well as the European Union's target of climate neutrality by 2050. After defining such goals, it is necessary to identify stakeholders capable of understanding the scale of the problem and possessing the power, expertise, and leadership required to propose concrete solutions. In many cases, this role can be assumed by local authorities-those with the territorial reach needed to act effectively at the local level. In Italy, the Aosta Valley represents a virtuous example in this respect: it is a region that, during certain time windows, achieves climate neutrality, thanks in part to factors such as high forest cover and a favorable socioeconomic context, characterized by low industrial impact, sustainable agricultural practices, and a strong focus on local environmental policies.

In this context, understanding the carbon sequestration potential of natural ecosystems, and in particular forests, becomes essential not only to evaluate their role in mitigating climate change, but also to support local and regional policy decisions. Forests are a major component of the terrestrial carbon sink, critical for mitigating climate change (Luyssaert et al., 2008). However, quantifying carbon uptake by natural ecosystems remains challenging due to spatial and temporal variability, complex species interactions, and environmental factors such as climate and soil properties (Pan et al., 2011). These uncertainties make it difficult to accurately assess regional carbon budgets. Understanding carbon fluxes at the regional scale is crucial for informing climate and land-management policies.

Given the importance of alpine forests in carbon sequestration, in this work, we focus on the Aosta Valley region, a typical alpine ecosystem that presents unique challenges for accurately estimating carbon fluxes. The region spans from 300 to 4800 m of elevation, with an average altitude of about 2100 m, and encompasses a highly heterogeneous landscape with steep terrain, diverse vegetation types, and strong climatic gradients. Snow cover is particularly persistent, lasting up to 200 days per year in some areas such as Torgnon, which makes it an ideal location to investigate the dynamics of carbon exchange under snow-dominated conditions. However, these same characteristics limit the representativeness within global datasets, complicating the extrapolation of large-scale carbon models to the local scale (Jung et al., 2020; Wei et

al., 2021). Furthermore, alpine ecosystems are marked by high spatial and temporal variability in snow-related metrics (Fugazza et al., 2021), including information derived from satellite data such as MODIS.

To address these limitations, ground-based observations play a central role in monitoring carbon fluxes in complex alpine ecosystems. Among these, eddy covariance measurements provide direct observations of net CO_2 exchange between ecosystems and the atmosphere. This exchange reflects the balance between carbon uptake and release processes making the reliable quantification of gross photosynthetic fluxes a fundamental first step toward assessing net ecosystem carbon exchange (e.g., Baldocchi, 2014). In this context, the eddy covariance tower at the Torgnon site provides half-hourly measurements of turbulent CO_2 fluxes, latent and sensible heat, and key meteorological variables such as temperature, humidity, and radiation. This tower is part of the FLUXNET network which include approximately 300 sites worldwide (Baldocchi et al., 2001). The turbulent CO_2 flux provide NEE (Net Ecosystem Exchange) directly, from which GPP (Gross Primary Productivity) and RECO (Ecosystem Respiration) are derived using flux-partitioning methods (Reichstein et al., 2005). These observations offer precise and continuous monitoring of ecosystem carbon dynamics.

Despite their accuracy, eddy covariance measurements are limited to the immediate footprint surrounding the tower, and thus they cannot capture the local variability of the wider region. Decision-makers require spatially continuous, high-resolution estimates across the entire territory to guide land and climate policies. Direct extrapolation of tower data or global-scale models often diverge from local flux observation in this context. This highlights the need for a data-driven approach that integrates local flux measurements with regional meteorological and remote sensing inputs. Such integration can yield more reliable, spatially explicit carbon flux maps (Tramontana et al., 2016).

Other complementary methods are also widely used. The LULUCF (Land Use, Land-Use Change, and Forestry) inventory, which estimates carbon stored in biomass and its changes due to land use, forest management, deforestation, and reforestation, according to IPCC (2006; 2019 refinement) guidelines. These data are used for national and

regional carbon balance assessments but generally have coarser spatial and temporal resolution compared to other modeling approaches.

On the other side, atmospheric inversions represent a top-down approach for estimating carbon fluxes, in which surface fluxes are inferred from atmospheric CO_2 concentration measurements collected at different altitudes through the inversion of atmospheric transport models. Such methods, adopted for example within the framework of the European Copernicus program (CAMS), provide an integrated view of net carbon fluxes at continental or global scales. However, their spatial resolution is often limited at the local level, particularly in complex areas such as alpine regions, as reported by (Gómez-Ortiz et al., 2025).

Process-based models are continuously being developed and are widely used to estimate ecosystem carbon dynamics by explicitly representing the ecological and physiological mechanisms controlling carbon exchanges between the biosphere and the atmosphere. At the canopy scale, models such as the Soil Canopy Observation, Photochemistry and Energy balance (SCOPE) model (van der Tol et al., 2009) couple radiative transfer, energy balance, and photosynthetic processes, enabling accurate simulation of canopy-level carbon fluxes, including GPP and solar-induced chlorophyll fluorescence (SIF). While providing instantaneous flux estimates that can be compared with point-scale measurements, it does not represent ecosystem respiration and lacks memory of carbon storage. The carbon cycle can be simulated by a land surface model such as the Community Land Model (CLM; Lawrence et al., 2019). It simulates the temporal evolution of carbon, water, and energy pools by numerically solving a coupled system of ordinary differential equations. At the ecosystem scale, dynamic global vegetation models such as Lund-Potsdam-Jena Global Vegetation Model (LPJ; Sitch et al., 2003) operate at coarser scales, simulating long-term ecosystem dynamics by combining time-evolving carbon pools with ecological rules governing vegetation structure and composition. Across scales, all process-based models require numerous well-calibrated parameters describing vegetation and soil properties, which are often poorly constrained in mountain regions or areas with persistent snow cover, potentially introducing systematic errors in estimates of carbon and carbon-equivalent fluxes.

An alternative approach is represented by hybrid models, which integrate process-based simulations with observational data, resulting in process-informed, data-driven estimates. For example, in the case of SMAP L4C (Jones et al., 2017), soil moisture measurements from microwave sensors (SMAP) are assimilated into a process-based model to estimate NEE. Finally, purely data-driven models also exist.

In recent years, machine learning (ML) approaches have gained increasing attention for their ability to capture non-linear relationships between environmental variables and eddy tower measurements. Data-driven upscaling frameworks such as FLUXCOM have enabled substantial progress in estimating large-scale carbon and water fluxes by combining eddy covariance observations, satellite remote sensing, and meteorological inputs. The recent X-BASE product from the FLUXCOM-X framework provides high-spatial (0.05°) and high-temporal (hourly) resolution maps of GPP and NEE, showing improvements in global intercomparisons and consistency with atmospheric inversions in many temperate and boreal regions (Nelson et al., 2024).

Within this context, our research question focuses on the Aosta Valley: we investigate whether a machine learning model, when trained on a carefully preprocessed and regionally weighted subset of sites, can outperform the same architecture trained at the global scale when applied within that region. Specifically, we ask whether such a model can better reproduce the daily and seasonal dynamics of GPP, reduce regional biases and improve performance metrics such as RMSE, NSE, and R^2 compared to the global model. We build upon the FLUXCOM approach, which is designed with methodological flexibility, to explore factors such as predictor selection and randomness stability. Our goal is to specialize this well-established ML framework for local-scale applications, identify the optimal training dataset for the region of interest, evaluate its performance at the Torgnon site, and assess its potential for regional upscaling.

3.3 DATA AND METHODS

The assessment of carbon exchange between terrestrial ecosystems and the atmosphere is essential for understanding the role of ecosystems in regulating the carbon cycle and

for supporting climate change mitigation strategies. Several approaches have been developed to estimate the carbon balance across spatial scales ranging from local to regional and global. The approach adopted in this work is based on direct measurements of CO_2 fluxes between the land surface and the atmosphere. The eddy covariance (EC) method enables high temporal resolution estimates of the NEE, while through flux partitioning techniques it is possible to derive GPP and RECO. These are point-based measurements that can be used to train empirical or statistically based models, including machine learning algorithms, which provide estimates by learning the relationships between observed fluxes and coincident environmental predictor variables. In this way, the model can be applied in prediction mode to extend the estimation of fluxes to locations where direct measurements are not available.

3.3.1 Eddy covariance method and dataset definition

The eddy covariance (EC) method is a micrometeorological technique that enables the measurement of gas and energy exchanges between ecosystems and the atmosphere (Baldocchi, 2020). It is based on determining the correlation between fluctuations of the vertical wind component and variations in the density of a scalar quantity, such as the concentration of a trace gas or air temperature (Aubinet et al., 2012). In this way, it is possible to quantify carbon fluxes and equivalent carbon exchanges as well as both sensible heat flux (related to direct heat transfer) and latent heat flux (associated with evapotranspiration). The EC method represents the primary approach for obtaining bottom-up estimates of the continental carbon balance, over temporal scales ranging from half-hourly to interannual (Baldocchi et al., 2001; Yu et al., 2006).

From a spatial perspective, one of its major limitations is related to the flux footprint that is the area from which the turbulent signals measured by the eddy covariance tower originate. The footprint represents the effective source area sampled by the tower (Wilson, 2015), typically extending from several hundred meters up to a few square kilometers, and it is not homogeneously centered around the tower itself. Proper footprint characterization is essential for interpreting flux measurements and improving

their accuracy, for instance by filtering out non-representative data (Foken & Leclerc, 2004). Although a detailed quantitative analysis of footprint effects is beyond the scope of this section, their implications are revisited in the conclusions, where we discuss how footprint heterogeneity may influence model representativeness and the spatial scaling of GPP estimates.

Eddy covariance measurements are acquired as high-frequency signals (10-20 Hz) and then aggregated into half-hourly averages, which constitute the point-based time series. The flux can thus be calculated from the covariance between the fluctuations of the vertical wind velocity (w') and those of the scalar associated with the flux, expressed as $Fs \approx w's'$. For example, in the case of CO_2 fluxes, the measured variable corresponds to the Net Ecosystem Exchange (NEE), derived from CO_2 concentration; similarly, latent heat flux is related to water vapor (weighted by the latent heat of evaporation), while sensible heat flux depends on air temperature and thermodynamic properties. These measurements, collected from multiple experimental sites, are harmonized and made available globally through the FLUXNET dataset, which aggregates data from more than 200 sites across 30 countries, representing diverse climates and ecosystems. The data undergo pre-processing first by local site teams and then through the ONE Flux data processing pipeline (Pastorello et al., 2020). Each dataset includes flux measurements, meteorological variables, and ancillary data.

The collected data are subjected to rigorous quality assurance and quality control (QA/QC) procedures, including instrumental corrections, low-turbulence filtering (u* filtering), and flux partitioning (Papale et al., 2006), before being integrated into the FLUXNET database. Recently, the Complementary Consistency Flags (C2F), also known as the Jung QC system (Jung et al., 2024), have been introduced to systematically identify inconsistencies and gaps in both flux and meteorological time series. This approach applies a series of data flags based on expected physical relationships between variables, daily and site-level coherence checks, and temporal continuity tests, thereby improving the reliability of multi-site analyses and machine learning-based models.

Gaps in the meteorological variables measured in situ may result from instrument malfunctions, unsuitable atmospheric conditions, or quality-control exclusions (Foken & Wichura, 1996). These gaps can last from a few hours to several months. The development of gap-filling methods that leverage global atmospheric reanalyses such as the ERA-Interim dataset, appropriately downscaled and debiased, has significantly improved the consistency and continuity of most FLUXNET predictors (Vuichard & Papale, 2015).

As complementary data, the dataset also includes products derived from MODIS remote sensing to characterize vegetation and photosynthesis. In this study, three indices derived from reflectance signals were used. The first is NIRv, the product of near-infrared reflectance and the Normalized Difference Vegetation Index (NDVI), computed from NIR and RED bands. NDVI is a proxy for vegetation cover (Carlson & Ripley, 1997), and NIRv modulates the reflectance of the canopy in the NIR, which is sensitive to leaf structure, making it a robust proxy for GPP (Badgley et al., 2017). The second index is the Enhanced Vegetation Index (EVI), which depends on RED, NIR, and BLUE bands and is calibrated following Huete et al. (2002). EVI provides a more linear and vegetation-sensitive index than NDVI. The inclusion of the BLUE band and calibration constants corrects some NDVI limitations, reducing saturation effects in dense vegetation and allowing the detection of GPP variations in highly vegetated areas. The third index is the Normalized Difference Water Index (NDWI), computed from NIR and shortwave infrared (SWIR), which increases with vegetation hydration (Gao, 1996). As water availability affects photosynthesis through stomatal closure, NDWI supports GPP modeling by providing information to distinguish variations due to biomass from those caused by water stress.

In this study, we selected a set of predictors and grouped them into different categories according to their nature, to evaluate their contribution to target predictions. This is shown in Table 1. Meteorological variables included air temperature (TA), vapor pressure deficit (VPD), friction velocity (u*), wind speed (WS), and air pressure (PA), describing local atmospheric and micrometeorological conditions. Flux-related variables included incoming shortwave radiation (SW_IN), potential incoming shortwave

radiation under clear-sky conditions (SW_IN_POT), its half-hourly derivative (dSW_IN_POT), and its daily excursion (dSW_IN_POT_day). These variables capture both instantaneous energy availability and seasonal changes in solar radiation. To account for the land surface thermal regime, both daytime (LST_day) and nighttime (LST_night) land surface temperature products were considered. From remote sensing, NIRv, EVI, NDWI, and plant functional type (PFT) as categorical information on vegetation cover were included. The study aimed to predict one of three main carbon flux variables NEE, GPP and RECO.

Class	Variable (acronym)	Description	Unit
Meteo	TA	Air temperature	°C
	VPD	Vapor pressure deficit	kPa
	u*	Friction velocity	$m\ s^{-1}$
	WS	Wind speed	$m\ s^{-1}$
	PA	Atmospheric pressure	kPa
Fluxes	SW_IN	Incoming shortwave radiation	$W\ m^{-2}$
	SW_IN_POT	Potential incoming shortwave radiation	$W\ m^{-2}$
	dSW_IN_POT	Temporal derivative of SW_IN_POT	$W\ m^{-2}\ day^{-1}$
	dSW_IN_POT_day	Daily excursion of SW_IN_POT	$W\ m^{-2}$
LST	LST_day	Daytime land surface temperature	K
	LST_night	Nighttime land surface temperature	K
Satellite	NDVI	Normalized Difference Vegetation Index	–
	EVI	Enhanced Vegetation Index	–
	NIRv	Near-infrared reflectance of vegetation	–
	NDWI	Normalized Difference Water Index (band 7)	–
Plant functional type	PFT	trees, shrubs, grasses, crops, others	categorical
Target	NEE	Net Ecosystem Exchange	$\mu mol\ CO_2$
	RECO	Ecosystem Respiration	$\mu mol\ CO_2$
	GPP	Gross Primary Production	$\mu mol\ CO_2$

Table 1: Predictors used, grouped by their respective categories.

These three carbon flux variables are not independent. In particular, when NEE is measured, flux partitioning into the two underlying processes is crucial. NEE represents the net balance of carbon uptake via photosynthesis (GPP) and carbon release through ecosystem respiration (RECO). This partitioning can be performed using various approaches (Reichstein et al., 2005). One typical approach relies on nighttime data, which are however biased due to suppression of turbulent fluxes at night and the dominance of advective flows (Feigenwinter et al., 2008). RECO is modeled as a

function of temperature and other variables, and GPP is obtained by subtracting RECO from NEE. The second approach examines light-response curves during daytime (Lasslop et al., 2010) to derive flux partitioning, although this method loses information about meteorological variables, particularly temperature (influencing RECO) and VPD (influencing GPP). Currently, these procedures are optimized using machine learning methods (Tramontana et al., 2020).

3.3.2 Machine learning model definition

In this study, we employed the Extreme Gradient Boosting (XGBoost) algorithm (Chen & Guestrin, 2016), a scalable implementation of gradient-boosted decision trees. It is particularly well suited for ecosystem flux prediction problems because it effectively handles non-linear relationships between predictors and target variables and captures complex interactions among predictors, also with high-dimensional datasets. Moreover it does not rely on assumptions about the nature of the predictors neither regarding their distribution nor their data type making so it is highly adaptable to heterogeneous inputs such as meteorological drivers, flux-derived variables, and satellite-based predictors. XGBoost is an ensemble learning method based on decision trees and operates through a boosting approach. It iteratively builds a sequence of shallow trees (weak learners), each dependent on the previous ones. At each iteration, the model optimizes a combination of the prediction error and a regularization term. These two components jointly improve predictive accuracy while controlling model complexity, thereby reducing overfitting and enhancing interpretability and generalization. The overall loss function used by XGBoost can be defined as:

$$L(\varphi) = \sum_{i=1}^{N} l(\mathrm{y_i}, \widehat{y_i}) + \sum_{k=1}^{K} R(\mathrm{f_k}) \qquad (1)$$

where given N data points y_i and $\hat{y}_i$ are the observed and predicted values of the target variable for the i-th sample, $l(y_i, \hat{y}_i)$ is the loss and R is the regularization term associated with the k-th tree f_k over a total of K trees. The regularization term in its basic form

includes three penalty components: one over the number of leaves (L0 type), a ridge L2 one, and an L1 lasso penalty:

$$R(f) = \rho T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 + \varphi \sum_{j=1}^{T} |\omega_j| \qquad (2)$$

where T denotes the number of leaves in the tree, ωj is the score of the j-th leaf and, ρ, λ, φ are the regularization coefficients for the L0, L2, and L1 terms respectively.

The prediction for the i-th data point at iteration t ($\hat{y}_i^t$) is obtained by adding the previous iteration's prediction ($\hat{y}_i^{t-1}$), whose the first guess value can be the mean of the target to the contribution of the t-th tree, scaled by the learning rate η: $\hat{y}_i^t = \hat{y}_i^{t-1} + \eta f_t$.

In practice, each tree does not predict the target directly but rather the residual correction to be added to the previous prediction in order to minimize the loss. Each tree is subject to the regularization constraint described above. The feature vector is mapped to a leaf weight by following the decision path through the tree's branches until reaching the corresponding leaf node. With proper configuration, XGBoost becomes a highly efficient learning algorithm. It achieves computational speed through parallelized computation at each boosting iteration, optimized memory management, and compressed data handling. The algorithm can read data in blocks directly from disk (out-of-core computation), thereby bypassing RAM limitations, and employs histogram-based feature binning, which reduces the computational complexity from O(#records) to O(#bins). Furthermore, it efficiently handles sparse datasets using actual pointers as made possible by the algorithm's core implemented in C++, while Python acts as a high-level wrapper, allowing easy integration with data science workflows.

The hyperparameters used in this study follow those adopted in previous works, particularly Nelson et al. (2024). The loss function selected for model optimization is the Mean Squared Error (MSE), suitable for continuous regression tasks such as carbon flux prediction. The number of boosting iterations was set to a high value of 1000, but in practice, the process is controlled by an early stopping of 10, preventing unnecessary overfitting and reducing training time. All available features were used at each node split

(so column sampled by node is set at 1), while a subsample ratio of 0.66666… ensures that each iteration is trained on two-thirds of the available data, introducing more stochasticity and improving model robustness. This corresponds to a bagging-like approach on the training data, while the number of parallel trees per iteration set to 1 indicates that a pure boosting scheme was applied.

A conservative learning rate eta of 0.05 was chosen to enhance model stability and generalization capability. Although a lower learning rate typically increases training time, the use of the histogram-based tree construction method efficiently mitigates the computational cost by aggregating data into histograms rather than processing each record individually. This approach greatly improves memory efficiency and scalability, especially with large datasets. The maximum tree depth of 10 was selected to allow the model to capture complex nonlinear interactions among predictors, which is particularly relevant when dealing with heterogeneous meteorological and ecological variables. Given the large training dataset, this configuration effectively prevents overfitting despite the model's expressive capacity. The minimum child weight of 5 sets the minimum sum of instance weights (i.e., the second derivative of the loss function) required to further split a leaf node. Under the assumption of independent predictions and MSE loss, this threshold approximately corresponds to the minimum number of data points required in a node for a split to occur, ensuring that each division contributes meaningfully to reducing residual error.

3.3.3 Training phase

The workflow followed is illustrated in Figure 1. The Full dataset corresponds to the FLUXNET database, which includes 294 experimental sites distributed worldwide and covering the period 2001–2020 (Table 2), consistent with the dataset used in Nelson et al. (2024).

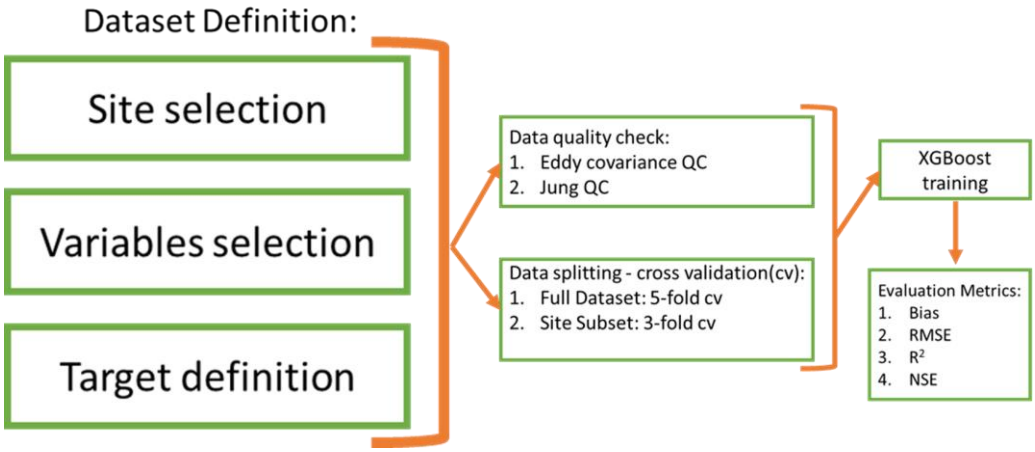

Figure 1: Algorithm training and evaluation pipeline.

AR-SLu (Garcia et al., 2016)	AR-TF1 (Kutzbach, 2021)	AR-Vir (Posse et al., 2016)	AT-Neu (Wohlfahrt et al., 2016)	AU-ASM (Cleverly and Eamus, 2016b)	AU-Ade (Beringer and Hutley, 2016c)
AU-Cpr (Meyer et al., 2016)	AU-Cum (Pendall and Griebel, 2016)	AU-DaP (Beringer and Hutley, 2016b)	AU-DaS (Beringer and Hutley, 2016f)	AU-Dry (Beringer and Hutley, 2016e)	AU-Emr (Schroder et al., 2016)
AU-Fog (Beringer and Hutley, 2016a)	AU-Gin (Macfarlane et al., 2016)	AU-RDF (Beringer and Hutley, 2016d)	AU-Rob (Liddell, 2016)	AU-TTE (Cleverly and Eamus, 2016a)	AU-Tum (Woodgate et al., 2016)
AU-Wac (Beringer et al., 2016b)	AU-Whr (Beringer et al., 2016a)	AU-Wom (Arndt et al., 2016)	AU-Ync (Beringer and Walker, 2016)	BE-Bra (Team and Centre, 2022)	BE-Dor (Team and Centre, 2022)
BE-Lcr (RI, 2021)	BE-Lon (Team and Centre, 2022)	BE-Maa (Team and Centre, 2022)	BE-Vie (Team and Centre, 2022)	BR-Npw (Vourlitis et al., 2022)	BR-Sa1 (Saleska, 2016)
BR-Sa3 (Goulden, 2016d)	CA-Cbo (Staebler, 2022)	CA-DB2 (Knox, 2022)	CA-DBB (Christen and Knox, 2022)	CA-ER1 (Wagner-Riddle, 2021)	CA-Gro (McCaughey, 2016)
CA-LP1 (Black, 2021)	CA-Man (Amiro, 2016b)	CA-NS2 (Goulden, 2016a)	CA-NS3 (Goulden, 2016b)	CA-NS4 (Goulden, 2016c)	CA-NS5 (Goulden, 2016g)
CA-NS6 (Goulden, 2016e)	CA-NS7 (Goulden, 2016f)	CA-Oas (Black, 2016b)	CA-Obs (Black, 2016a)	CA-Qfo (Margolis, 2016)	CA-SF1 (Amiro, 2016c)
CA-SF2 (Amiro, 2016a)	CA-SF3 (Amiro, 2016d)	CA-TP1 (Arain, 2016b)	CA-TP2 (Arain, 2016a)	CA-TP3 (Arain, 2022b)	CA-TP4 (Arain, 2016c)
CA-TPD (Arain, 2022a)	CG-Tch (Nouvellon, 2016)	CH-Aws (Team and Centre, 2022)	CH-Cha (Team and Centre, 2022)	CH-Dav (Team and Centre, 2022)	CH-Fru (Team and Centre, 2022)
CH-Lae (Team and Centre, 2022)	CH-Oe1 (Ammann, 2016)	CH-Oe2 (Team and Centre, 2022)	CN-Cha (Zhang and Han, 2016)	CN-Cng (Dong, 2016)	CN-Dan (Shi et al., 2016)
CN-Din (Zhou and Yan, 2016)	CN-Du2 (Chen, 2016k)	CN-Du3 (Shao, 2016b)	CN-HaM (Tang et al., 2016)	CN-Qia (Wang and Fu, 2016)	CN-Sw2 (Shao, 2016a)
CZ-BK1 (Team and Centre, 2022)	CZ-BK2 (Sigut et al., 2016)	CZ-KrP (Team and Centre, 2022)	CZ-Lnz (Team and Centre, 2022)	CZ-RAJ (Team and Centre, 2022)	CZ-Stn (Team and Centre, 2022)
CZ-wet (Team and Centre, 2022)	DE-Akm (Team and Centre, 2022)	DE-Geb (Team and Centre, 2022)	DE-Gri (Team and Centre, 2022)	DE-Hai (Team and Centre, 2022)	DE-HoH (Team and Centre, 2022)
DE-Hte (Team and Centre, 2020)	DE-Hzd (Team and Centre, 2022)	DE-Kli (Team and Centre, 2022)	DE-Lkb (Lindauer et al., 2016)	DE-Lnf (Knohl et al., 2016)	DE-Obe (Team and Centre, 2022)
DE-RuR (RI, 2022)	DE-RuS (Team and Centre, 2022)	DE-RuW (Team and Centre, 2022)	DE-Seh (Schneider and Schmidt, 2016)	DE-SfN (Klatt et al., 2016)	DE-Spw (Bernhofer et al., 2016)
DE-Tha (Team and Centre, 2022)	DE-Zrk (Sachs et al., 2016)	DK-Eng (Pilegaard and Ibrom, 2016)	DK-Fou (Olesen, 2016)	DK-Gds (RI, 2022)	DK-Sor (Team and Centre, 2022)
ES-Abr (Team and Centre, 2022)	ES-Agu (Team and Centre, 2022)	ES-Amo (Poveda et al., 2016)	ES-Cnd (Team and Centre, 2022)	ES-LJu (Team and Centre, 2022)	ES-LM1 (Team and Centre, 2022)
ES-LM2 (Team and Centre, 2022)	ES-LgS (Reverter et al., 2016b)	ES-Ln2 (Reverter et al., 2016a)	FI-Hyy (Team and Centre, 2022)	FI-Jok (Lohila et al., 2016)	FI-Ken (Team and Centre, 2022)
FI-Let (Team and Centre, 2022)	FI-Lom (Aurela et al., 2016a)	FI-Qvd (Team and Centre, 2022)	FI-Sii (Team and Centre, 2022)	FI-Sod (Aurela et al., 2016b)	FI-Var (RI, 2022)
FR-Aur (Team and Centre, 2022)	FR-Bil (Team and Centre, 2022)	FR-EM2 (RI, 2022)	FR-FBn (Team and Centre, 2022)	FR-Fon (Team and Centre, 2022)	FR-Gri (Team and Centre, 2022)
FR-Hes (Team and Centre, 2022)	FR-LBr (Berbigier and Loustau, 2016)	FR-LGt (RI, 2022)	FR-Lam (Team and Centre, 2022)	FR-Pue (Ourcival, 2016)	FR-Tou (RI, 2022)
GF-Guy (Team and Centre, 2022)	GH-Ank (Valentini et al., 2016b)	GL-Dsk (RI, 2022)	GL-NuF (Hansen, 2016)	GL-ZaF (Lund et al., 2016b)	GL-ZaH (Lund et al., 2016a)
IE-Cra (Team and Centre, 2022)	IL-Yat (Team and Centre, 2022)	IT-BCi (Team and Centre, 2022)	IT-BFt (RI, 2022)	IT-CA1 (Sabbatini et al., 2016c)	IT-CA2 (Sabbatini et al., 2016a)
IT-CA3 (Sabbatini et al., 2016b)	IT-Col (Matteucci, 2016)	IT-Cp2 (Team and Centre, 2022)	IT-Cpz (Valentini et al., 2016a)	IT-Isp (Gruening et al., 2016b)	IT-La2 (Cescatti et al., 2016)
IT-Lav (Team and Centre, 2022)	IT-Lsn (RI, 2022)	IT-MBo (Team and Centre, 2022)	IT-Noe (Spano et al., 2016)	IT-PT1 (Manca and Goded, 2016)	IT-Ren (Team and Centre, 2022)
IT-Ro1 (Valentini et al., 2016c)	IT-Ro2 (Papale et al., 2016)	IT-SR2 (Team and Centre, 2022)	IT-SRo (Gruening et al., 2016a)	IT-Tor (Team and Centre, 2022)	JP-MBF (Kotani, 2016b)
JP-SMF (Kotani, 2016a)	MX-Tes (Yepez and Garatuza, 2021)	MY-PSO (Kosugi and Takanashi, 2016)	NL-Hor (Dolman et al., 2016a)	NL-Loo (Team and Centre, 2020)	PA-SPn (Wolf et al., 2016b)
PA-SPs (Wolf et al., 2016a)	PE-QFR (Griffis and Roman, 2021)	RU-Che (Merbold et al., 2016)	RU-Cok (Dolman et al., 2016b)	RU-Fy2 (Team and Centre, 2022)	RU-Fyo (Team and Centre, 2022)
RU-Ha1 (Belelli et al., 2016)	SD-Dem (Ardö et al., 2016)	SE-Deg (Team and Centre, 2022)	SE-Htm (Team and Centre, 2022)	SE-Lnn (Team and Centre, 2020)	SE-Nor (Team and Centre, 2022)
SE-Ros (Team and Centre, 2022)	SE-Svb (Team and Centre, 2022)	SJ-Adv (Christensen, 2016)	SJ-Blv (Boike et al., 2016)	SN-Dhr (Tagesson et al., 2016)	US-A32 (Billesbach et al., 2022)
US-AR1 (Billesbach et al., 2016b)	US-AR2 (Billesbach et al., 2016a)	US-ARM (Biraud et al., 2022)	US-ARb (Torn, 2016b)	US-ARc (Torn, 2016a)	US-Atq (Zona and Oechel, 2016a)
US-BZB (Euskirchen, 2022b)	US-BZF (Euskirchen, 2022c)	US-BZS (Euskirchen, 2022d)	US-BZo (Euskirchen, 2022a)	US-Bi1 (Rey-Sanchez et al., 2022b)	US-Bi2 (Rey-Sanchez et al., 2022a)
US-Blo (Goldstein, 2016)	US-CF1 (Huggins, 2021)	US-CF2 (Huggins, 2022c)	US-CF3 (Huggins, 2022a)	US-CF4 (Huggins, 2022b)	US-CRT (Chen and Chu, 2016b)
US-CS1 (Desai, 2022a)	US-CS2 (Desai, 2022c)	US-CS3 (Desai, 2022d)	US-CS4 (Desai, 2022b)	US-Cop (Bowling, 2016)	US-EDN (Oikawa, 2021)
US-GBT (Massman, 2016)	US-GLE (Massman, 2022)	US-Goo (Meyers, 2016b)	US-HB1 (Forsythe et al., 2021)	US-HWB (Goslee, 2022)	US-Ha1 (Munger, 2016)
US-Hn3 (Liu et al., 2022)	US-Ho2 (Hollinger, 2022)	US-IB2 (Matamala, 2016)	US-ICs (Euskirchen et al., 2022a)	US-ICt (Euskirchen et al., 2022b)	US-Ivo (Zona and Oechel, 2016b)
US-Jo2 (Vivoni and Perez-Ruiz, 2022)	US-KFS (Brunsell, 2022a)	US-KLS (Brunsell, 2022b)	US-KS1 (Drake and Hinkle, 2016a)	US-KS2 (Drake and Hinkle, 2016b)	US-KS3 (Hinkle, 2022)
US-LWW (Meyers, 2016a)	US-Lin (Fares, 2016)	US-Los (Desai, 2016c)	US-MMS (Novick and Phillips, 2022)	US-MOz (Wood and Gu, 2022)	US-Me1 (Law, 2016c)
US-Me2 (Law, 2022)	US-Me3 (Law, 2016a)	US-Me4 (Law, 2016e)	US-Me5 (Law, 2016d)	US-Me6 (Law, 2016b)	US-Mpj (Litvak, 2021)
US-Myb (Sturtevant et al., 2016)	US-NGB (Torn and Dengel, 2021)	US-NR1 (Blanken et al., 2022)	US-Ne1 (Suyker, 2022)	US-Ne2 (Suyker, 2016b)	US-Ne3 (Suyker, 2016a)
US-ONA (Silveira, 2021)	US-ORv (Bohrer, 2021)	US-OWC (Bohrer and Kerns, 2022)	US-Oho (Chen et al., 2016)	US-PFa (Desai, 2016d)	US-Prr (Kobayashi and Suzuki, 2016)
US-Rms (Flerchinger, 2022c)	US-Ro1 (Baker et al., 2022)	US-Ro4 (Baker and Griffis, 2022a)	US-Ro5 (Baker and Griffis, 2021)	US-Ro6 (Baker and Griffis, 2022b)	US-Rwe (Flerchinger and Reba, 2022)
US-Rwf (Flerchinger, 2022a)	US-Rws (Flerchinger, 2022b)	US-SRC (Kurc, 2022)	US-SRG (Scott, 2016a)	US-SRM (Scott, 2016b)	US-Sne (Shortt et al., 2022)
US-Snf (Kusak et al., 2022)	US-Sta (Ewers and Pendall, 2016)	US-Syv (Desai, 2016b)	US-Ton (Baldocchi and Ma, 2016)	US-Tw1 (Valach et al., 2021)	US-Tw2 (Sturtevant et al., 2022)
US-Tw3 (Chamberlain et al., 2022)	US-Tw4 (Sanchez et al., 2016)	US-Tw5 (Valach et al., 2022)	US-Twt (Baldocchi, 2016)	US-UM3 (Bohrer, 2022)	US-UMB (Gough et al., 2016)
US-UMd (Gough et al., 2022)	US-Var (Baldocchi et al., 2016)	US-WCr (Desai, 2016a)	US-WPT (Chen and Chu, 2016a)	US-Whs (Scott, 2016d)	US-Wi0 (Chen, 2016g)
US-Wi1 (Chen, 2016e)	US-Wi2 (Chen, 2016j)	US-Wi3 (Chen, 2016b)	US-Wi4 (Chen, 2016d)	US-Wi5 (Chen, 2016a)	US-Wi6 (Chen, 2016h)
US-Wi7 (Chen, 2016i)	US-Wi8 (Chen, 2016c)	US-Wi9 (Chen, 2016f)	US-Wjs (Litvak, 2022)	US-Wkg (Scott, 2016c)	US-xBR (Network, 2022)

Table 2: List of experimental sites used, from Nelson et al. (2024).

The complete experimental dataset represents only one of the site configurations used in this work. With the specific goal of investigating the Aosta Valley region and the Torgnon site, we partitioned the available sites according to their similarity to Tor,

generating different site subsets. This intentional reduction in data availability was designed to evaluate model performance under conditions representative of sites with atypical characteristics, such as Torgnon. Among the tested configurations, the most promising subsets were: Local (geographically closest sites), Alps (sites located in Alpine regions), Mounts (mountain sites, with elevation ≥ 1200 m), Grassland and All (all available sites). Following preliminary tests the Local subset was excluded due to its high redundancy with the Alps, whereas the Mounts subset shared only five sites with Alps (less than half of each), ensuring complementary sampling. The final subdivision of experimental sites is summarized in Table 3.

Alps		Mountain			Grassland								
CH-Cha	IT-Lav	AU-Tum	IT-Lav	US-Me2	AT-Neu	AU-Stp	CN-Cng	DE-RuR	IT-MBo	US-AR1	US-KFS	US-Sne	US-xDC
CH-Dav	IT.Mbo	CH-Dav	IT-MBo	US-NR1	AU-Cpr	AU-Ync	CN-Dan	DK-Eng	IT-Tor	US-AR2	US-KLS	US-Snf	US-xKA
CH-Lae	IT-PT1	DE-Lkb	IT-Ren	US-SRG	AU-DaP	BE-Dor	CN-Du2	ES-Abr	NL-Hor	US-ARb	US-LWW	US-Var	US-xKZ
CH-Oe1	IT-Ren	ES-LJu	IT-Tor	US-Sta	AU-DaS	CG-Tch	CN-Du3	ES-LM1	PA-SPs	US-ARc	US-NGC	US-Wjs	US-xNG
CH-Oe2	IT-Ro1	ES-LgS	US-Blo	US-Wkg	AU-Dry	CH-Aws	CN-HaM	ES-LM2	RU-Ha1	US-Cop	US-ONA	US-Wkg	
DE-SfN	IT-SR2	ES-Ln2	US-Cop	US-Whs	AU-Emr	CH-Cha	CN-Sw2	FR-Mej	SD-Dem	US-Goo	US-Ro4	US-xAE	
IT-Isp	IT-Sro	IT-Col	US-GLE		AU-GWW	CH-Fru	CZ-BK2	FR-Tou	SN-Dhr	US-Hn2	US-SRG	US-xCL	
IT-La2	IT-Tor	IT-La2	US-GBT		AU-Rig	CH-Oe1	DE-Gri	GL-ZaH	US-A32	US-IB2	US-Seg	US-xCP	

Table 3: List of experimental site subsets used in the study.

For the analysis of input features, predictors were grouped into the five classes previously described and summarized in Table 1. Each class could be either included or excluded in individual experiments, resulting in $2^5 = 32$ theoretical possible feature combinations, of which approximately half were explored. This categorization was based on the physical origin of the variables and aimed to identify the optimal predictor configuration for each modeling task. In view of potential spatial applications of the model beyond the eddy covariance tower scale, it is worth to note that not all variables are universally available. A future refinement could involve a further separation of variable classes into those locally available and those not directly measurable. The current model is single output based, with GPP used as the target variable. However, since eddy covariance measurements provide NEE that can be partitioned into GPP and Reco, a logical next step would be to train an additional model to estimate one of the remaining fluxes (e.g. NEE) so requiring a dedicated model configuration.

For the training phase, we adopted a leave-one-fold-out cross-validation approach specifically designed to account for the spatial structure of the dataset. Rather than applying cross validation to individual data records, we divided the considered ensemble of site into k folds and iteratively left out one entire fold at time. At each iteration, all data records belonging to the sites included in the excluded fold were used as the test set, while the model was trained on the data from the remaining k-1 folds. In this way every site contributes once to model validation, and all its associated temporal data are used consistently either for training or for testing, so never for both simultaneously. To avoid spatial autocorrelation, nearby towers were never split across folds, so all sites located within 0.05° of each other were assigned to the same fold. The model was trained k times, each time using data from k-1 folds as the training set (randomly aggregated) and evaluated on the excluded fold. The computed performance metrics can be averaged across all folds to assess generalizability while capturing both spatial and temporal variability. The number of folds was not fixed a priori but instead depended on the number of sites in each dataset configuration. For robustness and comparability, a k-fold cross-validation scheme was applied consistently across datasets. We used k = 10 for the Full dataset (≈ 300 sites), which provided a balanced trade-off between computational efficiency and statistical reliability, ensuring sufficient diversity within each fold while maintaining independence between training and validation data. For the smaller subsets that contains between 10 and 30 sites, we used k = 3, reflecting the reduced sample size and the need to preserve an adequate number of sites in each fold. A smaller k also reduces the risk of high variance in validation metrics that can arise from over-partitioning small datasets. At each iteration, the model was trained on k-1 folds and evaluated on the remaining fold.

	Eddy QC	Jung QC
NEE	✓	
GPP		✓
TA	✓	✓
VPD	✓	✓
WS	✓	
PA	✓	
SW_IN	✓	✓

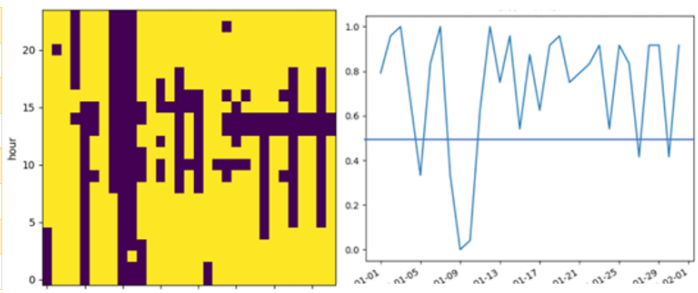

Figure 2: On the left, a list of variables for which quality control can be applied is shown. An example of a daily case is also provided: for each hour of January 2010 at Torgnon, data are assigned a quality flag of 0 or 1. The results associated with a given day are included in the calculation of metrics only if at least half of the hourly data for that day meet the quality criteria. For instance, around January 10th, some days were discarded due to insufficient data quality, whereas toward the end of the month, data quality improved.

Only records passing both QC/QA and Jung QC quality control procedures were retained for each variable. These two independent filters ensure that only physically consistent and statistically reliable observations are included in the analysis. In the case of aggregated data (e.g., daily), we adopted a conservative filtering strategy to maintain data integrity. Specifically, an aggregated value was retained only if at least 50% of the contributing original half-hourly or hourly records met the quality criteria. This approach minimizes the propagation of uncertainty from low-quality measurements into the aggregated dataset while preserving sufficient temporal coverage for model training and evaluation. Figure 2 illustrates a typical example of this filtering process, applied for the Torgnon site (January 2010). Days for which fewer than half of the hourly measurements were flagged as valid were excluded from subsequent analyses, while periods with consistently high-quality data were retained. This ensures that the resulting dataset reflects true ecosystem dynamics rather than artifacts introduced by data gaps or sensor noise.

So each performed experiment consisted of training the model to predict a specific target variable using a selected combination of input predictors, with data record extracted from one of the previously defined subsets of experimental sites. The experiments were designed to assess both the contribution of different feature groups and the spatial transferability of the model across varying environmental and climatic conditions. Model performance was evaluated through several statistical metrics commonly used in ecosystem flux modelling. The Root Mean Square Error (RMSE) quantifies the average

magnitude of prediction errors, providing a direct measure of model accuracy and being particularly sensitive to large deviations between observed and predicted values. The Coefficient of Determination (R^2) expresses how much of the observed variance is explained by the model, thus reflecting its ability to reproduce temporal and spatial variability in the data. The bias assesses the systematic deviation between predictions and observations, revealing whether the model tends to overestimate or underestimate fluxes on average. Finally the Nash-Sutcliffe Efficiency (NSE) offers a more comprehensive measure of model skill by comparing the predictive performance to that of a model that simply uses the mean of observations as prediction. Unlike RMSE or R^2, which focus on accuracy or explained variance, NSE evaluates the overall efficiency of the model in reproducing the observed dynamics. Values close to 1 indicate high predictive skill, whereas values near or below 0 suggest that the model performs no better than the mean of the observed data. These metrics were computed for each fold and then can be averaged over all folds to obtain a robust estimate of model generalization.

3.3.4 Local dataset

Understanding and quantifying carbon fluxes in alpine ecosystems is critical for assessing regional carbon budgets and their response to climate variability. The Torgnon site is an alpine grassland ecosystem located in the Aosta Valley (north-western Italian Alps) at an altitude of 2160 m a.s.l. It is part of the FLUXNET network and hosts an eddy-covariance (EC) tower that provides continuous, high-frequency measurements of CO_2 exchange including NEE. The site is characterized by long periods of snow cover (≈ 200 days per year) and low mean annual temperature (≈ 3 °C), which restrict the growing season and strongly influence carbon flux dynamics. Vegetation is dominated by alpine grassland species, with limited tree cover, making the site particularly suitable for studying carbon exchange processes in high-elevation environments.

A characterization of the Torgnon dataset, including meteorological variables, carbon fluxes, and vegetation indices, is reported in Table 4. It summarizes the main statistical properties of the daily aggregated, quality-controlled measurements.

Variable (unit)	Min	Median	Max
TA (°C)	–15.19	3.25	17.99
SW_IN_POT (W m^{-2})	115.9	311.41	488.4
SW_IN (W m^{-2})	1.09	153.15	392.1
VPD (kPa)	0.02	2.34	13.02
PA (kPa)	74.94	78.4	79.96
WS (m s^{-1})	0.22	1.23	5.65
u*(m s^{-1})	0.042	0.119	0.483
NEE (µmol CO_2 m^{-2} s^{-1})	–7.78	0.52	3.71
RECO (µmol CO_2 m^{-2} s^{-1})	0.21	1.94	7.34
GPP (µmol CO_2 m^{-2} s^{-1})	–1.21	1.09	12.57
NDVI	0.094	0.219	0.81

Table 4: Summary statistics for meteorological, flux, and vegetation variables measured at the Torgnon eddy-covariance site. For each variable, the minimum, maximum, and median values are reported. Only quality-controlled data were included.

The main challenge addressed in this study is the development of a predictive model capable of estimating carbon fluxes at Torgnon based on meteorological variables, remote sensing indices, and other environmental drivers. By leveraging the EC measurements for both training and validation, a ML model can learn the relationships between environmental conditions and carbon flux dynamics at a local scale. This approach allows for an assessment of model performance under the highly variable alpine conditions, characterized by long snow periods, low temperatures, and high diurnal and seasonal variability in radiation and moisture.

However, training a model exclusively on data from Torgnon would inevitably lead to overfitting. To extend the applicability of the model to nearby areas within the Aosta

Valley, it is therefore necessary to include a larger and more diverse training dataset that captures the spatial and ecological variability of the region (D'Amico et al., 2020). The inclusion of Torgnon within the FLUXNET network facilitates this approach, allowing the adaptation of frameworks such as FLUXCOM and XBASE to develop a robust, data-driven model integrating EC, meteorological, and satellite data for quantitative estimation of CO_2 fluxes at both site and regional scales.

Moreover snow cover is a dominant feature of the Torgnon site, typically persisting for more than half of the year and strongly influencing the length of the growing season, and consequently carbon flux dynamics. Training ML models on a subset of sites with environmental characteristics similar to Torgnon can help account for this snow-related variability, thereby improving model generalization to Torgnon-like conditions and Aosta Valley. Furthermore satellite-derived vegetation indices such as NDVI and so NIRv, provide valuable information on vegetation distribution and seasonal productivity patterns in alpine grasslands (Filippa et al., 2022), which can support the interpretation of carbon flux dynamics at the Torgnon site. Such indices allow the model to better capture phenological dynamics and carbon uptake patterns in high-altitude grasslands, supporting more accurate site-specific and regional predictions.

3.4 RESULTS AND DISCUSSION

In this study, our primary objective is to train a model capable of estimating GPP, providing a quantitative assessment of CO_2 fluxes at the local scale with particular focus on the Aosta Valley region, where the Torgnon site is located. Unless otherwise specified, all results presented refer to the prediction scenario in which Torgnon is included in the left-out fold during cross-validation. This setup allows us to evaluate model capacity to generalize to unseen sites with similar ecological and climatic characteristics. The analysis can be systematically replicated for any other FLUXNET site. Two main modeling frameworks are considered: a global model, trained across the full set of sites to identify the configuration yielding the best overall performance, and a local model, specifically designed to rely on predictors that are potentially available at

regional or local scales. The latter aims to support future applications of the model for regional upscaling, enabling repeated estimations of carbon fluxes across the Aosta Valley using locally accessible meteorological and remote sensing data.

3.4.1 Analysis of Best Feature Set and Training Configuration in Daily Prediction

The results of the training process are illustrated in Figure 3, focusing on the comparison between different combinations of training datasets and predictor sets. As an example of a local configuration, we considered the case including all predictors except land surface temperature (LST) and plant functional type (PFT) variables. Model performance under different training conditions, with the associated metric values reported for each case.

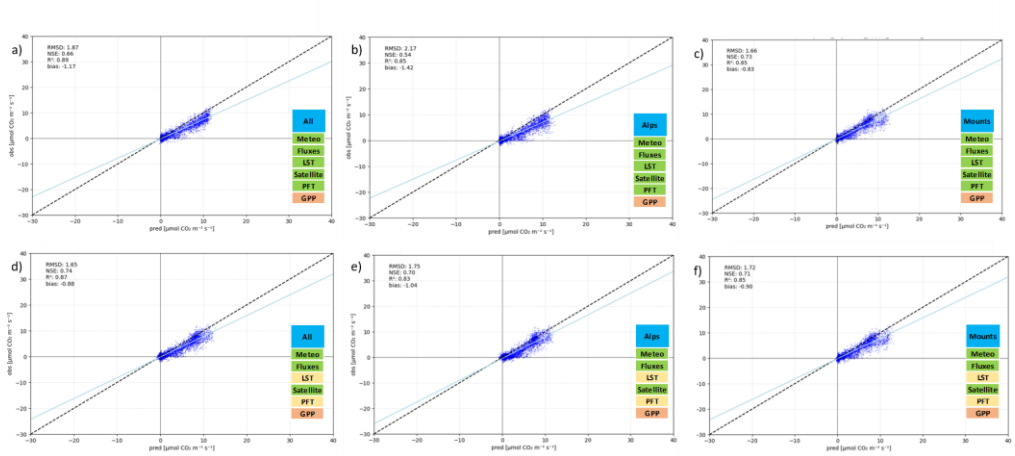

Figure 3: Scatterplots of observed versus predicted daily GPP at the Torgnon site under four training conditions. The top row (a, b, c) shows predictions obtained using the full set of predictors, while the bottom row (d, e, f) excludes LST and PFT variables. The left panels correspond to models trained on all experimental sites, the middle to the Alps subset, and the right to the Mounts subset. Main performance metrics (RMSE, R^2, bias, NSE) are reported for each panel.

Predictions correspond to daily estimates for all days in which the predictors were available and met both quality control criteria. In all scenarios, a negative bias is observed, indicating a general overestimation of predicted values compared to observations. Overall, model performance is higher when using the full set of predictors compared to the reduced set excluding LST and PFT variables. From this preliminary analysis, the local model configuration (Alps subset with selected features) appears to outperform the global model (all sites, all predictors).

The scatterplots indicate that the model captures the daily dynamics of GPP at Torgnon reasonably well. Daily GPP predictions consistently fall within the range of approximately -2 to 13 μmol CO_2 m^{-2} s^{-1}, in line with expectations for the studied ecosystem. However, the persistent negative bias highlights the overestimation of carbon uptake, particularly under extreme climatic conditions such as those found at high-elevation sites like Torgnon, which are underrepresented in global training datasets. This behaviour aligns with the findings of Tramontana et al. (2016), who reported similar limitations of machine learning models within the FLUXCOM framework.

This first result suggests that model generalizability can be improved by restricting the training to a locally informed subset of sites with ecological conditions similar to Torgnon. This approach enhances the model's ability to reproduce site-specific phenological patterns and mitigates the overestimation observed in global models. However, it is likely highly sensitive to random fluctuations involved in the training process, both in the construction of the k folds and in the iterative model fitting itself. To assess the stability of the results, we therefore repeated the training procedure using multiple initialization seeds for the random number generator. As shown in Figure 4, the obtained statistics are strongly influenced by the randomness: for instance, RMSE values for the Full dataset (All) vary from just below 1.5 to nearly 3.5 μmol CO_2 m^{-2} s^{-1} depending on the random seed used.

It is worth to note that selecting a seed a posteriori to obtain the model with the best performance is not valid, as it would effectively interfere with the random number generation process.

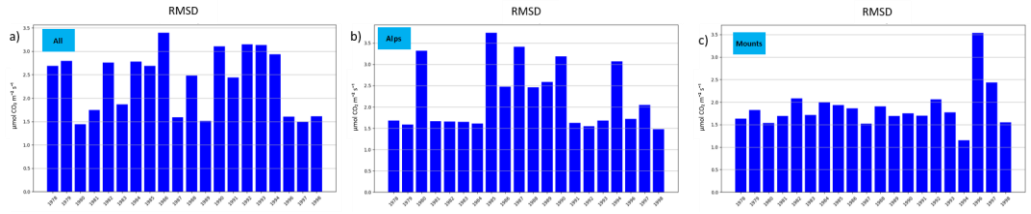

Figure 4: shows a bar plot of RMSD for Torgnon site predictions at daily frequency, illustrating the variability induced by different random seeds. Three site configurations specific setups were explored: a) All, b) Alps, c) Mounts.

The influence of stochastic variables cannot be ignored. These stochastic effects can lead to over optimistic or under optimistic evaluations of model capabilities, mainly when the training process is so sensitive to initialization or random sampling. Similar observations have been reported in different studies, even if in contexts different from the present work. Vos et al. (2025) showed that random seed selection can significantly influence predictive performance, feature importance, and model interpretability. By repeating trials with different seeds and aggregating the results, they were able to obtain more stable and reproducible estimates. Similarly Schader et al. (2024) demonstrated that the choice of random seed can yield divergent scientific interpretations from the same dataset, and proposed methods to stabilize results across seeds. Although these studies focus on different applications, their findings converge on the general principle that stochastic variability in machine learning models must be explicitly considered.

We analyzed the effect of fold composition on model performance at the Torgnon site with the alps dataset, using RMSE = 1.7 as a threshold between higher-performing (RMSE $\leq$ 1.7) and lower-performing runs (RMSE > 1.7). To assess which sites contribute most to predictive accuracy, we considered the frequency with which other sites appeared outside the fold of IT-Tor, i.e., available for training and able to provide informative signals.

For higher-performing runs, the most informative sites were IT-SRo and IT-SR2, appearing outside the fold in 8 cases each, along with IT-Ren (7) and IT-PT1 (6). Despite some environmental differences compared to Torgnon as lower elevation (IT-SRo, IT-SR2, IT-PT1) or different vegetation type (forest instead of grassland), the presence of these sites in the training folds helped the model capture general GPP patterns and reduce prediction errors. IT-Ren, due to its higher elevation and montane/subalpine context, is the site most similar to Torgnon, but the other sites also contributed significantly by providing additional variability for model learning.

In contrast, during lower-performing runs, these informative sites were less frequently available for training (IT-SRo and IT-SR2: 4 times each; IT-Ren: 5; IT-PT1: 5), coinciding with higher RMSE values. Less similar sites, such as CH-Oe1, CH-Oe2, and IT-Isp, appeared outside the fold in 4 to 6 cases for high-performance runs, indicating a moderate contribution to predictions. In particular, CH-Oe1 and CH-Oe2 have intermediate elevations and forest vegetation: they are not alpine grasslands like Torgnon but provide partially relevant information on montane dynamics. IT-Isp, with low elevation and very different vegetation, contributes less as an "informative" site for predicting GPP at Torgnon.

A further analysis therefore requires the use of a robust evaluation method, for example by comparing the medians of the performance distribution obtained by varying only the random seed. This approach allows assessing the performance while accounting for stochastic variability in model training.

A comparative analysis of the results obtained for the Torgnon site is shown in Figure 5. It can be observed that the GPP at Torgnon is statistically better captured when using regionally restricted datasets (Alps and Mounts) compared to the global dataset (All). Specifically when the model is trained using all predictors, the Alps dataset provides the best performance. However, when the number of predictors is reduced, the Mounts dataset yields better results. In this latter configuration, the Alps and All datasets show similar performance, with a slight reduction in overestimation for the Alps subset.

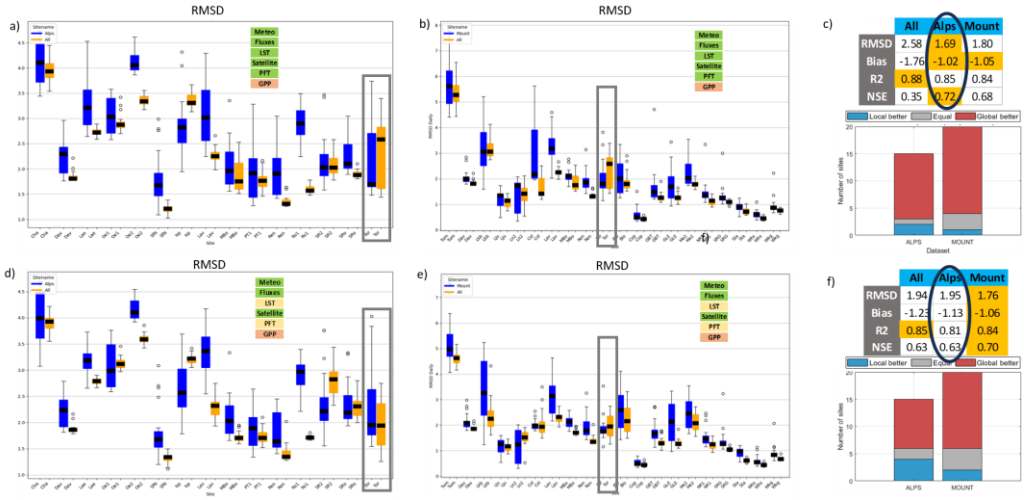

Figure 5: RMSE boxplots comparing pairs of site subsets used for training (Alps vs. All in panels a and d, Mounts vs. All in panels b and e) across different experimental sites, shown for two variable configurations: full set on the top panels (a,b) and reduced set on the bottom ones (d,e). In each boxplot, the black central line represents the median, the colored box shows the interquartile range (IQR, 50% of the data), and the whiskers extend to the last data point within 1.5 × IQR. Outliers are indicated by empty circles. For each variable configuration, two additional panels are shown on the right. The first one (c) highlights statistics for the Torgnon site, with the best-performing result statistically highlighted. The second one(f) shows the number of sites where a model trained on a subset of sites (local) outperforms the model trained on all sites (global). The stacked bars indicate, for each comparison, the number of sites where the local model performs better (blue), similarly (grey), or worse (red) than the global model.

Looking more closely at the other experimental sites, it is possible to quantify how often local models outperform the global one. Figure 5 also shows two panels illustrating that, in general, the model trained on all sites performs better, but there are exceptions. In particular, when considering the Alps subset with the full predictor set, only two sites (IT-Tor and ISP) outperform All, with one additional site showing comparable performance. When reducing the predictor set, the local model improves slightly,

outperforming the global model in four sites (Excluding IT-Tor) and achieving similar results in two. However, differences are not very pronounced except for three sites (IT-Isp, IT-SR2, IT-Tor). This improvement for IT-Tor is also reflected in a smaller interquartile range for the Alps model compared to All when using the full predictor set. Otherwise, the All model demonstrates greater stability, which is reasonable given the smaller training sample size in the Alps subset. The comparison between the All and Mount subset is even more unbalanced: under the full predictor set, the local model performs better in only one case (IT-Tor) and similarly in three, while with reduced predictors it performs better in two sites (IT-Tor and ES-Ln2, although ES-Ln2 is considerably less stable) and achieves similar performance in four out of twenty sites. In general, All models show greater stability compared to the Mounts subset when accounting for stochastic variability in training.

It is worth to note that the three sites IT-Tor, IT-SR2, and IT-Isp, present distinct environmental contexts: IT-Tor is a grassland site with a long snow cover period (~ 200 days/year) and low mean annual temperature (~ 3 °C), whereas IT-Isp, and IT-SR2 are forested (deciduous broadleaf and evergreen needleleaf, respectively) with very limited snow cover (< 5 days/year) and mean annual temperatures above 12 °C.

Overall, using a local dataset constructed to match characteristics similar to Torgnon results in slightly better statistical performance compared to using the full dataset, along with reduced computational costs during training. This analysis can be replicated for any site, but it is advisable to first assess the characteristics of the target site and explore different configurations of site subsets and predictor sets. Notably, preliminary analyses were also conducted using datasets composed exclusively of sites sharing the same grassland PFT as Torgnon.

These yielded poor model performance and are therefore not shown in full. Figure 6 illustrates two representative plots. Panel (a) presents the RMSD values obtained for multiple experimental runs with different random seeds. The results reveal that the RMSE remains nearly constant across random initializations, yet overall performance is unsatisfactory. This suggests that the characteristics of the Torgnon site are poorly

represented within the available training data, leading to inconsistent folds and limited generalization. Panel (b) shows the corresponding scatterplot of observed versus predicted GPP for one representative run. The distribution clearly indicates a systematic overestimation across almost the entire range of observed values. These findings suggest that model performance is not primarily constrained by PFT similarity, but rather by broader biome-level differences, which exert a stronger influence on predictive accuracy.

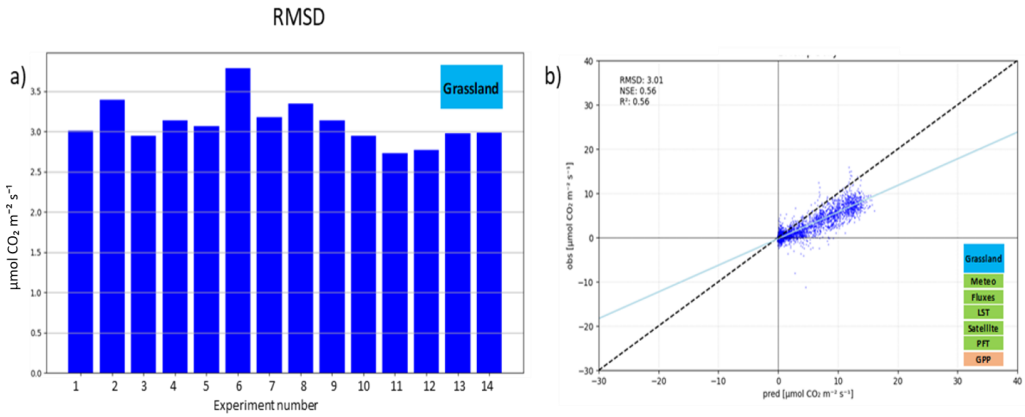

Figure 6. Grassland results. (a) Bar plot of RMSD values for daily GPP predictions at the Torgnon site, obtained using different random seeds. (b) Scatterplot of observed versus predicted values for one representative experiment (corresponding to one of the random seeds shown in panel a).

These observations also support conclusions from Jung et al. (2020) and Wei et al. (2021), who emphasized that local environmental representativeness and the spatial density of flux tower networks are critical factors influencing the reliability of upscaled flux estimate, particularly in complex alpine terrain with sharp topographic and vegetation gradients, as highlighted by Coates et al. (2021). The improved performance obtained using a regionally filtered dataset suggests that carefully selecting sites with similar environmental conditions can reduce structural biases inherited from global-scale models.

3.4.2 Time series analysis

The GPP time series shown in Figure 7 highlight two distinct temporal patterns. These series were constructed by overlaying estimated and observed daily values throughout the year, excluding days that did not meet the QC/QA criteria. Only results obtained with the local models are presented-specifically, the Alps based model trained with the full set of predictors, and the model trained on the Mountain subset using the reduced predictor set.

No clear interannual trends are observed, with GPP conditions at the Torgnon site remaining relatively stable across years. The Mount model (panel c) exhibits a slight but consistent overestimation, whereas the Alps model (panel a) tends to produce larger fluctuations, including sharper peaks and a faster decrease toward zero during periods of negligible productivity (GPP $\approx$ 0). This systematic overestimation was also evident in the time-independent analysis, where models with higher accuracy consistently displayed smaller biases.

At the hourly scale (panel b), the Alps model shows good temporal stability but again reveals a noticeable overestimation, which propagates into higher daily mean values. This bias is particularly pronounced during the spring and autumn transition periods, when daily mean GPP tends to be overpredicted. This effect becomes even clearer in the yearly cycle derived by averaging daily results across all selected cases. A similar behaviour is observed for the Mount model (panel d), though with a weaker tendency toward overestimation.

In both models, the overall temporal pattern is comparable: the Mount model exhibits smoother variations in both peak and low-productivity periods, while the Alps model produces sharper responses. Consequently, during peak growing-season conditions the Alps model tends to overestimate GPP relative to Mount, whereas in nighttime or winter periods, Mount remains slightly overestimated but smoother, while Alps more rapidly approaches zero.

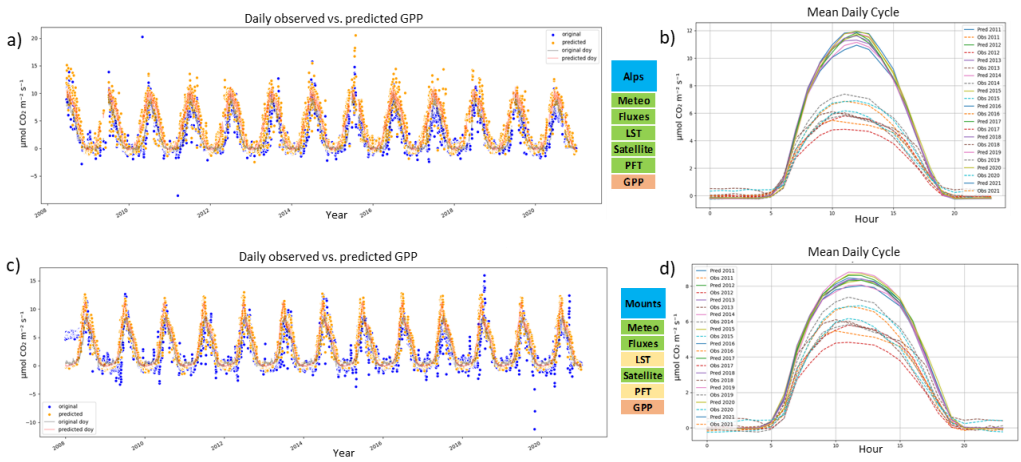

Figure 7: Comparison of the temporal evolution between observed and predicted GPP values, obtained via cross validation. The top panels (a and b) refer to the Alps-based model trained with all predictors, while the bottom panels (c and d) correspond to the Mount configuration using the reduced predictor ensemble. Panels (a and c) display daily GPP values for each day of the year, whereas panels (b and d) show the mean daily cycle, representing the average of estimated and observed values across all years for each hour of the day.

The annual cycle of GPP at the Torgnon site, obtained by averaging daily estimates across the day of year (DOY), is presented in Figure 8. The figure includes the full model trained on all sites with both predictor sets, as well as the two local configurations, namely Alps with all predictors and Mount with the reduced predictor set. In all four cases, the models exhibit some degree of overestimation. The two panels corresponding to the local datasets (b, d) better capture the phenology, with overestimation largely confined to the spring period. In contrast, in the two panels on the left (a, c), overestimation remains substantial even in autumn. A plausible explanation is that these biases are related to the incomplete representation of snow-related processes. Transitioning from the Alpine to the Mountain subset leads to further improvements, particularly when using the reduced set of predictors. While including all predictors

allows the model to partially account for snow effects through Land Surface Temperature, this is insufficient to fully resolve the observed biases.

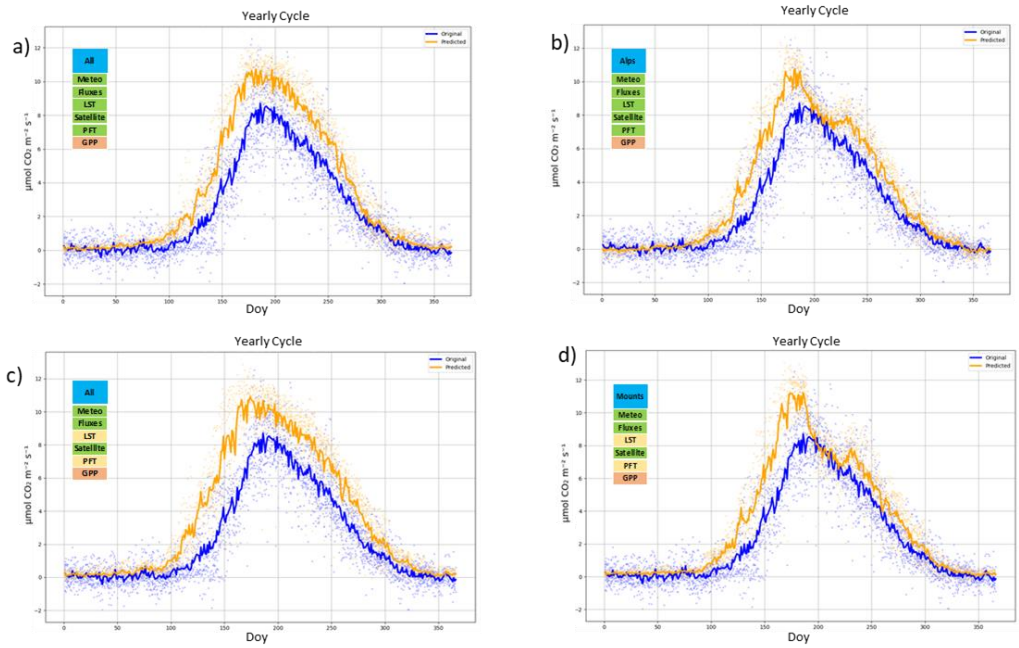

Figure 8: Yearly cycle as daily data average comparison of prediction and measured. The top row (a,b) corresponds to models trained with all predictors, while the bottom row (c,d) corresponds to the reduced predictor set. The left column (a,c) shows results using the full set of experimental sites, while the right column (b,d) shows results using local datasets.

Snow dynamics play a fundamental role in shaping the functioning of alpine ecosystems such as Torgnon. The site is typically snow-covered for more than half of the year, strongly limiting photosynthetic activity and carbon exchange. As reported by Galvagno et al. (2013), an exceptionally short snow season resulted in an extended CO_2 uptake period and nearly doubled the annual carbon sequestration. This underscores the high sensitivity of alpine grasslands to both the timing and duration of snow cover. More

generally, Torgnon is characterized by persistent and recurrent snow presence, which presents a challenge for predictive modeling. The observed tendency of our model to overestimate carbon uptake during snow-covered and transitional periods reflects the difficulty of data-driven approaches in accurately capturing snow-related processes and their influence on photosynthetic dynamics. When the model was trained exclusively on sites with environmental conditions similar to Torgnon (local subset), the overestimation was largely confined to the spring period, indicating an improved capacity to reproduce the site's phenological patterns.

3.5 CONCLUSIONS

In this study, we developed and evaluated the performance of a machine learning algorithm designed to estimate GPP in the Aosta Valley, with a specific focus on the high-altitude alpine grassland site of Torgnon. The training procedure followed the FLUXCOM pipeline, but we introduced a targeted sub-site and feature selection approach, demonstrating that locally trained models can improve predictive performance. In particular this result focus on ecosystems that are poorly represented in global datasets such as Torgnon, characterized by long snow-cover periods and strong seasonal variability.

Our results highlight several key findings. The model trained on the local dataset more effectively captured site-specific phenological dynamics at Torgnon, while also reducing the overestimation of GPP during transitional periods. This improvement was shown to be robust against the stochastic variability inherent in the training process. When varying the distribution of experimental sites across training folds and the initialization conditions of the algorithm, model performance exhibited significant variation, with RMSE values differing by up to a factor of 2.5 between the best and worst cases.

Feature selection also emerged as a crucial step, revealing that using a reduced predictor set can outperform the global model. In particular the integration of meteorological, flux, and remote-sensing drivers produced promising results, suggesting the potential for spatial upscaling applications. The performance of the global model at Torgnon was

linked to its limited ability to reproduce snow-driven dynamics, which strongly affect carbon exchange in alpine ecosystems. Instead the reduced and regionally focused dataset provided improving prediction accuracy by mitigating this limitation.

This approach shows that training while leveraging on a subset of ecologically similar sites helps reduce the sparsity of the ill-posed inversion problem, enabling a more realistic reproduction of snow-driven phenology. Overall, the proposed framework provides a promising pathway for generating high-resolution local carbon flux maps, supporting regional carbon budgeting and climate mitigation strategies. Moreover, the methodology could be extended to other regions with complex environmental conditions by identifying subsets of experimental sites with similar characteristics. Future developments could also include explicit snow related metrics such as the Normalized Snow and Vegetation Index (NSVI), to further enhance model efficiency in data-driven upscaling across the Aosta Valley.

3.6 References

Aubinet, M.; Vesala, T.; Papale, D., Eds. 2012. Eddy Covariance: A Practical Guide to Measurement and Data Analysis. Springer: Dordrecht, The Netherlands. https://doi.org/10.1007/978-94-007-2351-1

Badgley, G., Field, C. B., Berry, J. A., 2017. Canopy near-infrared reflectance and terrestrial photosynthesis. Science Advances, vol. 3, no. 3, e1602244. https://doi.org/10.1126/sciadv.1602244

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw U, K. T., Pilegaard,

K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Journal of Geophysical Research: Atmospheres, vol. 106, no. D17, pp. 2415–2415. https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2

Baldocchi, D.; Aubinet, M.; Vesala, T.; Papale, D. 2014. Measuring fluxes of trace gases and energy between ecosystems and the atmosphere—The state and future of the eddy covariance method. Global Change Biology, 20, 3600–3609. https://doi.org/10.1111/gcb.12649

Baldocchi, D. D., 2020. How eddy covariance flux measurements have contributed to our understanding of Global Change Biology. Global Change Biology, vol. 26, no. 1, pp. 242–260. https://doi.org/10.1111/gcb.14807

Broadbent, A. A. D., Snell, H. S. K., Michas, A., Pritchard, W. J., Newbold, L., Cordero, I., Goodall, T., Schallhart, N., Kaufmann, R., Griffiths, R. I., Schloter, M., Bahn, M., Bardgett, R. D., 2021. Climate change alters temporal dynamics of alpine soil microbial functioning and biogeochemical cycling via earlier snowmelt. The ISME Journal, vol. 15, pp. 2264–2275. https://doi.org/10.1038/s41396-021-00922-0

Carlson, T. N., Ripley, D. A., 1997. On the relation between NDVI, fractional vegetation cover, and leaf area index. Remote Sensing of Environment, vol. 62, no. 3, pp. 241–252. https://doi.org/10.1016/S0034-4257(97)00104-1

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proceedings

of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.48550/arXiv.1603.02754

Coates, T. W., Alam, M., Flesch, T. K., Hernandez-Ramirez, G., 2021. Field testing two flux footprint models. Atmospheric Measurement Techniques, vol. 14, pp. 7147–7152. https://doi.org/10.5194/amt-14-7147-2021

D'Amico, M. E., Pintaldi, E., Sapino, E., Colombo, N., Quaglino, E., Stanchi, S., Navillod, E., Rocco, R., Freppaz, M., 2020. Soil types of Aosta Valley (NW-Italy). Journal of Maps, vol. 16, no. 2, pp. 755–765. https://doi.org/10.1080/17445647.2020.1821803

Feigenwinter, C., Bernhofer, C., Eichelmann, U., Heinesch, B., Hertel, M., Janous, D., Kolle, O., Lagergren, F., Lindroth, A., Minerbi, S., Moderow, U., Mölders, M., Montagnani, L., Queck, R., Rebmann, C., Vestin, P., Yernaux, M., Zeri, M., Ziegler, W., Aubinet, M., 2008. Comparison of horizontal and vertical advective CO_2 fluxes at three forest sites. Agricultural and Forest Meteorology, vol. 148, pp. 12–24. https://doi.org/10.1016/j.agrformet.2007.08.013

Filippa, G., Cremonese, E., Galvagno, M., Bayle, A., Choler, P., Bassignana, M., Piccot, A., Poggio, L., Oddi, L., Gascoin, S., Costafreda-Aumedes, S., Argenti, G., Dibari, C., 2022. On the distribution and productivity of mountain grasslands in the Gran Paradiso National Park, NW Italy: A remote sensing approach. International Journal of Applied Earth Observation and Geoinformation, vol. 108, art. 102718. https://doi.org/10.1016/j.jag.2022.102718

Foken, Th., Wichura, B., 1996. Tools for quality assessment of surface-based flux measurements. Agricultural and Forest Meteorology, vol. 78, pp. 83–105

Foken, T., Leclerc, M.Y., 2004. Methods and limitations in validation of footprint models. Agricultural and Forest Meteorology, vol. 127, pp. 223–234. https://doi.org/10.1016/j.agrformet.2004.07.015

Fugazza, D., Manara, V., Senese, A., Diolaiuti, G., Maugeri, M., 2021. Snow cover variability in the Greater Alpine Region in the MODIS era (2000–2019). Remote Sensing, vol. 13, no. 15, art. 2945. https://doi.org/10.3390/rs13152945

Galvagno, M., Wohlfahrt, G., Cremonese, E., Rossini, M., Colombo, R., Filippa, G., Julitta, T., Manca, G., Siniscalco, C., Morra di Cella, U., Migliavacca, M., 2013. Phenology and carbon dioxide source/sink strength of a subalpine grassland in response to an exceptionally short snow season. Environmental Research Letters, vol. 8, art. 025008. https://doi.org/10.1088/1748-9326/8/2/025008

Gao, B.-C., 1996. NDWI: A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space. Remote Sensing of Environment, vol. 58, pp. 257–266

Gómez-Ortiz, C., Monteil, G., Basu, S., e Scholze, M., 2025. A CO2–É14CO2 inversion setup for estimating European fossil CO2 emissions. Atmospheric Chemistry and Physics, vol. 25, pp. 397–424. https://doi.org/10.5194/acp-25-397-2025

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., e Gao, X., Ferreira, L. G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment, vol. 83, pp. 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Jones, L. A., Kimball, J. S., Reichle, R. H., Madani, N., Glassy, J., Ardizzone, J. V., Colliander, A., Cleverly, J., Desai, A. R., Eamus, D., Euskirchen, E. S., Hutley, L., Macfarlane, C., e Scot, R. L., 2017. The SMAP Level 4 Carbon Product for Monitoring Ecosystem Land–Atmosphere CO2 Exchange. IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 11, pp. 6517. https://doi.org/10.1109/TGRS.2017.2729343

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., Gans, F., Goll, D. S., Haverd, V., Köhler, P., Ichii, K., Jain, A. K., Liu, J., Lombardozzi, D., Nabel, J. E. M. S., Nelson, J. A., O'Sullivan, M., Pallandt, M., Papale, D., Peters, W., Pongratz, J., Rödenbeck, C., Sitch, S., Tramontana, G., Walker, A., Weber, U., e Reichstein, M., 2020. Scaling carbon fluxes from eddy covariance sites to the globe: synthesis and evaluation of the FLUXCOM approach. Biogeosciences, vol. 17, pp. 1343–1365. https://doi.org/10.5194/bg-17-1343-2020

Jung, M., Nelson, J., Migliavacca, M., El-Madany, T., Papale, D., Reichstein, M., Walther, S., e Wutzler, T., 2024. Technical note: Flagging inconsistencies in flux tower data. Biogeosciences, vol. 21, pp. 1827–1846. https://doi.org/10.5194/bg-21-1827-2024

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Tareth, A., Barr, A., Stoy,

P., e Wohlfahrt, G., 2010. Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: Critical issues and global evaluation. Global Change Biology, vol. 16, no. 1, pp. 187–208. https://doi.org/10.1111/j.1365-2486.2009.02041.x

Lawrence, D.M.; Fisher, R.A.; Koven, C.D.; Oleson, K.W.; Swenson, S.C.; Bonan, G.; et al. 2019. The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. Journal of Advances in Modeling Earth Systems, 11, 4245–4287. https://doi.org/10.1029/2018MS001583

Luyssaert, S., Schulze, E.-D., Börner, A., Knohl, A., Hessenmöller, D., Law, B. E., Ciais, P., e Grace, J., 2008. Old-growth forests as global carbon sinks. Nature, vol. 455, 11 September 2008. https://doi.org/10.1038/nature07276

Migliavacca, M., Grassi, G., Bastos, A., Ceccherini, G., Ciais, P., Janssens-Maenhout, G., Lugato, E., Mahecha, M. D., Novick, K. A., Peñuelas, J., Pilli, R., Reichstein, M., Avitabile, V., Beck, P. S. A., Barredo, J. I., Forzieri, G., Herold, M., Korosuo, A., Mansuy, N., Mubareka, S., Orth, R., Rougieux, P., e Cescatti, A., 2025. Securing the forest carbon sink for the European Union's climate ambition. Nature, published online 30 July 2025. https://doi.org/10.1038/s41586-025-08967-3

Nelson, J. A., e Walther, S., et al., 2024. X-BASE: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, FLUXCOM-X. Biogeosciences, 21, 5079–5115. https://doi.org/10.5194/bg-21-5079-2024

Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., Phillips, O.

L., Shvidenko, A., Lewis, S. L., Canadell, J. G., Ciais, P., Jackson, R. B., Pacala, S. W., McGuire, A. D., Piao, S., Rautiainen, A., Sitch, S., & Hayes, D., 2011. A Large and Persistent Carbon Sink in the World's Forests. Science, 333(6045), 988–993. https://doi.org/10.1126/science.1201609

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., Vesala, T., & Yakir, D., 2006. Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation. Biogeosciences, 3, 571–583. https://doi.org/10.5194/bg-3-571-2006

Pastorello, G., Trotta, C., Canfora, E., et al., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data, 7, 225. https://doi.org/10.1038/s41597-020-0534-3

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.-M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., Valentini, R., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Global Change Biology, 11, 1424–1439. https://doi.org/10.1111/j.1365-2486.2005.001002.x

Schader, L., Song, W., Kempker, R., Benkeser, D., 2024. Don't Let Your Analysis Go to Seed: On the Impact of Random Seed on Machine Learning-based Causal

Inference. Epidemiology, 35(6), 764–778. https://doi.org/10.1097/EDE.0000000000001782

Sitch, S.; Smith, B.; Prentice, I.C.; Arneth, A.; Bondeau, A.; et al. 2003. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. Global Change Biology, 9, 161–185. https://doi.org/10.1046/j.1365-2486.2003.00569.x

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., Papale, D., 2016. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. Biogeosciences, 13, 4291–4313. https://doi.org/10.5194/bg-13-4291-2016

Tramontana, G., Migliavacca, M., Jung, M., Reichstein, M., Keenan, T. F., Camps-Valls, G., Ogee, J., Verrelst, J., Papale, D., 2020. Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. Global Change Biology, 26, 3321–3340. https://doi.org/10.1111/gcb.15203

Van der Tol, C., Verhoef, W., Timmermans, J., Verhoef, A., Su, Z., 2009. An integrated model of soil-canopy spectral radiances, photosynthesis, fluorescence, temperature and energy balance. Biogeosciences, vol. 6. https://doi.org/10.5194/bg-6-3109-2009

Vos, G., van Eijk, L., Sarnyai, Z., Rahimi Azghadi, M., 2025. Stabilizing machine learning for reproducible and explainable results: A novel validation approach to

subject-specific insights. Computer Methods and Programs in Biomedicine, 269, 108899. https://doi.org/10.1016/j.cmpb.2025.108899

Vuichard, N., Papale, D., 2015. Filling the gaps in meteorological continuous data measured at FLUXNET sites with ERA-Interim reanalysis. Earth System Science Data, 7, 157–171. https://doi.org/10.5194/essd-7-157-2015

Wei, D. Qi, Y., Ma, Y., Wang, X., Ma, W., Gao, T., Huang, L., Zhao, H., Zhang, J., & Wang, X. (2021). Plant uptake of CO2 outpaces losses from permafrost and plant respiration on the Tibetan Plateau. Proceedings of the National Academy of Sciences of the United States of America, 118(33), e2015283118. https://doi.org/10.1073/pnas.2015283118

Wilson, J. D. (2015). Computing the flux footprint. Boundary-Layer Meteorology, 156(1), 1–14. https://doi.org/10.1007/s10546-015-0017-9

Yu, G.-R., Wen, X.-F., Sun, X.-M., Tanner, B. D., Lee, X., & Chen, J.-Y. (2006). Overview of ChinaFLUX and evaluation of its eddy covariance measurement. Agricultural and Forest Meteorology, 137(1-2), 125–137. https://doi.org/10.1016/j.agrformet.2006.02.011

Chapter 4

Conclusions

In this thesis, I developed and applied a set of machine learning–based methodologies aimed at improving the estimation of key vegetation structural parameters and the quantification of carbon fluxes in alpine ecosystems. These methods were conceived to bridge the gap between theoretical modeling and operational applications in Earth observation, by integrating remote sensing data, radiative transfer simulations, meteorological information, and in situ eddy covariance measurements. The work was structured into two main research lines, both addressing the need for accurate and spatially consistent information to better understand ecosystem functioning and carbon cycle dynamics under changing environmental conditions.

The first research line focused on the development and validation of a data-driven algorithm for retrieving the FC and the LAI from simulated spectral data, in the context of the forthcoming SBG-TIR mission, jointly developed by the ASI and NASA/JPL. The algorithm was trained using synthetic data generated with the SCOPE radiative transfer model, allowing a controlled simulation of the spectral response of vegetation canopies and soil backgrounds. Results demonstrated that models trained on such physically consistent datasets were able to retrieve FC and LAI on previously unseen synthetic validation sets. When FC and LAI values were mapped in [0,1] domain, the models achieved RMSE% values of 4.6% for FC and 5.2% for LAI. These results are in agreement with finding reported in previous studies (e.g., Weiss et al., 2016; Garcia-Haro et al., 2018).

When applied to real data, model performance decreases. Using observed FC and LAI values, the retrieval yielded an RMSE of 0.19 for FC and 1.02 for LAI. These results are consistent with those obtained on the same validation dataset using the SNAP Biophysical Processor (Weiss et al., 2020). This confirms that the performance degradation primarily reflects the increased complexity and noise of real-world data rather than limitations of the modelling framework itself. At the same time it confirms

the reliability and transferability of machine learning approaches when applied to spectrally complex and spatially heterogeneous environments. These findings highlight the potential of combining physical modeling and data-driven learning to derive vegetation parameters for next-generation satellite missions, contributing to operational monitoring of biosphere–atmosphere interactions.

The second research line addressed the quantification of carbon fluxes in alpine ecosystems, using a local optimization of the well-established FLUXCOM framework. The analysis was realized over the Alpine site of Torgnon, in the Aosta Valley, a high-altitude grassland ecosystem characterized by extreme climatic conditions and long snow-cover periods, which are rarely represented in global datasets. The study assessed the capacity of localized models to reproduce GPP in such an environment, comparing their performance against the global FLUXCOM model trained on more than 300 eddy covariance sites worldwide. Results showed that models trained exclusively on alpine or mountainous sites outperformed global models, even when limited predictor variables were available. This indicates that regionally optimized approaches are particularly valuable for improving the representation of under-sampled ecosystems, such as those at high elevations, and for refining carbon budget estimates at the regional scale.

The results of this work provide valuable insights into the use of machine learning in environmental modeling. They demonstrate how data-driven approaches, when properly constrained by physical knowledge and high-quality reference data, can complement traditional modeling techniques and enhance our capacity to monitor ecological processes. In particular, the methodologies developed here contribute to a better understanding of vegetation structure and carbon dynamics in mountain ecosystems, where environmental gradients and extreme conditions make conventional parameterizations less reliable.

Beyond their immediate applications, the approaches proposed in this thesis open the way for further developments in Earth system observation and modeling. Future research could extend these methods by incorporating multi-source satellite data, better accounting for snow-related processes and by integrating temporal dynamics to track

ecosystem responses across seasons and years. Moreover, extending local optimization strategies to other mountain regions could enhance global products by accounting for ecological variability and climatic extremes. Ultimately, these methodologies contribute to advancing the capacity of remote sensing and machine learning to provide reliable, scalable, and physically meaningful estimates of ecosystem processes, supporting both scientific understanding and informed environmental decision-making.

La borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4 – Componente 1 – Riforma 4.1 Riforma dei Dottorati – Inv. 4.1
Borse PNRR patrimonio Culturale –*CUP H41J22000170002*