ESA Contract Report

ESA Contract (4000144712/24/I-DT-bgh)

Contract Report to the European Space Agency

D1- Towards enhanced fire fuel estimation with satellite-derived predictive models

Authors: Siham El Garroussi and Joe McNorton
Contract officer: Claudia Vitolo

December 2024

Series: ECMWF ESA Contract Report Series

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/publications/

Contact: library@ecmwf.int

Executive summary

This report presents a technical note on developing predictive models that link satellite observations, such as vegetation optical depth and solar-induced fluorescence, to key fuel variables, specifically fuel load and fuel moisture content, which are crucial for wildfire forecasting.

The extreme gradient boosting model, a scalable and efficient decision tree-based machine learning algorithm, was implemented as a non-linear observation operator to integrate satellite-derived observations into an offline land data assimilation framework. Data preparation ensured temporal alignment with satellite observations and involved preprocessing steps to improve physical consistency. Feature importance analysis quantified global predictor contributions, while Shapley additive explanations analysis offered detailed insights into predictor impact and directionality at a granular level. Both analyses confirmed the model's reliance on physically meaningful relationships, consistent with known vegetation-water-energy dynamics.

The model performed well in capturing spatial and seasonal patterns, particularly in regions with clear phenological cycles, such as crops and savannas. Some challenges persist in areas with dense canopies or sparse vegetation, where signal saturation or soil-vegetation decoupling can reduce prediction accuracy. Filtering for orography, snow cover, and steep slopes is included as a preprocessing step in the land data assimilation framework.

The resulting predictive models, along with a sample of data, are provided as open-source code through this GitHub repository https://github.com/selgarroussi/fuelity.

Contents

1 Introduction

Major ecosystems of the world—boreal forests, shrublands, grasslands, and savannas—experience recurrent fires driven by natural causes or human activities (Di Giuseppe et al., 2016, 2021; Hantson et al., 2022). While fire weather components, such as temperature, show an increasing trend from the early 1980s to the present (Burton et al., 2024; El Garroussi et al., 2024), burnt area trends have declined over the same period (Burton et al., 2024). This offset between worsening fire weather and reduced burnt area reflects the influence of human activities, such as agricultural expansion, land-use changes, and fire suppression, in modulating fire outcomes. These contrasting trends have generated significant interest in understanding fire behaviour in relation to fuel characteristics, availability, and distribution (Carmona-Moreno et al., 2005; Hood et al., 2022). When large fires occur, fire managers require fuel maps based on consistent data and mapping methods. Unfortunately, such maps are often unavailable. Moreover, fuel characteristics and distribution represent a missing component in atmospheric monitoring systems that incorporate biomass burning emissions and their impact on air quality (Fleming et al., 2009).

Recently, the European Centre for Medium-Range Weather Forecasts (ECMWF) developed a diagnostic fuel framework that combines land surface modelling, meteorological variables, and satellite observations (McNorton and Di Giuseppe, 2024). This framework provides daily updates on vegetation characteristics globally at a high spatial resolution of 9 kilometres. The characteristics of the fuel extend beyond load and moisture, with further attribution to the foliage and woody components of both live and dead vegetation. However, the predictive quality of these data could be enhanced by assimilating Earth Observation (EO) observations into the system.

The synergistic assimilation of multiple satellite-derived variables has been recognised as a promising approach to enhancing the representation of vegetation in land surface models. The Fuelity project seeks to integrate diverse satellite observations, Vegetation Optical Depth (VOD) and Solar-Induced Fluorescence (SIF), to improve the estimation of key fuel variables, namely fuel load and fuel moisture content. Data assimilation (DA) is the process of producing an optimal estimate of the Earth's system state by integrating observations with model background information while accounting for both observation and model background errors (Courtier et al., 1994). A central component of DA is the observation operator, which translates model variables into observational space. This enables direct comparison and integration of satellite observations with Earth System Models (ESMs).

The choice of observation operator significantly impacts the quality of assimilation and encompasses different approaches, including physical forward models, empirical observation operators, and data-driven emulators. Physical forward models rely on well-established physical principles to simulate the measurement process, such as radiative transfer models that capture the interaction of radiation with vegetation canopies and soils. While these models offer high fidelity and consistency with underlying physical processes, they are computationally intensive, making them challenging to implement for global-scale or high-resolution applications. Empirical observation operators, on the other hand, utilise statistical relationships derived from historical data, such as cumulative distribution function (CDF) matching, to link model states with observations. These approaches are computationally efficient and straightforward to apply; however, they may oversimplify the complex, non-linear relationships inherent in some datasets, leading to potential inaccuracies in dynamic or heterogeneous processes. Finally, data-driven emulators, including machine learning (ML) techniques, have emerged as powerful alternatives, offering the advantage of not requiring a physical description of processes. This is particularly valuable for vegetation dynamics, where the underlying interactions are often highly complex and challenging to model explicitly.

These ML emulators can approximate complex observation processes by training on large datasets and

can handle diverse types of observations. However, their effectiveness depends on the quality and diversity of the training dataset, as well as the information content of the model predictors. Insufficient or biased training data can undermine model accuracy, limiting the effectiveness of assimilation efforts. In addition, training ML models at high spatial resolution and global scale requires substantial computational resources, especially for dense neural networks. In this context, decision-based methods, such as extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016; Calvet et al., December 2024), are explored here, which offer an effective balance between computational efficiency and robustness.

In this report, we examine the feasibility of using the XGBoost machine learning algorithm to develop an observation operator that emulates satellite-derived datasets. Specifically, we focus on VOD metrics at C-band, L-band, and X-band (C-VOD, L-VOD, X-VOD), along with SIF. VOD provides insights into vegetation water content, biomass, and structural characteristics, with its sensitivity varying across microwave frequencies. At the same time, SIF, retrieved from hyperspectral satellite sensors, serves as a proxy for photosynthetic activity and vegetation stress. Section 2 outlines how the fuel characteristic model generates the relevant fuel variables and describes the available observations considered pertinent for informing these fuel variables. Section 4 explains the modelling approach used to assess the contribution of each observation to the relevant fuel variables. Finally, Section 5 highlights the robustness of the predictive models developed.

2 Data

2.1 Fuel data

Effective wildfire management and prevention strategies rely on accurate forecasts of fire occurrence and spread. Fuel load and fuel moisture content contain essential variables for these forecasts and can be estimated using ESMs (McNorton et al., 2024). These variables can be modelled either in real-time to support operational wildfire forecasting or in historical mode to provide insights into long-term fuel trends driven by climate variability.

We use a global fuel characteristic model that provides frequent (sub-daily), high-resolution (∼9km) data in both historical and real-time operational modes. The model outputs eight key variables—four related to fuel load and four related to fuel moisture—which are summarised in Figure 1. Several of these variables are informed by a combination of model outputs and satellite observations.

The fuel load is initialised using data from the ESA-CCI Biomass product version 3 (Santoro et al., 2021). However, this dataset alone is insufficient for informing fuel variables relevant to wildfire forecasting due to its infrequent updates and lack of real-time availability. Additionally, the dataset does not provide information on the partitioning of fuel among different pools, which we address using land surface modelling techniques.

For frequent fuel load updates, we use modelled net ecosystem exchange (NEE) generated by the ECLand surface model (Boussetta et al., 2021). ECLand is the surface component of the operational Integrated Forecasting System (IFS) used by ECMWF. NEE is modelled using the A-gs scheme, which is driven by meteorological and land surface variables (e.g., temperature, soil moisture) as well as plant-specific physiological traits (Boussetta et al., 2013).

Since model NEE is susceptible to biases, we incorporate an online bias correction scheme from the Copernicus Atmosphere Monitoring Service, which provides operational atmospheric CO_2 forecasts (Agustí-Panareda et al., 2019). This scheme applies scaling factors to the NEE values based on a cli-

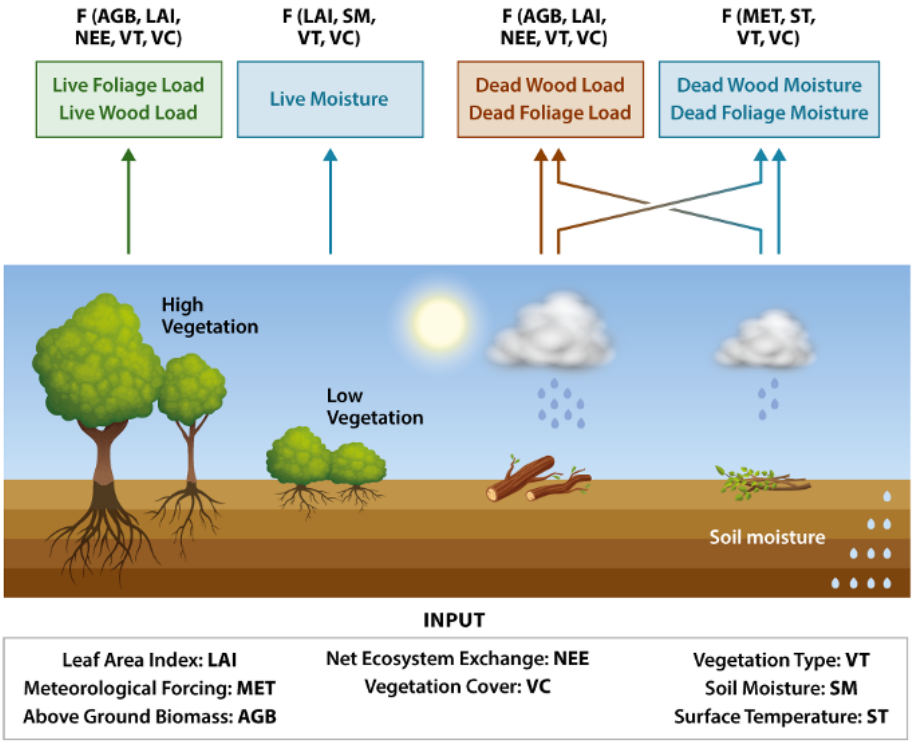

Figure 1: A schematic of the fuel characteristic model, showing the required input data and the dependencies for the output variables.

matology of optimised fluxes obtained from atmospheric inversions (Chevallier et al., 2010; Agustí-Panareda et al., 2016).

The modelled fuel load is further subdivided into foliage and wood mass, accounting for both live and dead components, as these are critical for fire modelling methodologies (McNorton and Di Giuseppe, 2024). Dead foliage and wood is partitioned based on the 20 different PFTs using fixed ratios as shown by table X1. The calculation of dead fuel moisture as 1- , 10- , 100- and 1000- h fuels is also determined by PFT as shown by table 2. The allocation of live mass to foliage varied temporally and was determined using vegetation-specific leaf mass per unit area (LMA) and observed LAI following the methodology described by Harper et al. (2016). Our definition of live foliage includes leaf mass only, while dead foliage includes small branches for the purpose of the fuel moisture model.

The temporal variability in live wood mass, which is a representation of seasonal short-lived wood mass (e.g. twigs), was determined as being proportional to LAI using the power law relationship proposed by Enquist et al. (1998) and vegetation-specific coefficients provided by Eq.(4) in Harper et al. (2018). The remaining component of live wood, which is a representation of long-lived wood mass which remains regardless of LAI state (e.g. trunk), was assumed to be constant over time, preserving the live-to-dead fuel load ratio (L:D) when averaged over the entire time series.

The majority of total fuel load was allocated to long-lived wood mass with relatively little seasonal variability in live wood mass, particularly in low latitude regions. The LAI dependence resulted in an increased fraction of live mass during the growing season, with relative increases largest in foliage but absolute increases largest in wood.

Seasonal variability in the four fuel categories (live foliage, live wood, dead foliage, and dead wood) is

Vegetation Type	Live Percentage (%)	Dead Foliage Percentage (%)	Leaf Mass per Unit Area	Allometric Coefficient†	Reference
Crops	85	100	0.1370	0.005	-
Short grass	85	100	0.0495	0.005	Fan et al. (2007), Peichl et al. (2011), Perez et al. (2000)
Evergreen needleleaf trees	65	45	0.2263	0.65	Pan et al. (2011)
Deciduous needleleaf trees	55	60	0.1006	0.80	Pan et al. (2011)
Deciduous broadleaf trees	60	90	0.0823	0.78	Michaelian et al. (2010), Pan et al. (2011), Peichl et al. (2011)
Evergreen broadleaf trees	85	10	0.1039	0.845	Pan et al. (2011)
Mixed Crops	60	100	0.1370	0.005	Fan et al. (2007), Guo et al. (2005), Li et al. (2020)
Desert	50	50	0.1370	0.005	-
Tundra	75	50	0.0495	0.005	-
Irrigated crops	85	100	0.1370	0.005	-
Semidesert	65	50	0.1370	0.005	-
Bogs and marshes	65	50	0.1370	0.005	-
Evergreen shrubs	70	50	0.1515	0.13	Anderson et al. (2015), Baeza et al. (2006)
Deciduous shrubs	65	90	0.0709	0.13	Li et al. (2020), Taylor et al. (2021)
Broadleaf Savannah	80	90	0.1370	0.13	-
Interrupted forest	60	90	0.1039	0.78	-

Table 1: ECLand vegetation types and parameter values used to derive categorised fuel load. Values estimated where references are not given. †relates LAI to wood mass (See equation 4, Harper et al. (2018)).

modulated by either climatological or time-varying LAI values.

Fuel state, in addition to quantity, is a key component for modelling fire danger and risk. This includes, amongst other things, fuel arrangement, structure, and moisture content. Here, we focus on model-derived moisture content for both live (LFMC) and dead (DFMC) fuels, with DFMC further subdivided between foliage and wood components.

To estimate LFMC for both high and low vegetation types, we use a semi-empirical model based on key variables identified through a random forest approach (McNorton and Di Giuseppe, 2024). The semi-empirical method was chosen because it imposes physical constraints on the modelled LFMC, necessary to account for modelled LFMC ranges that were not adequately covered by the training data. The model is trained on the Globe-LFMC dataset, which is based on in-situ destructive sampling measurements (Yebra et al., 2019). Model optimisation was performed using the Trust Region Reflective algorithm, under the assumption that LFMC varies with soil moisture and LAI following an asymptotic regression. This approach accounts for theoretical maximum LFMC values per vegetation type and incorporates vegetation-specific factors such as drought resistance, growth phase, and dormancy. In the operational

Vegetation Type	1-h Fuel (% of dead foliage)	10-h Fuel (% of dead foliage)	100-h Fuel (% of dead wood)	1000-h Fuel (% of dead wood)
Crops	50	50	100	0
Short grass	100	0	100	0
Evergreen needleleaf trees	53	47	30	70
Deciduous needleleaf trees	53	47	30	70
Deciduous broadleaf trees	53	47	30	70
Evergreen broadleaf trees	53	47	30	70
Mixed Crops	50	50	100	0
Desert	65	35	100	0
Tundra	65	35	100	0
Irrigated crops	50	50	100	0
Semidesert	65	35	100	0
Bogs and marshes	65	35	100	0
Evergreen shrubs	65	35	100	0
Deciduous shrubs	50	50	100	0
Broadleaf Savannah	50	50	50	50
Interrupted forest	50	50	50	50

Table 2: Allocation of vegetation-specific fuel into the 4 classes defined by the National Fire Danger Rating System.

IFS, this process returns two LFMC values: one for low vegetation, such as grasses, and one for high vegetation, such as trees.

Daily DFMC is estimated by generalising the 'Nelson model', a physically based approach originally developed for the moisture content of 10-hour fuels (Nelson, 2000). This model uses hourly inputs of air temperature, humidity, radiation, and precipitation to simulate heat and moisture transfer both within the fuel and at its surface. Following the extension by Carlson et al. (2007), we expanded the model to include 1-hour, 10-hour, 100-hour, and 1000-hour fuels. These four classifications are based on the response time of the fuel to changes in moisture and are associated with the diameter of the fuel.

Given that the fuel model differentiates between dead foliage and wood, we further categorised the DFMC fuel types. For foliage, including small branches, we applied the moisture content derived from 1 hour and 10 hour fuels, while the 100 hour and 1000 hour fuel moisture was used for wood. The weighting of these moisture values varies according to the specific type of vegetation.

2.2 Observations

In this work, we consider four distinct satellite observations- L-, C-, and X-band Vegetation Optical Depth (VOD), along with Solar-Induced Fluorescence (SIF) -to assimilate fuel variables (see Table 3, Figure 2). While these observations form the foundation of our current analysis, additional observations could and should be included in the future to enhance fuel representation.

Satellite observations that exhibit significant sensitivity to vegetation parameters can be classified into two primary categories. The first category includes optical sensors measuring at visible frequencies of the

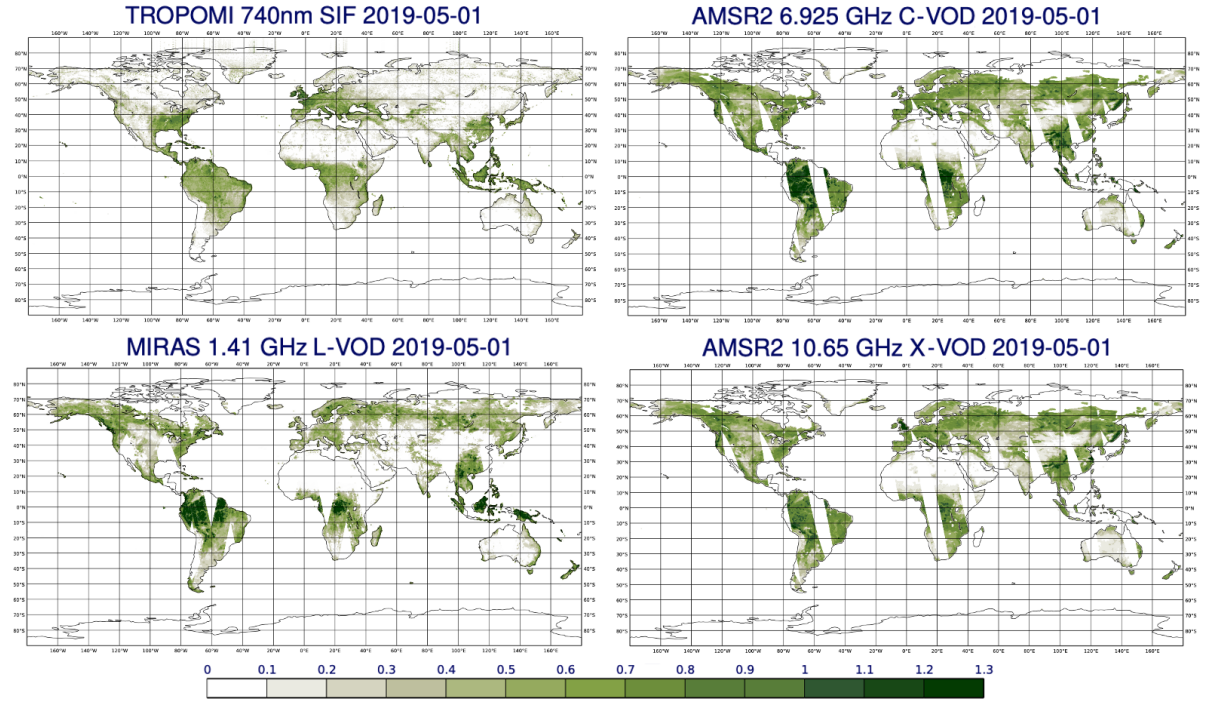

Figure 2: Daily coverage maps of TROPOMI Solar-Induced Fluorescence (SIF) at 740 nm and AMSR2 Vegetation Optical Depth (VOD) observations on 1st May 2019. The maps show SIF at the top left, AMSR2 VOD at 6.9 GHz (C-band) at the top right, AMSR2 VOD at 1.41 GHz (L-band) at the bottom left, and AMSR2 VOD at 10.65 GHz (X-band) at the bottom right.

electromagnetic spectrum, such as the Ocean and Land Cover Imager (OLCI) on the Sentinel-3 series of satellites. These instruments provide direct measurements of vegetation parameters; however, the availability of usable observations is limited to daylight hours and is affected by cloud cover. As a result, it can take many days to achieve reliable global coverage. Observations from these sensors contribute to the monthly climatologies currently used in the numerical weather prediction (NWP) system.

The second category comprises passive microwave (MW) sensors, such as the Advanced Microwave Scanning Radiometer-2 (AMSR-2) onboard the Global Change Observation Mission for Water (GCOM-W) satellite. The emitted microwave radiation is sensitive to the water content of the vegetation, thus providing indirect information about the vegetation parameters themselves. At the lowest MW frequencies, there is little to no sensitivity to atmospheric conditions and no sensitivity to the time of day, allowing these instruments to provide near-global coverage every 12 hours. This is the primary reason for assimilating MW observations to generate a daily analysis of fuel variables.

Observation operators that transform land surface model parameters into simulated top-of-atmosphere MW radiances are not yet mature or accurate enough to consider directly assimilating level 1 MW radiances for analysing vegetation parameters (or other land surface model variables). This limitation arises from the heterogeneity of land-surface conditions and the complex relationship between surface characteristics and MW emissivity and penetration depth. Therefore, we utilise a derived level 2 vegetation product instead of the raw level 1 radiances. This level 2 product is the VOD, which is retrieved from the measured MW radiances and quantifies how much the vegetation attenuates the MW radiation emitted by the surface. VOD is sensitive to the water content, type, and density of above-ground vegetation, with its sensitivity varying across different frequencies.

At higher frequencies, such as the C-band (4–8 GHz) and X-band (8–12 GHz), VOD is particularly responsive to vegetation cover and shows strong correlations with the LAI. In contrast, at lower frequencies like the L-band (1–2 GHz), VOD is more sensitive to above-ground biomass (AGB) and correlates more effectively with it (Rodríguez-Fernández et al., 2018). Furthermore, L-VOD does not saturate as quickly as C- and X-band measurements, enabling deeper canopy penetration (Ulaby et al., 1981). The specific sensitivity of VOD also varies with vegetation structure. L-VOD is more responsive to coarse woody components, such as trunks, stems, and branches. Meanwhile, C- and X-VOD are more sensitive to finer vegetation elements, such as leaves and thin stems (Guglielmetti et al., 2007).

Table 3: Characteristics of the global satellite observations considered in the study.

Observations	**L-VOD**	**X-VOD**	**C-VOD**	**SIF**
Satellite	SMOS	GCOM-W1	GCOM-W1	Sentinel-5 Precursor
Sensor	MIRAS	AMSR-2	AMSR-2	TROPOMI
Frequency band and value (GHz)	L (1.41)	X (10.65)	C (6.925)	O2-A (403-395)
Swath (km)	1000	1445	1445	2600
Spatial resolution (km)	25	42×24	62×35	3.5×5.5
Temporal resolution	daily	daily	daily	8-daily
Reference	Al Bitar et al. (2017)	Moesinger et al. (2020)	Moesinger et al. (2020)	Guanter et al. (2021)

In this study, retrieved VOD from three different frequencies and two different instruments is used. L-band VOD (1.41 GHz) is obtained from the Microwave Imaging Radiometer using Aperture Synthesis (MIRAS) instrument onboard the ESA Soil Moisture Ocean Salinity (SMOS) satellite. In addition, C-band VOD (6.925 GHz) and X-band VOD (10.65 GHz) are used from the AMSR-2 instrument onboard the GCOM-W satellite. Both these satellites are in sun-synchronous polar orbits and therefore provide near-complete global coverage every 24 hours.

The solar-induced fluorescence (SIF), also considered in this study, was derived from the TROPOMI (TROPOspheric Monitoring Instrument) onboard the Copernicus Sentinel-5P satellite. TROPOSIF provides global estimates of SIF at 740 nm, obtained from the 743–758 nm near-infrared window. SIF captures the fluorescence emitted by chlorophyll during photosynthesis, offering real-time insights into plant stress and photosynthetic activity.

The canopy SIF signal is primarily influenced by three biophysical processes: (1) the fluorescence yield of leaves, closely linked to photosynthetic activity; (2) the absorbed photosynthetically active radiation (APAR), which is determined by solar radiation and vegetation structure; and (3) the escape probability of emitted photons, dependent on multiple scattering processes within the canopy. High SIF values are typically associated with healthy vegetation, while reduced SIF may indicate water stress, correlating with increased fuel dryness.

3 Methodology

We carried out a preliminary assessment to evaluate the added value of different satellite datasets in capturing and enhancing fuel variables. This assessment aimed to identify which satellite observations provide the most meaningful contributions to monitoring and predicting fuel conditions. In the second phase of our analysis, we focused on the development of an ML-based observation operator that translates fuel variables into the corresponding satellite observations. This ML forward model is intended to bridge the gap between model-derived fuel estimates and satellite measurements, enabling a more seamless integration of observational data into the offline land data assimilation system (Rosnay et al., 2013).

For both phases of the analysis, we employed the XGBoost model (see Figure 3) to investigate the

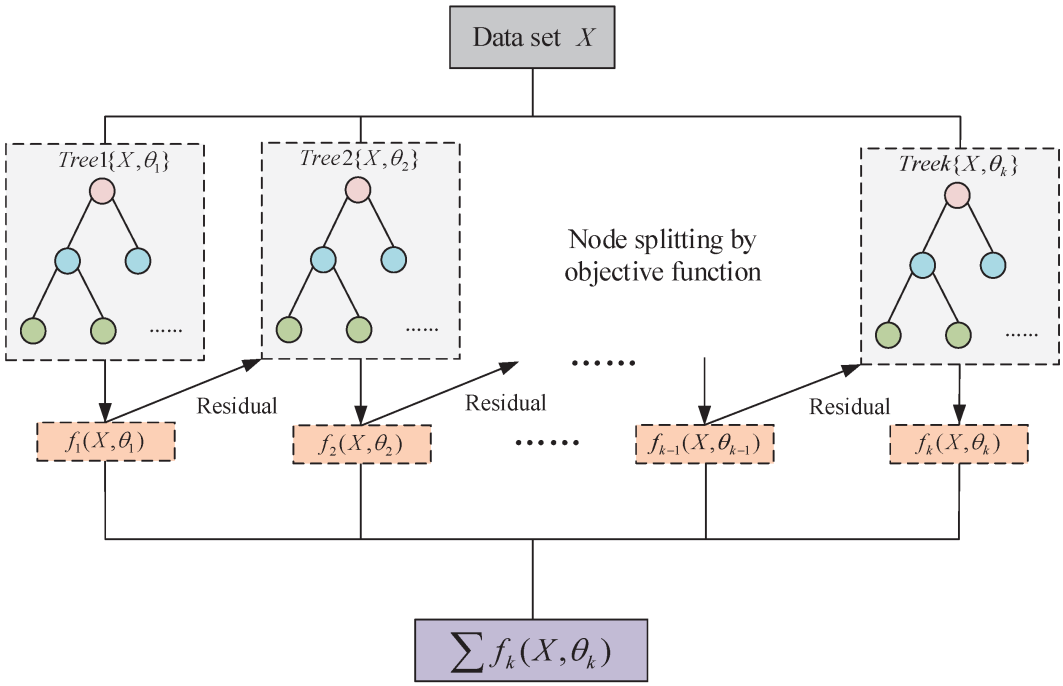

Figure 3: XGBoost framework, adopted from Guo et al. (2020)

relationships between the four satellite observations (see Table 3) and fuel variables (see Figure 1).

XGBoost, short for eXtreme Gradient Boosting (Chen and Guestrin, 2016), is a high-performance machine learning algorithm based on gradient boosting decision trees. It is widely recognised for its speed, scalability, and accuracy, making it a go-to tool for structured data analysis. XGBoost iteratively constructs an ensemble of decision trees, where each subsequent tree corrects the errors made by the previous ones, gradually minimising the prediction loss. It also incorporates advanced regularisation techniques: L_1, which penalises the absolute values of model coefficients to encourage sparsity, and L_2, which penalises the squared values of coefficients to shrink them and reduce overfitting. This enables robust predictions even in the presence of noisy or sparse data. Furthermore, XGBoost offers flexibility for feature selection, making it ideal for complex, multi-variable systems.

3.1 Initial assessment of the potential of satellite data to inform fuel variables

This involves identifying the distinct signal that each fuel variable contributes to each type of satellite observation. The methodology begins by constructing an XGBoost model that simulates satellite observations using eight daily fuel variables, aggregated to the spatial resolution of the observations, for the entire year of 2020. In this tree-based model, each time a feature is selected to split the decision tree, the improvement it brings in reducing model error, as measured by a loss function, indicates the feature's importance. By considering the contribution across all trees and splits, we determine the overall feature importance. This metric reflects the sensitivity of the satellite observation to perturbations in each of the fuel variables, within the variability and limits of the training data.

We also considered SHAP (SHapley Additive exPlanations) analysis, a rigorous and scientifically grounded method for interpreting machine learning models, with fuel properties as predictors and satellite observations as targets. SHAP analysis quantifies the contribution of each fuel-related feature to the model's

predictions of satellite-observed outcomes by leveraging Shapley values from cooperative game theory. These values allocate the total "payoff" (in this case, the model's prediction of satellite observations) among the fuel predictors based on their average marginal contributions across all possible subsets. This approach ensures that both the individual effects of fuel properties and their interactions are accounted for in a fair and consistent manner. The additive property of SHAP values ensures that their sum corresponds exactly to the difference between the model's prediction for a given instance and a baseline value (typically the mean satellite observation), offering a detailed and interpretable breakdown of how specific fuel properties influence the satellite-derived targets.

Because different vegetation types are expected to exhibit distinct responses to VOD and SIF observations, we train separate models for each vegetation type to better capture the feature importance. This approach provides deeper insights into which fuel variables in different ecosystems are more strongly influenced by various satellite observations. In an operational context, it is anticipated that the model and assimilation will involve all features and observations simultaneously within one framework.

Several considerations accompany this approach. A low ranking in feature importance analysis does not necessarily imply that a variable cannot be updated by an observation. For example, if a variable is well-correlated with another variable that significantly improves model performance, it may still be updated through observation assimilation. The challenge arises when multiple variables require updating but are constrained by a limited number of observations. This situation can potentially lead to an underconstrained problem with several viable solutions when assimilating observational data.

XGBoost models were trained on the entire 2020 dataset, with fuel characteristics sampled at the spatiotemporal resolution of the observations. For VOD, this involved daily sampling at approximately 25 km resolution, while for SIF, observations were sampled every 8 days at the same resolution. Multiple models were trained for various vegetation types, both high and low. We considered tropical, temperate, and boreal forests for high vegetation types, based on land cover classifications used in the operational IFS. For low vegetation, we considered only grassland or cropland. To distinguish whether a grid cell is considered high or low vegetation dominant, we examined land cover fractions from the IFS.

3.2 Development of an observation operator for linking fuel variables to satellite observations

The development of an observation operator to link fuel variables with satellite observations began with the need to harmonise the datasets. As the data originated from different sources and resolutions, regridding was essential to create a common spatial grid. Various interpolation methods were tested to ensure consistency (see Figure 4), and the most suitable method was selected to minimise errors, optimise the handling of missing data, and maintain accuracy across the datasets.

The initial XGBoost model, built solely on fuel variables, performed poorly due to strong intercorrelations and insufficient signal to predict the four satellite observations. To improve performance, additional relevant variables were incorporated based on insights from a literature review and exploratory tools like correlation analysis (see Figure 5).

In selecting predictors, we prioritised minimising intercorrelations and ensuring model parsimony. While LFMC could be a key predictor, it was not directly included in the observation operator. Instead, we relied on LAI and soil moisture levels 1 and 4, as these variables were identified as the most significant predictors. Soil moisture at levels 2 and 3 showed weaker correlations and was therefore excluded to limit redundancy and ensure computational efficiency. Fuel load components were aggregated due to their higher intercorrelation. In operational mode, the aggregated fuel load will be disentangled using

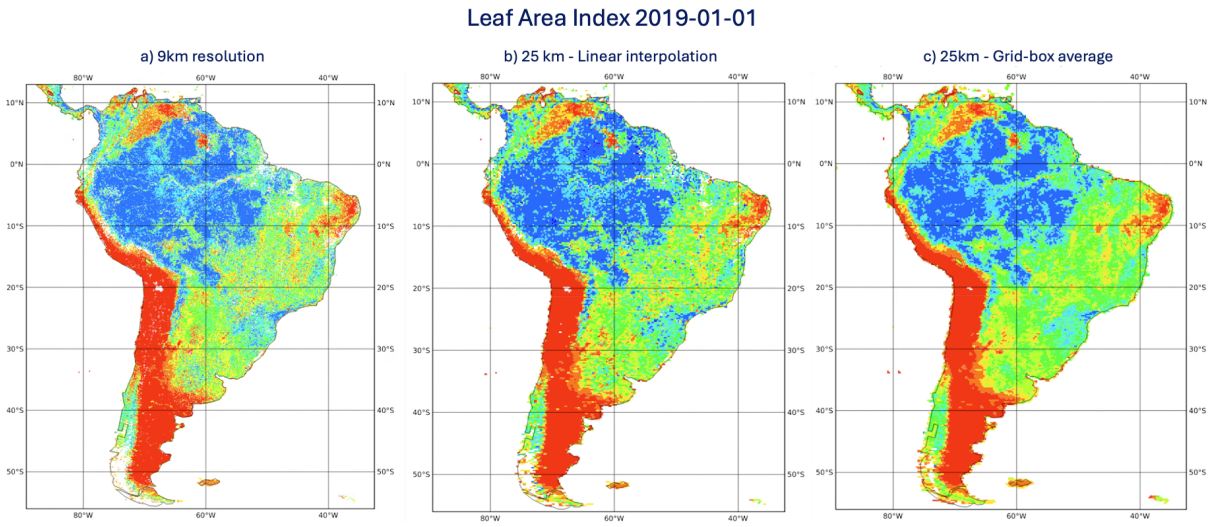

Figure 4: Comparison of spatial data interpolation methods for LAI on 1st January 2019: (a) 9 km resolution preserves high spatial detail and precision; (b) 25 km linear interpolation ensures smooth transitions between grid points at coarser resolution; (c) 25 km gridbox averaging provides a representative mean value within each grid box, summarising spatial variability.

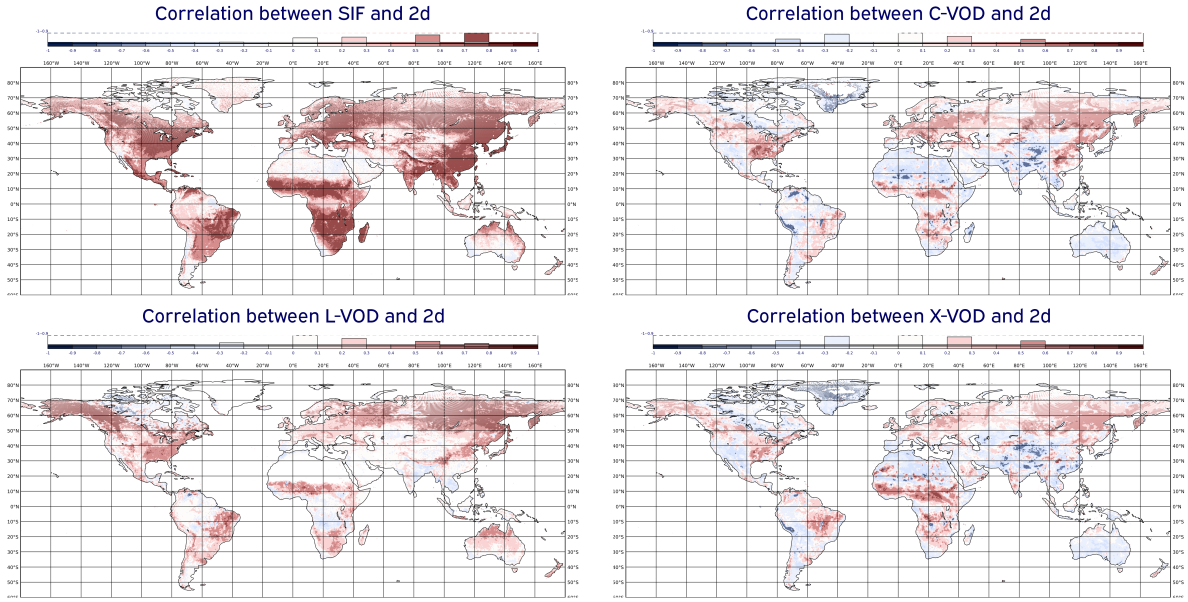

Figure 5: Spatial map of temporal correlation between 2-metre dewpoint temperature (2d) and satellite observations over the period 2019–2021. The top-left panel shows the correlation with SIF, the top-right panel with C-VOD, the bottom-left panel with L-VOD, and the bottom-right panel with X-VOD. Red shading indicates positive correlations, while blue shading indicates negative correlations.

LAI as a reference.

Data preparation relied on the Anemoi datasets (https://anemoi-datasets.readthedocs.io/en/latest/) framework to streamline the handling and integration of data. This ensured that all relevant predictors were efficiently managed and consistently formatted for modelling.

To improve the forward model's performance, we trained the observation operator using LAI Copernicus data, which provides more dynamic vegetation information compared to the monthly climatological LAI from IFS. However, as the LAI Copernicus dataset has an 8-day temporal resolution, we regridded other predictors, such as fuel variables, to match this temporal frequency.

A common temporal period across all datasets was identified, with 2019–2020 used for training and 2021 for testing. This allowed us to validate the forward model's generalisability and robustness using unseen data.

For validation, we employed cross-validation techniques across the training dataset to assess the model's reliability and minimise the risk of overfitting. Cross-validation involves dividing the training data into multiple subsets, iteratively training the model on a portion of the data while validating it on the remaining subsets. Root Mean Square Error (RMSE) and the Predictive Coefficient (R^2) were used as performance metrics. RMSE quantified the model's ability to predict target variables by measuring the average magnitude of the prediction errors, while the predictive coefficient evaluated how well the model captured the variance in the test data.

4 Results

4.1 Feature importance

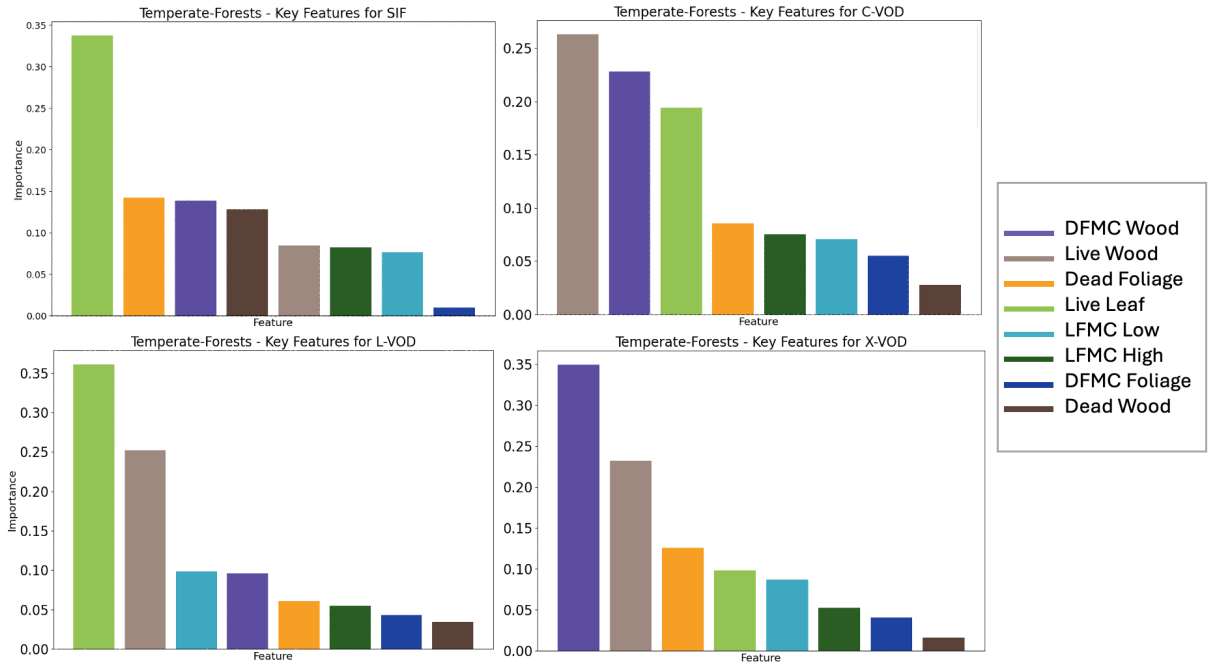

Figure 6: Feature importance of modelled observations in Temperate Forest regions based on global daily data for 2020.

We have evaluated model feature importance for four observation types: C-, L- and X- Band VOD and SIF. Results are illustrated in Figure 6, summarised in Table 4, and detailed in Appendix Figures A1-5. These results indicate which features can be informed by each observation type, considering the previously mentioned factors.

Table 4: Summary of feature importance for the eight fuel variables across all models when predicting observations for different biomes. Average importance values greater than 0.15 are categorised as highly informative (green), values between 0.05 and 0.15 as partially informative (yellow), and values less than 0.05 as not informative (red).

	C-Band VOD						L-Band VOD						X-Band VOD						SIF					
	Boreal F.	Temp. F.	Trop. F.	Savannah	Cropland	Grassland	Boreal F.	Temp. F.	Trop. F.	Savannah	Cropland	Grassland	Boreal F.	Temp. F.	Trop. F.	Savannah	Cropland	Grassland	Boreal F.	Temp. F.	Trop. F.	Savannah	Cropland	Grassland
Live Leaf Load																								
Live Wood Load																								
Dead Foliage Load																								
Dead Wood Load																								
LFMC High																								
LFMC Low																								
DFMC Foliage																								
DFMC Wood																								

Before exploring these details, it is important to address the correlations between variables. The strongest correlation exists between dead foliage and dead wood load for a given vegetation type. This is due to their perfect temporal correlation, with dead fuel varying in magnitude with LAI and a fixed ratio allocating that load between foliage and wood. The differences observed in the feature importance are the result of interference caused by model dead foliage and wood from the non-targeted vegetation type, whether high or low. For example, the total dead foliage over a grid cell will vary depending on both vegetation types present within the grid cell, high and low, and it is this value considered by the model.

Similarly, live wood and leaf load depend on LAI, though they are based on assumptions about static wood load, resulting in a high but not perfect correlation. Live Fuel Moisture Content (LFMC) is influenced by LAI and soil moisture, but because the model calculation is vegetation-type dependent, the correlation is high but not perfect. DFMC for foliage and wood are a function of the same variables but are not perfectly correlated due to slight variations in the calculations. These correlations can make disentangling signals challenging, although insights into the differences between fuel load and moisture can still be achieved.

Our results reveal both similarities and differences in the average signal provided by variables in modelled observations. For all vegetation types, except savannah regions, live leaf load is the most informative variable for SIF, reflecting its role as a proxy for photosynthetic activity. In contrast, for L-band VOD, feature importance varies with vegetation type. For instance, in savannah regions, it informs LFMC, dead foliage load, and live wood load, while in cropland areas, it is a key observation for inferring live leaf load.

Figure 6 highlights how, on average, a combination of observation types can inform all fuel model variables in temperate forests. Live wood load is influenced by C-, L-, and X-band VOD, while live leaf load is informed by C- and L-band VOD and SIF. Dead wood and foliage load, being well-correlated for a given vegetation type, can be informed by X-band VOD and SIF. For moisture variables, LFMC is informed by L-band VOD, whereas DFMC is informed by C- and X-band VOD.

Table 4 outlines the potential information content of each observation for all fuel variables by biome. Interestingly, while SIF primarily serves as a proxy for photosynthetic activity, it appears to be partially influenced by dead fuel load and moisture, likely due to correlated information.

Feature importance reflects the average gain from each variable, but it is important to note that the significance of each variable can vary with individual model predictions of VOD or SIF. Consequently,

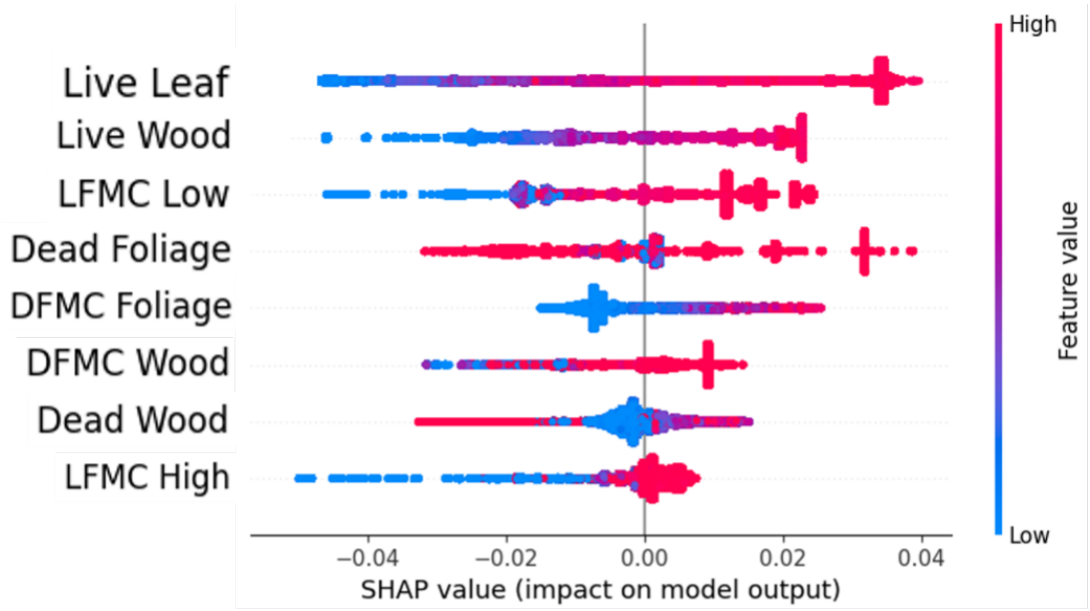

Figure 7: A beeswarm SHAP plot showing the SHAP values for each variable for every prediction made globally in 2020 for C-Band VOD in tropical forests.

a low average feature importance does not imply that a variable lacks informative value for specific observations. This variability is further illustrated by calculating SHAP values for each prediction (see Example Figure 7). SHAP values offer a feature-specific metric to evaluate the contribution to each final prediction, revealing that under certain conditions, the contribution of a specific feature can noticeably vary.

Our analysis reveals that while fuel load is generally significant for most observations, anomalies observed in VOD may often be linked to variations in moisture content, if wood mass remains relatively constant throughout the year. Therefore, the potential information content from observations for variables with low average importance might be more substantial than indicated by average feature importance alone (e.g., the red squares in Table 4).

For example, as previously mentioned, dead wood and foliage load are well correlated for a given vegetation type, and therefore information on one can inform the other. Similarly, Low and High LFMC are functions of soil moisture and LAI, specific to each vegetation type, and thus can inform each other to some extent. An example of these correlations in the relative contributions of each variable to model predictions is illustrated in Figure 8. Consequently, all observation types can offer insight into modelled fuel variables and should be considered for inclusion in any assimilation system. However, it is important to recognise that specific increments provided by such a system may be under-constrained. This is due, in part, to the relatively large number of fuel variables (8) compared to the fewer observation types (4), and to the challenges posed by the correlations between these variables.

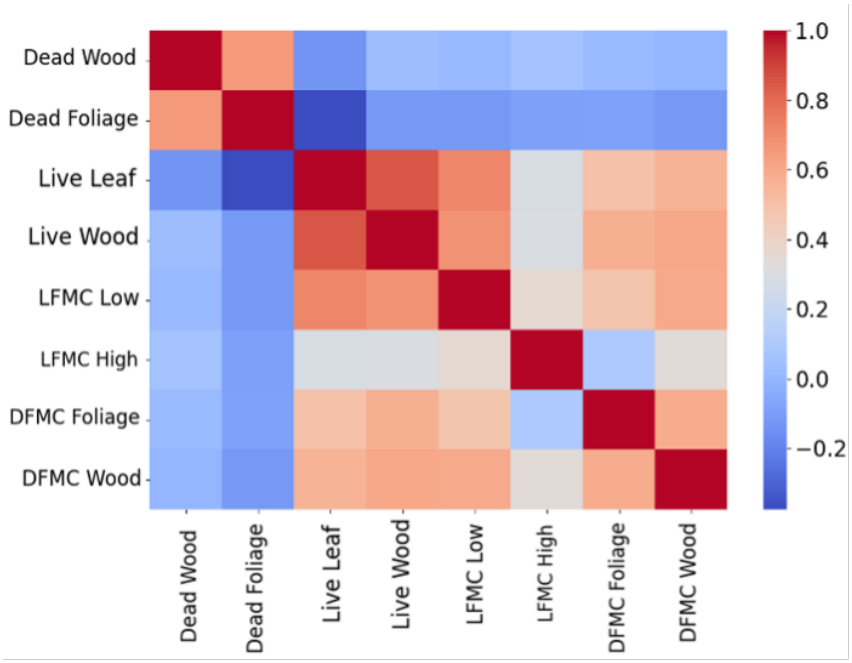

Figure 8: A correlation heatmap of SHAP values for each fuel variable globally in 2020 for C-Band VOD in tropical forests.

4.2 XGBoost forward model

4.2.1 Model optimisation

We trained the model using data from 2019–2020 with an 8-day temporal resolution, reflecting the frequency of SIF observations, and incorporated LAI from Copernicus Global Land Service (CGLS), which provides global 10-daily data. Testing was conducted on 2021 data, with the same model applied across SIF, C-VOD, X-VOD, and L-VOD observations. The spatial resolution is TCO1279 (∼9 km globally), and the temporal resolution is 8 days.

Different sets of predictors were tested to optimise the emulation and generalisation across satellite observations. Accumulating solar radiation to match the SIF temporal frequency slightly improved performance by better capturing energy flux dynamics; however, for operational simplicity, accumulation was restricted to a single day during inference. Adding latitude and longitude marginally enhanced generalisation, but their exclusion prompted the model to prioritise physical processes, such as relying on solar radiation for SIF, rather than learning spatial proximity effects. While incorporating vegetation type and cover modestly improved generalisation, the inclusion of orography introduced a dominant static influence, likely linked to its strong association with rainfall patterns, sunlight exposure, temperature, and cloud cover, all of which significantly affect vegetation health and structure. Attempts to include trigonometric representations of the Julian day (cosine and sine) led to degraded results, suggesting limited benefit for capturing seasonal variations within the tested framework. Fuel moisture content for dead vegetation added noise and degraded performance, reflecting its limited relevance to SIF dynamics. In contrast, summing fuel load improved predictions across all variables (SIF, C-VOD, L-VOD, and X-VOD).

The selected features (also referred to as predictors), summarised in Table 5, were chosen based on their

relevance in minimising the model's RMSE while ensuring physical consistency and interpretability. These predictors include key variables such as total fuel load, moisture content of dead fuel in wood and foliage, leaf area index (LAI), 2-metre dew point (2d), 2-metre temperature (2t), surface solar radiation downwards (ssrd), soil moisture (swvl1 for layer 1: 0–7 cm and swvl4 for layer 4: 100–289 cm), soil temperature (stl1) measured at the midpoint of layer 1 (3.5 cm), and vegetation-related parameters, including high and low vegetation types (tvh and tvl) and their respective cover fractions (cvh and cvl).

Table 5: Summary of XGBoost features.

Variable	Input Category	Frequency	Source	Reference
2m Temperature (2t)	Weather	Daily	ERA5-Land	Muñoz Sabater et al. (2021)
2m Dewpoint Temperature (2d)	Weather	Daily	ERA5-Land	Muñoz Sabater et al. (2021)
Surface Short-wave (solar) Radiation Downwards (ssrd)	Weather	Daily (acc.)	ERA5-Land	Muñoz Sabater et al. (2021)
Soil Moisture (swvl)	Weather/Fuel	Daily	ERA5-Land	Muñoz Sabater et al. (2021)
Soil Temperature (stl)	Weather	Daily	ERA5-Land	Muñoz Sabater et al. (2021)
Total Fuel Load	Fuel	Daily	Fuel Model	McNorton and Di Giuseppe (2024)
Dead Fuel Moisture Content	Fuel	Daily	Fuel Model	McNorton and Di Giuseppe (2024)
Leaf Area Index (LAI)	Fuel/Vegetation	10-daily	Satellite (Copernicus Global Land Service, LAI 300m v1.0)	Fuster et al. (2020)
Type of Vegetation	Vegetation	Fixed	ECLand	Boussetta et al. (2021)
Vegetation Cover	Vegetation	Fixed	ECLand	Boussetta et al. (2021)

Besides optimising the set of predictors, we also focused on fine-tuning the model parameters to avoid overfitting and enhance generalisation performance.

The convergence plots in Figure 9 compare the training and validation RMSE for the L-VOD model before (left) and after (right) applying regularisation. Before regularisation, the model converges rapidly for a number of trees equal to 250. However, the close alignment of the training and validation RMSE suggests limited control over overfitting (the RMSE validation curve, in red, being below the RMSE training curve, in green), as the model may be fitting noise rather than generalising well to unseen data.

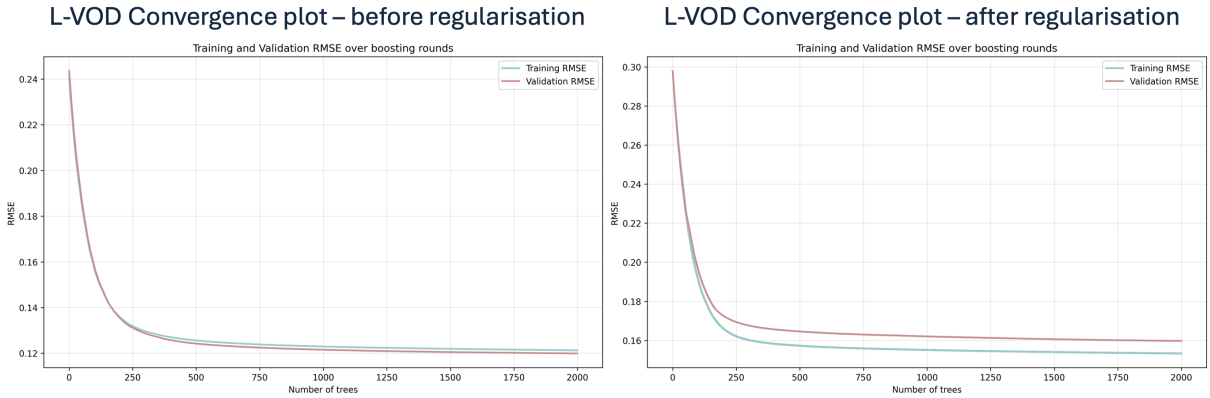

Figure 9: Comparison of L-VOD convergence plots before (left) and after (right) regularisation.

To address this, regularisation techniques were applied by incorporating constraints into the XGBoost model hyperparameters. Specifically, gamma = 2.0 was used to penalise overly complex splits, lambda = 5.0, and alpha = 2.0 to enforce L_2 and L_1 regularisation, respectively.

Additionally, setting the *minimum child weight* to 10 ensures that splits occur only when a sufficient amount of data contributes to the decision, preventing overfitting to small, less informative subsets. Subsampling, set to 60% (*subsample* = 0.6), randomly selects a portion of the training data for each boosting iteration, reducing variance and promoting model generalisation. Similarly, feature sampling (*colsample by tree* = 0.6) limits the number of predictors considered for each tree, introducing randomness and decreasing dependency on specific variables. The learning rate was set to 0.01, enabling a gradual and stable optimisation process across 2000 boosting rounds. The gamma parameter, set to 2.0, imposes a penalty on overly complex splits by requiring a minimum reduction in loss before a split is made. Regularisation terms such as L_2 regularisation (*lambda* = 5.0) and L_1 regularisation ($alpha$ = 2.0) act as constraints on the model weights, further discouraging overfitting by smoothing extreme values. After applying these regularisation techniques, the right plot highlights improved separation between training and validation RMSE. The validation curve stabilises at a slightly higher RMSE than the training curve, reflecting reduced overfitting and enhanced generalisation performance, while maintaining a consistent trade-off between bias and variance.

4.2.2 Model performance

The frequency plots in Figure 10 illustrate XGBoost performance for the four satellite observations (SIF, C-VOD, L-VOD, and X-VOD) on both training (left) and test (right) datasets. Each panel compares predicted values against observed values using density heatmaps, where denser regions are shown in darker shades. The diagonal blue line indicates the ideal 1:1 relationship between predictions and observations.

For SIF, the training dataset shows strong agreement between predicted and observed values, with a high R^2 and low error metrics. However, on the test data, the model slightly underestimates larger values, as indicated by the deviation below the 1:1 line at higher observed values.

For C-VOD, predictions closely match observations in both training and test datasets, maintaining a strong R^2 and low RMSE. The model generalises well across the data, with predictions evenly distributed around the diagonal.

In the L-VOD case, performance on the training set shows high alignment with observed values. On

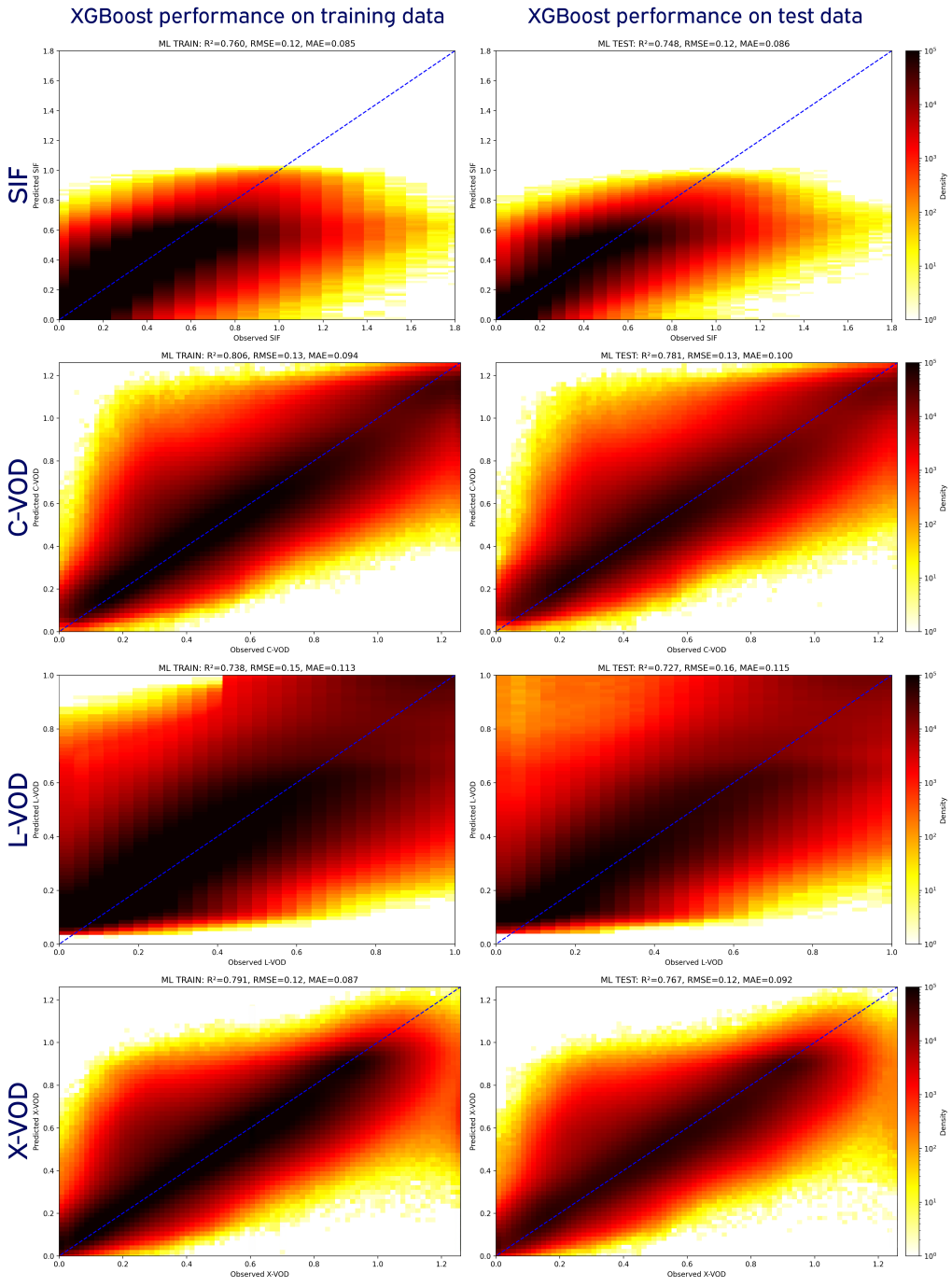

Figure 10: XGBoost performance for SIF, C-VOD, L-VOD, and X-VOD on training (left) and test (right) datasets. Density plots show predicted values versus observed values, with performance metrics (R^2, RMSE, and MAE) annotated for each case.

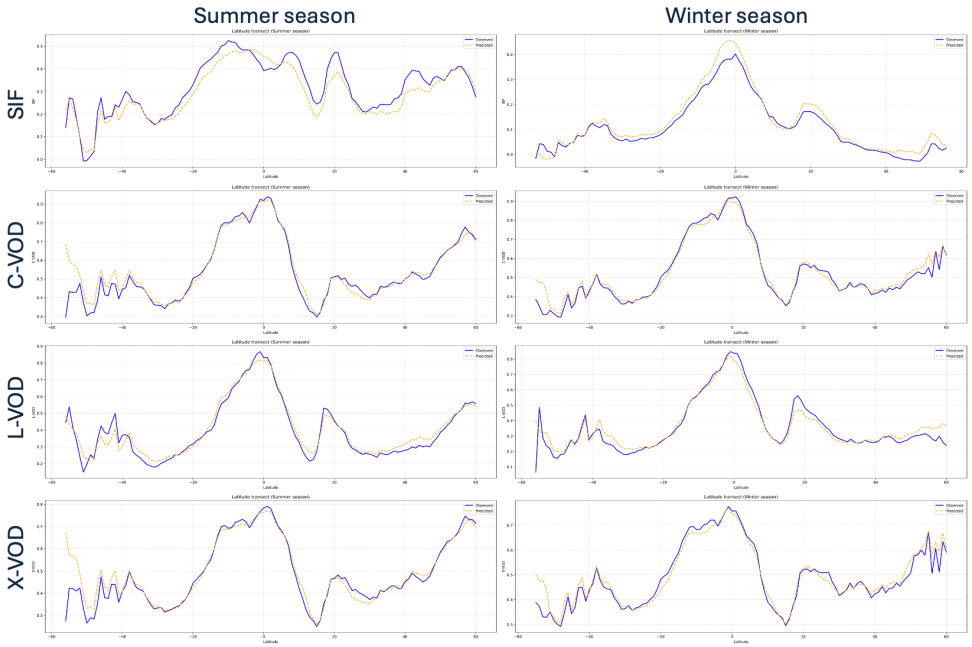

Figure 11: Latitude transects of observed (solid blue) and predicted (dashed orange) values for SIF, C-VOD, L-VOD, and X-VOD during the summer (left) and winter (right) seasons for the validation year 2021.

the test set, while the spread remains centred around the diagonal, there is a slight underestimation for larger observed values. For X-VOD, the model achieves consistent performance across training and test datasets, as shown by the tight clustering of predictions around the 1:1 line.

Across all variables, the heatmaps of the distribution of values, predicted and observed, confirm that the model effectively captures the overall patterns and seasonal dynamics in the training data (see Figure 11), with only minor deviations appearing in the test data, particularly for extreme values. This could imply a limitation in capturing extreme dynamics or outliers, which may have been smoothed by the model. This effect is likely due to the absence of physical filtering for orographic features, snow-covered areas, and steep slopes. Such filters will be applied in the preprocessing step of the DA framework to better handle these regions.

Figure 12 provides further spatial insight into the model's performance, focusing on L-VOD temporal RMSE (top panel) and anomaly correlation (bottom panel). The top panel demonstrates that the temporal RMSE remains low across most regions globally, with minimal error observed in areas of dense vegetation, such as the Amazon, Congo Basin, and Southeast Asia. These regions, where vegetation dynamics dominate, exhibit more predictable temporal patterns. Conversely, regions with higher RMSE values, highlighted in yellow, are sparsely distributed and coincide with areas of complex terrain, snow cover, or arid conditions where the model struggles to capture temporal variability. This aligns with the earlier limitations noted regarding the lack of physical filtering. The bottom panel highlights the anomaly correlation between the predicted and observed L-VOD, revealing stronger agreement (blue shades) across temperate and boreal regions, including North America, Europe, and parts of Asia. This suggests that the model effectively captures vegetation anomalies in these areas, where seasonal dynamics are well defined.

In contrast, weaker or even negative correlations (red shades) are observed in regions such as parts of the Amazon Basin, Africa, and central Australia. This could be attributed to several factors. First,

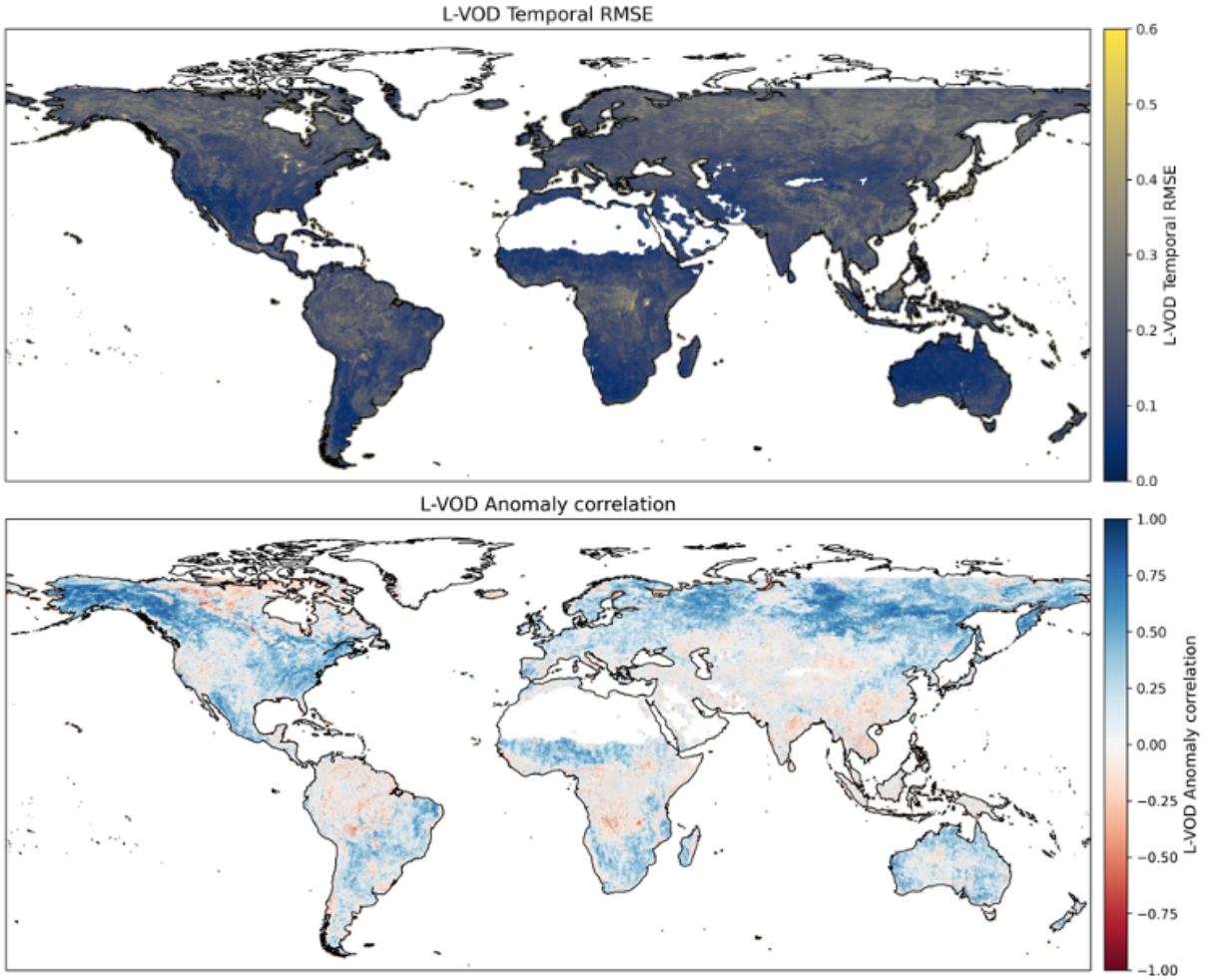

Figure 12: Global distribution of L-VOD temporal RMSE (top) and anomaly correlation (bottom) over the independent validation year of 2021. Similar assessments for other variables (SIF, C-VOD, and X-VOD) are provided in the appendix.

these regions are characterised by extreme environmental variability, including abrupt droughts, fires, and intense seasonal fluctuations. In the Amazon Basin and parts of Africa, dense vegetation cover, especially tropical rainforests, introduces strong signal attenuation and high moisture levels complicating microwave signal retrievals. This can cause saturation of the L-band signal, where increasing vegetation biomass no longer translates linearly into L-VOD values, limiting the model's ability to resolve such dense canopies. Second, soil moisture-vegetation decoupling is a significant issue in arid and semi-arid regions like central Australia and parts of Africa. In these areas, sparse vegetation and low biomass generate weak L-VOD signals, which are often confounded by surface soil moisture variability. L-VOD retrieval from SMOS relies on microwave penetration into the canopy, but when vegetation is minimal, the signal is dominated by soil moisture rather than vegetation dynamics. This can reduce the signal-to-noise ratio, making it challenging for the model to isolate vegetation contributions.

Third, land surface heterogeneity and complex topography can degrade model performance. In regions with steep slopes or highly variable land cover (e.g., Africa's savanna-forest transitions and mountainous terrain), SMOS-derived L-VOD retrievals may contain errors due to mixed signals and sub-pixel variability. The limited spatial resolution of SMOS further amplifies this effect, as signals from vegetation,

soil, and surface roughness are blended within a single observation.

Finally, non-linear vegetation responses to water stress are not always well represented in the data, particularly under extreme conditions such as droughts and fires. These processes can cause rapid changes in vegetation biomass, which are difficult for models to capture without additional predictors or physical constraints. In tropical regions like the Amazon, persistent cloud cover also hampers the quality of ancillary inputs (e.g., optical vegetation indices), adding uncertainty to the L-VOD predictions.

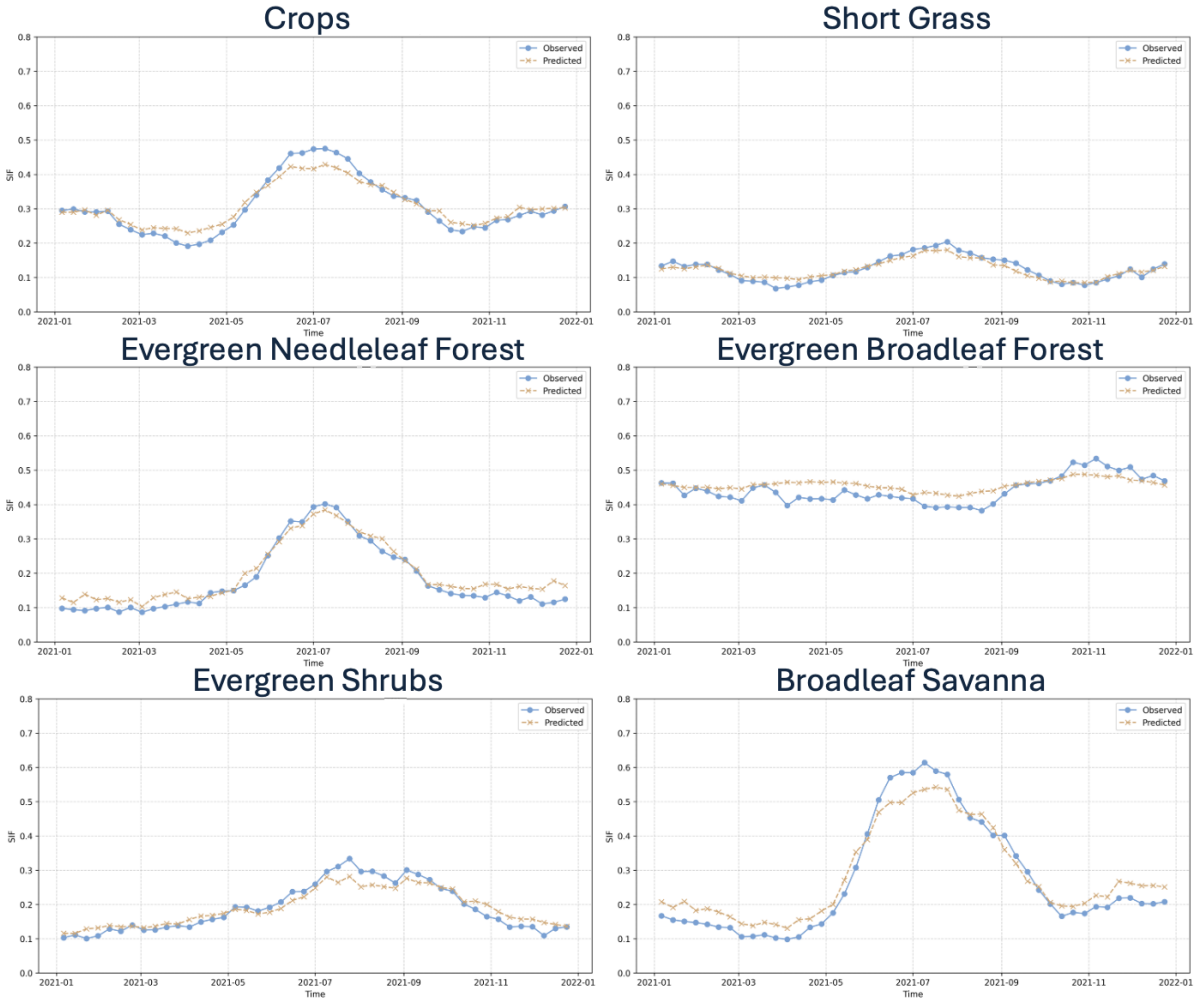

Figure 13: Assessment of XGBoost performance for SIF predictions across vegetation types, including crops, short grass, evergreen needleleaf forest, evergreen broadleaf forest, evergreen shrubs, and broadleaf savanna. Observed (solid blue) and predicted (dashed orange) SIF values are compared over time for the validation year 2021. Similar assessments for other variables (C-VOD, L-VOD, and X-VOD) are provided in the appendix.

The model's performance for SIF was evaluated across different vegetation types (see Figure 13), demonstrating overall good agreement with observed seasonal dynamics. For crops and broadleaf savanna, the model captures the seasonal peaks during the growing season, aligning with photosynthetic activity and biomass production. However, a slight underestimation at the peak suggests unmodelled stress factors, such as drought or agricultural management.

In short grass and evergreen shrubs, the model maintains reasonable agreement but underrepresents ob-

served variability, likely due to low SIF magnitudes and noise from sparse vegetation and soil contributions.

For evergreen needleleaf and broadleaf forests, the model underpredicts SIF, likely due to signal saturation in dense canopies for needleleaf forests and attenuation in moisture-rich environments for broadleaf forests, which limits its ability to capture seasonal peaks and subtle variations, respectively.

Overall, the model performs well in vegetation types with strong seasonal dynamics, such as crops and savanna, but faces challenges in regions with low variability, sparse cover, or dense canopies. Further improvements, including refined predictors and physical constraints, could enhance performance. Similar evaluations for C-VOD, L-VOD, and X-VOD are presented in the appendix.

Limiting the number of predictors is essential to ensure both computational efficiency and physical consistency in the observation operator within a DA framework. Since XGBoost is a non-linear decision tree-based model, computing the Jacobian relies on finite difference methods, which can become inefficient and less reliable when an excessive number of predictors is included. Redundant or irrelevant predictors may introduce noise, reduce sensitivity to key variables, and obscure the physical relationships that the model aims to capture.

4.2.3 Model insights

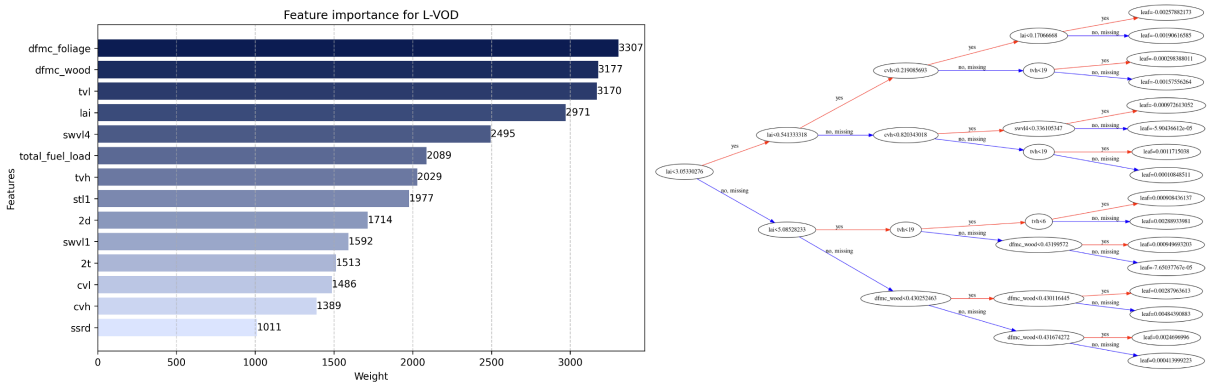

Figure 14: Feature importance (left) and a decision tree split example (right) for L-VOD predictions.

To gain insights into the model's behaviour and improve explainability, feature importance analysis was conducted for the four satellite observations. Figure 14 presents the results for L-VOD, where fuel-related variables, particularly dead fuel moisture content in foliage ($dfmc\ foliage$) and wood ($dfmc\ wood$), emerge as the dominant predictors. This dominance highlights the strong physical connection between fuel moisture and vegetation optical depth, as fuel variables reflect the water status in plant tissues and surface vegetation layers, which directly influence the L-band microwave signal captured by satellites. LAI, soil moisture at deeper layers (swvl4), and total fuel load also emerge as critical contributors. These variables collectively account for vegetation structure and water availability, highlighting the role of canopy density, biomass, and soil-vegetation interactions in modulating L-VOD signals. The decision tree example further illustrates how the model prioritises splits based on fuel moisture content and vegetation structure, with LAI and dead fuel moisture content in wood acting as key thresholds for decision-making.

While feature importance based on weight provides a global ranking of predictors and insights into their overall contributions to the model, it does not indicate how individual predictors influence specific

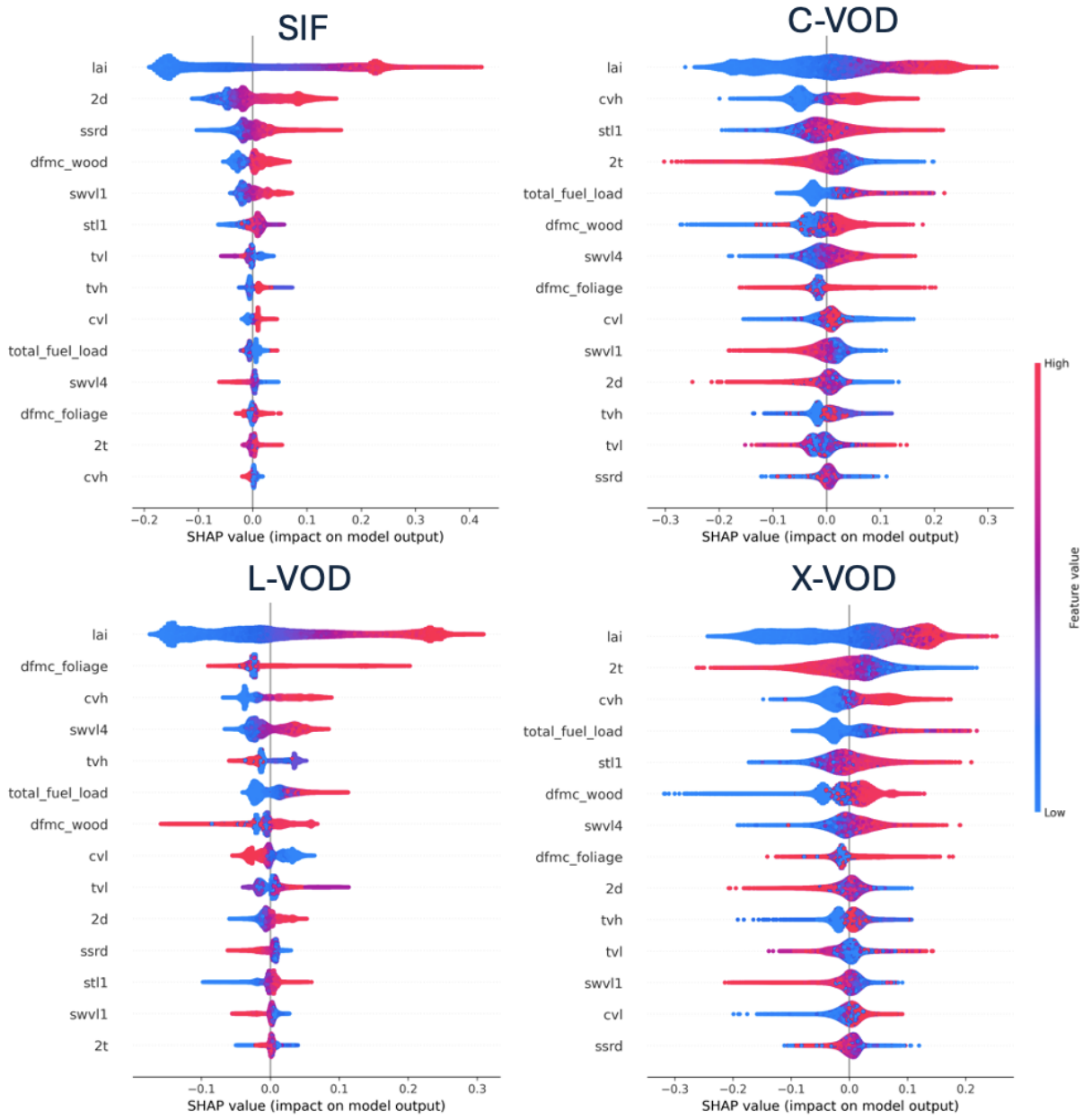

Figure 15: SHAP analysis for SIF, C-VOD, L-VOD, and X-VOD predictions. Each plot shows the SHAP values (x-axis) representing the impact of predictors (y-axis) on model output. Feature values are color-coded, with red indicating high values and blue indicating low values. The horizontal spread indicates the magnitude and direction of each predictor's influence on the model's predictions.

predictions or the direction of their impact. To address this limitation, SHAP analysis was conducted to further explore the model's behaviour (see Figure 15). SHAP values quantify the magnitude of each predictor's contribution and clarify whether a predictor increases or decreases the model's output under different conditions.

For SIF, LAI stands out as the most influential predictor. Higher LAI values (in red) correspond to higher SIF values, reflecting denser vegetation and more active photosynthesis. Other significant contributors include solar radiation (ssrd) and dead fuel moisture content in wood (*dfmc wood*), which affect vegeta-

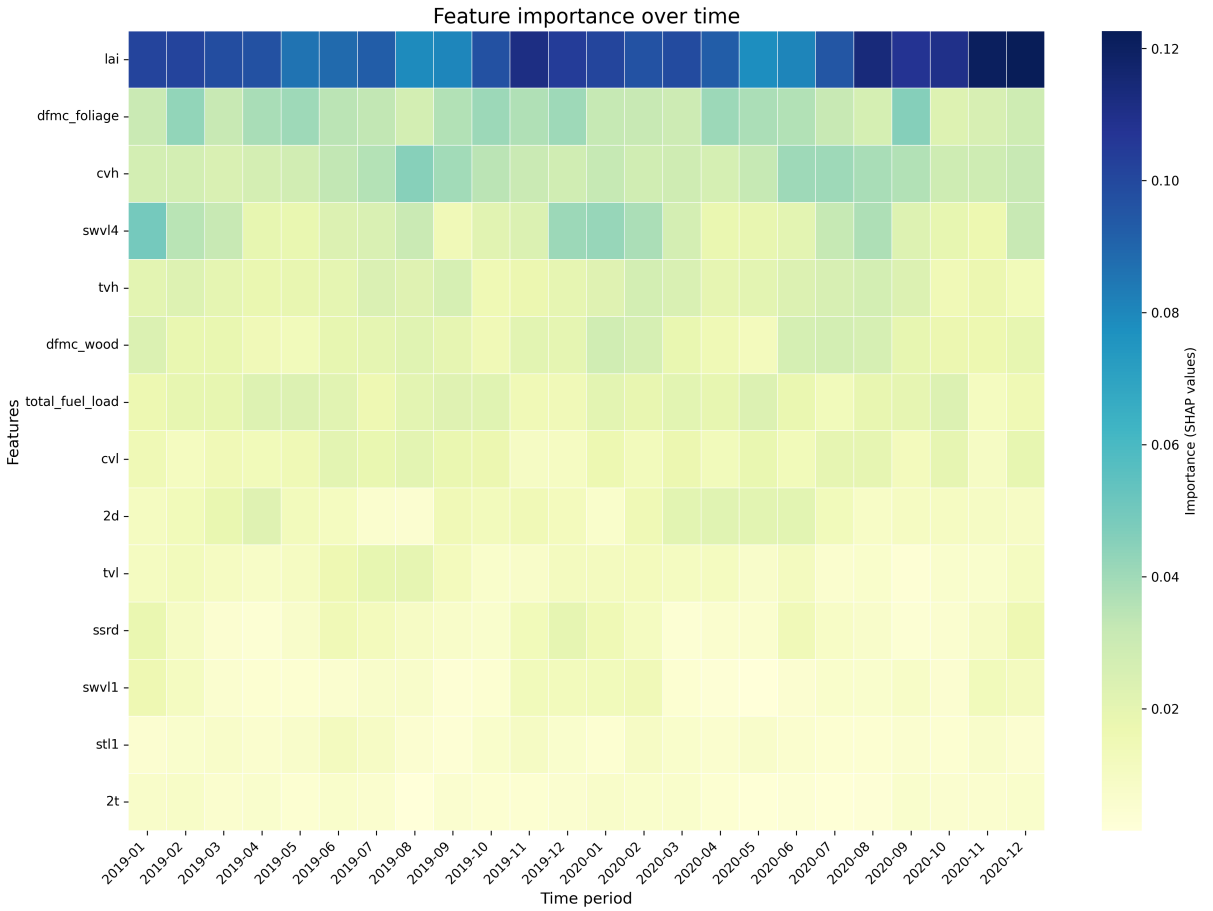

Figure 16: Feature importance over time derived using SHAP values for an XGBoost model built separately for each month to predict L-VOD. The analysis spans the training period from 2019 to 2020, using monthly satellite observations. The heatmap highlights seasonal patterns. Darker blue shades represent features with greater influence on model predictions. Similar assessments for other variables (C-VOD, X-VOD, and SIF) are provided in the appendix.

tion water status and stress. Notably, predictors such as swvl1 (top layer soil moisture) and stl1 (top layer soil temperature) emphasise the critical roles of both moisture and energy availability in modulating the SIF signal.

In the case of C-VOD, LAI and high vegetation cover (cvh) are the primary predictors. Strong contributions also arise from stl1 (surface temperature) and soil moisture variables like swvl4 (deep soil moisture), reflecting the interaction between vegetation, soil water availability, and energy balance. Physically, C-VOD is sensitive to vegetation water content and structure, particularly in moderate biomass regions where C-band signals do not easily saturate.

For X-VOD, similar trends are observed, with LAI, 2t (2-metre temperature), and cvh (high vegetation cover) as the dominant predictors. Vegetation structure, biomass, and surface energy balance determine X-band optical depth (see Figure 15), with additional contributions coming from fuel moisture variables and deep soil moisture (swvl4).

For L-VOD, LAI remains the most significant predictor, consistent with its role in representing canopy density and biomass. Fuel-related variables, such as dead fuel moisture content in foliage and wood,

also show substantial contributions, indicating their relevance in capturing vegetation water content and stress, especially in dry conditions. Soil moisture variables (swvl4 and swvl1) and total fuel load are also important, highlighting the connection between L-VOD and both canopy water content and soil moisture dynamics. L-VOD's sensitivity to deep biomass and water stress aligns well with the signals observed in L-band retrievals. The heatmap in Figure 16 illustrates how feature importance, derived from SHAP values, varies over time in predicting L-VOD, with clear seasonal patterns reflecting vegetation growth dynamics. The dominant influence of vegetation-related variables, such as LAI and Dead Fuel Moisture Content in foliage, indicates that the model relies heavily on live vegetation density and moisture content to make predictions. These features maintain high importance throughout the year, highlighting their critical role in capturing both growing and dormant season dynamics.

Seasonality is particularly evident, as the importance of soil moisture (swvl4) increases during the mid-year months, likely corresponding to peak vegetation growth in spring and summer when water availability plays a pivotal role in sustaining biomass. Conversely, during colder months, vegetation-related variables such as LAI remain influential, reflecting the model's sensitivity to structural vegetation characteristics even when active growth subsides.

Meteorological features and aggregated fuel load variables exhibit relatively low importance, suggesting that while these factors indirectly affect vegetation, they are less directly tied to L-VOD dynamics. This pattern underscores the model's focus on vegetation health and moisture as primary predictors, capturing the seasonal interplay between soil water availability, vegetation structure, and L-VOD response.

5 Summary and conclusion

The XGBoost model was developed as an observation operator for offline land DA of satellite-derived observations (SIF, C-VOD, L-VOD, and X-VOD) to improve the estimation of fuel variables. Data preparation ensured temporal alignment with observations and included preprocessing steps to enhance consistency. Key predictors, such as fuel moisture content, LAI, soil moisture, and vegetation cover, were selected based on their physical relevance.

Feature importance analysis highlighted fuel moisture and vegetation indices as the dominant drivers of L-VOD and X-VOD. SHAP analysis provided further insights into the predictors' impact, confirming consistency with known vegetation-water-energy relationships. The model effectively captured spatial and seasonal patterns, particularly in regions with clear phenological cycles, but showed limitations in dense canopies and areas with sparse vegetation, where signal saturation or soil-vegetation decoupling can occur. The DA framework will allow the application of physical filters, such as those for orography and snow cover, to better handle noisy data.

As an observation operator, the XGBoost model provides a flexible and computationally efficient tool for assimilating satellite data. Limiting the number of predictors within the XGBoost model is essential for ensuring both computational efficiency, especially when dealing with high-resolution global data, and physical consistency. Since XGBoost is a non-linear decision tree-based model, computing the Jacobian matrix relies on finite difference methods, which can become inefficient and less reliable when too many predictors are included. Redundant or irrelevant predictors may introduce noise, reduce sensitivity to key variables—specifically fuel variables—and obscure the physical relationships that the model aims to capture.

This analysis does not address how well-constrained the information is, given the presence of eight fuel variables, which are compressed to six fuel variables for optimal XGBoost performance, and only four

observation types. Given this constraint, it is anticipated that the correlations between fuel variables will provide adequate information to update them with reasonable accuracy within an assimilation framework.

To enhance this framework, incorporating a broader range of observations could offer varying sensitivities across model fuel variables. Such diversity would improve the assimilation system's capability to update fuel variables more comprehensively. This analysis highlights the potential value of constructing an assimilation system to inform fuel variables effectively.

A key addition to the system, again not explored here, is the potential for intermittent nudging of the model; for example, providing occasional global biomass estimates to prevent model drift. This would not have to be provided at the model timestep resolution but could be an annual estimate used to help constrain the model.

Data availability

The forward model presented in this work can be rebuilt using the Jupiter notebook workflow available on GitHub https://github.com/selgarroussi/fuelity.

Appendix

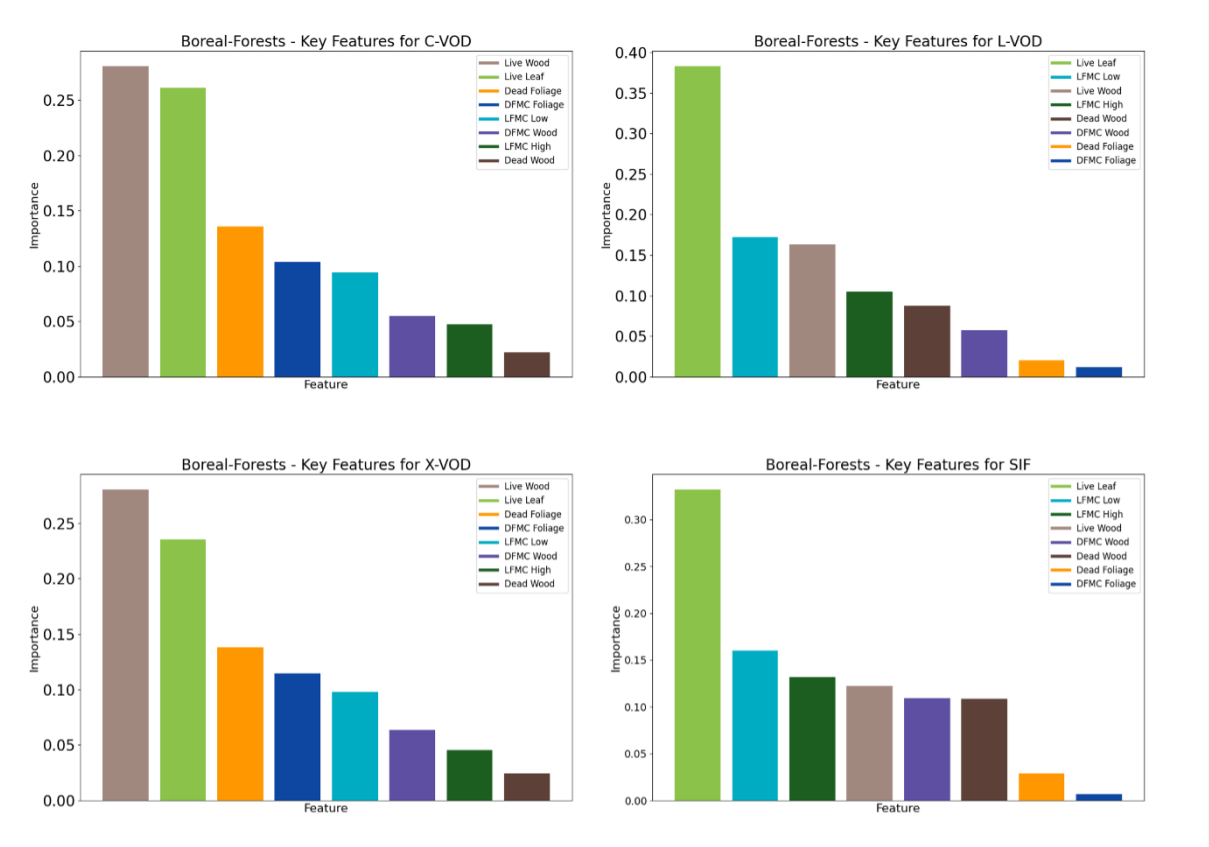

Figure A.1: Feature importance of modelled observations in Boreal Forest regions based on global daily data for 2020.

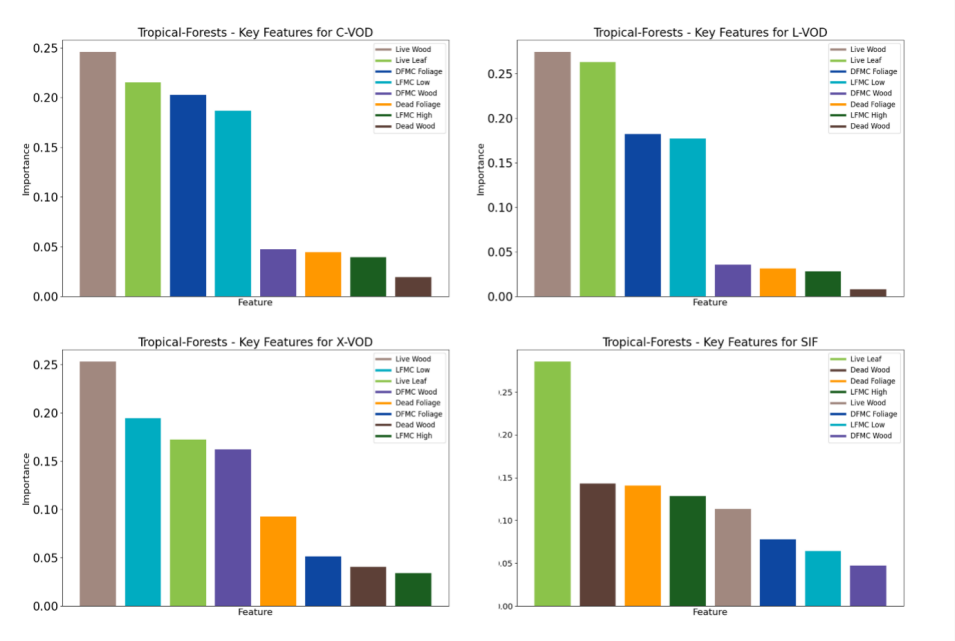

Figure A.2: Feature importance of modelled observations in Tropical Forest regions based on global daily data for 2020.

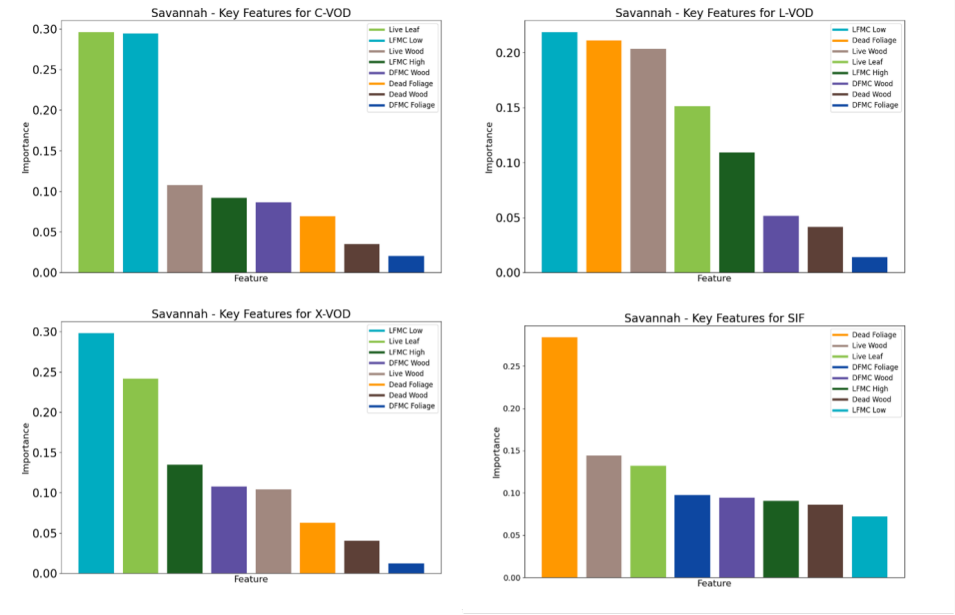

Figure A.3: Feature importance of modelled observations in Savannah regions based on global daily data for 2020.

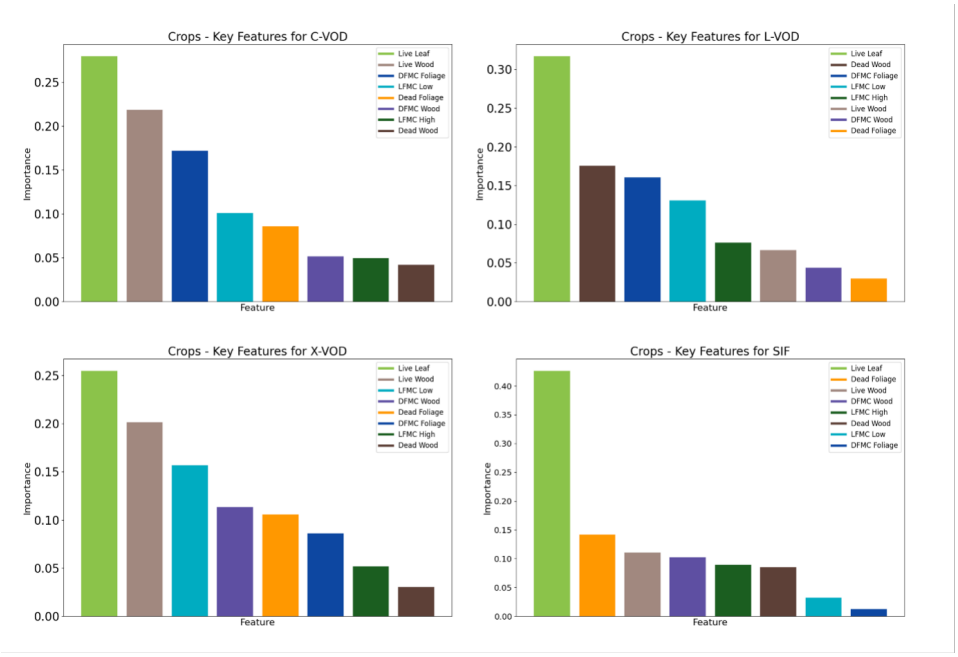

Figure A.4: Feature importance of modelled observations in Crop regions based on global daily data for 2020.

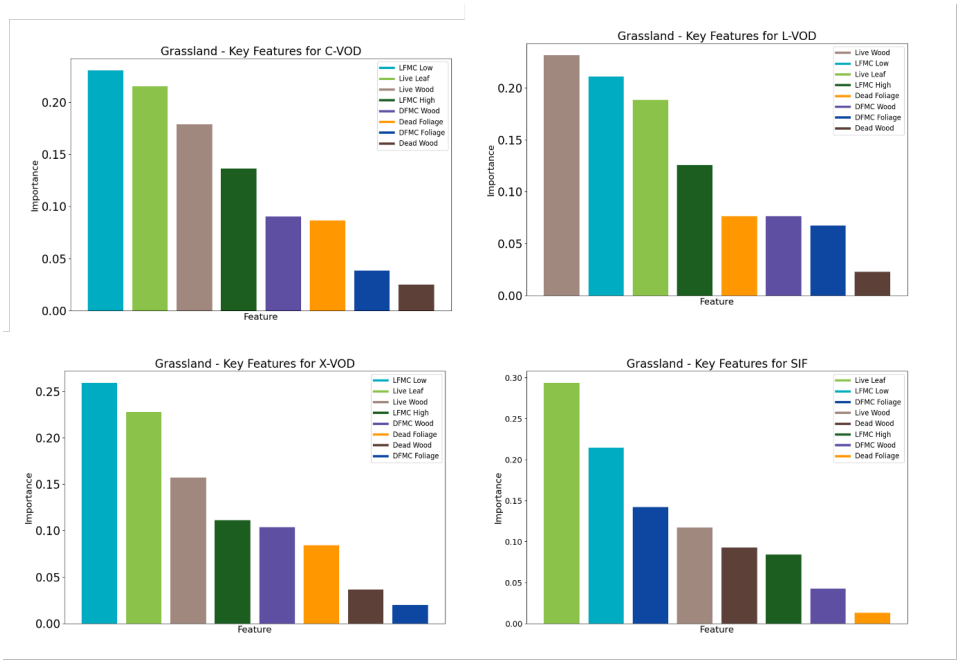

Figure A.5: Feature importance of modelled observations in Grassland regions based on global daily data for 2020.

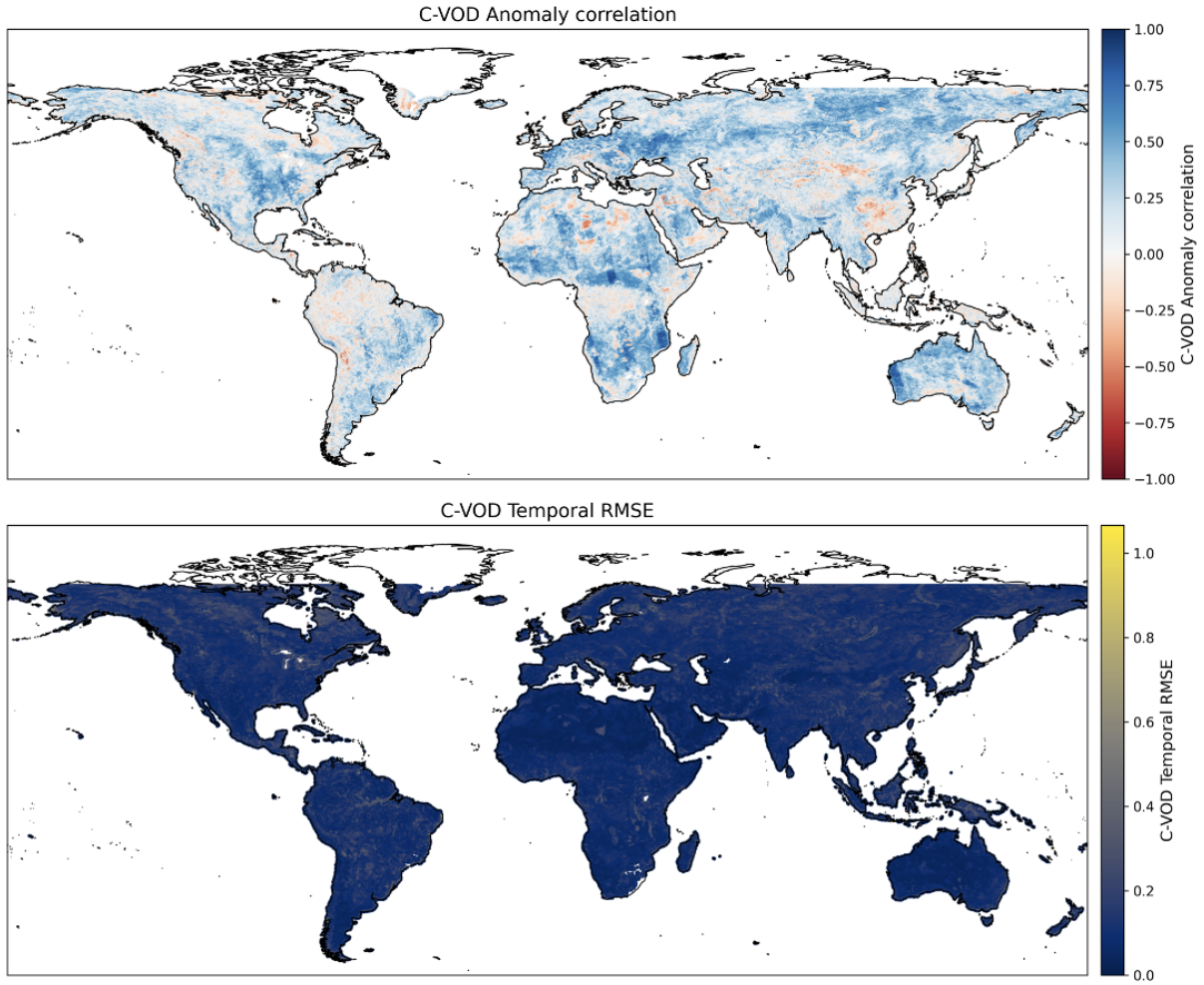

Figure A.6: Global distribution of C-VOD temporal RMSE (top) and anomaly correlation (bottom) over the independent validation year of 2021.

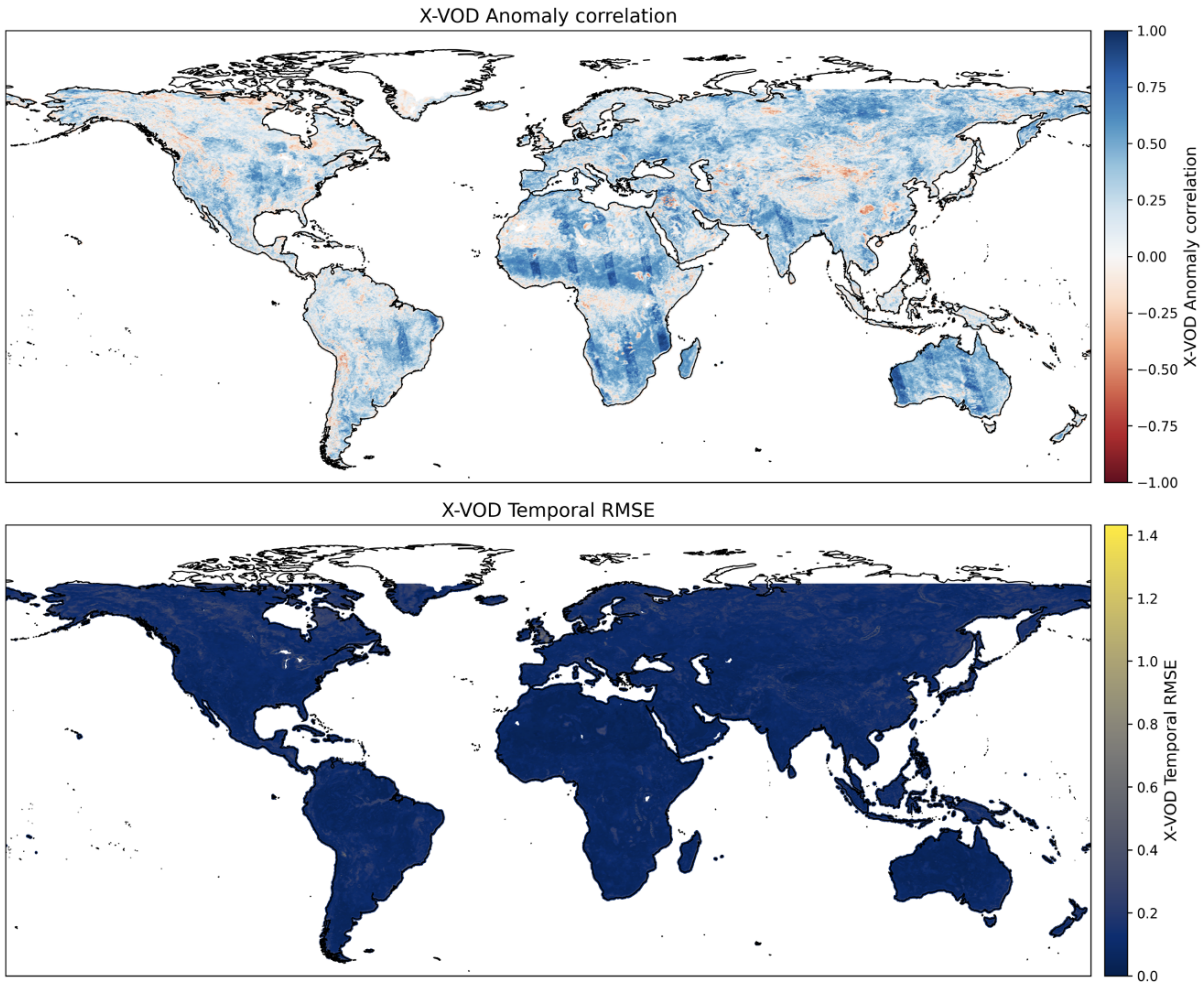

Figure A.7: Global distribution of X-VOD temporal RMSE (top) and anomaly correlation (bottom) over the independent validation year of 2021.

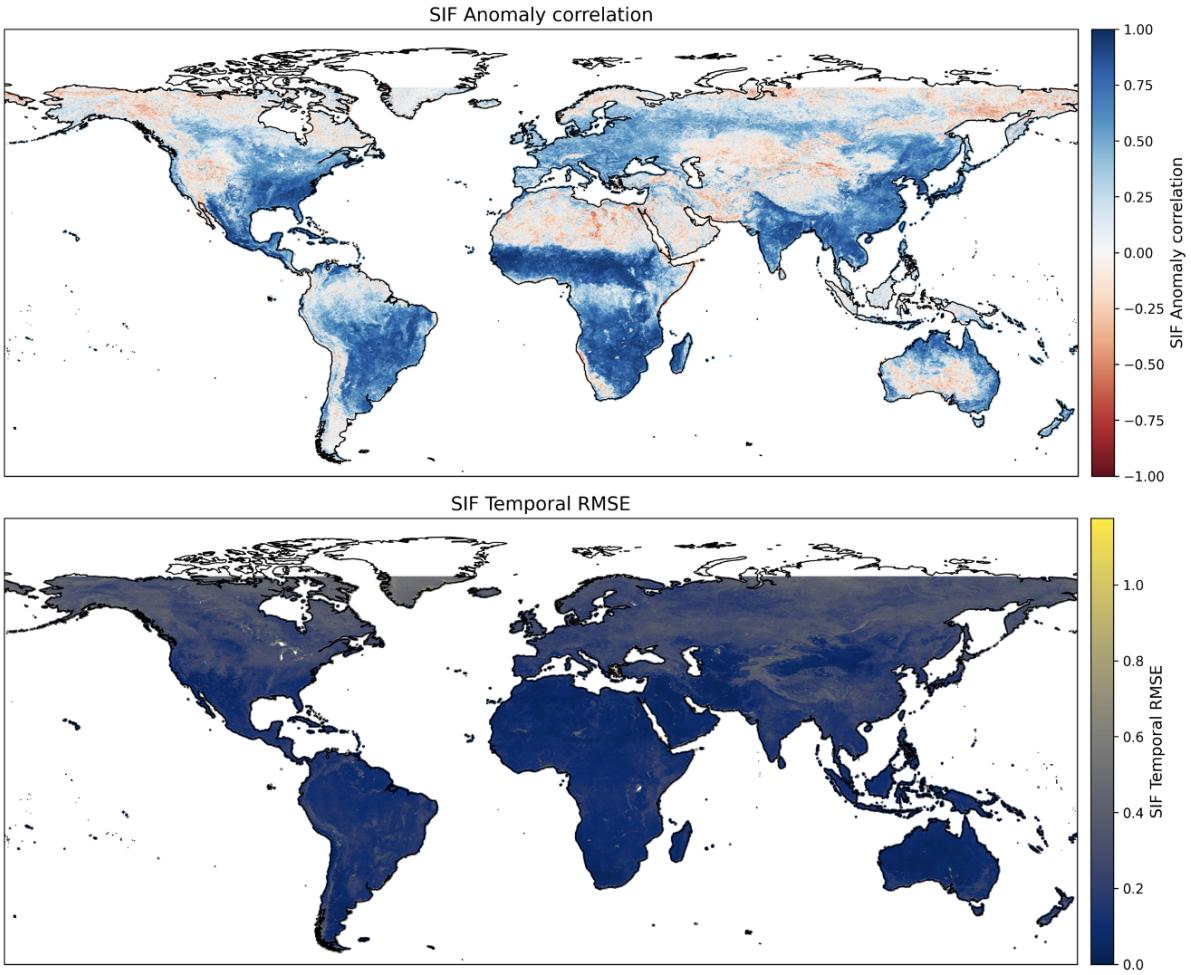

Figure A.8: Global distribution of SIF temporal RMSE (top) and anomaly correlation (bottom) over the independent validation year of 2021.

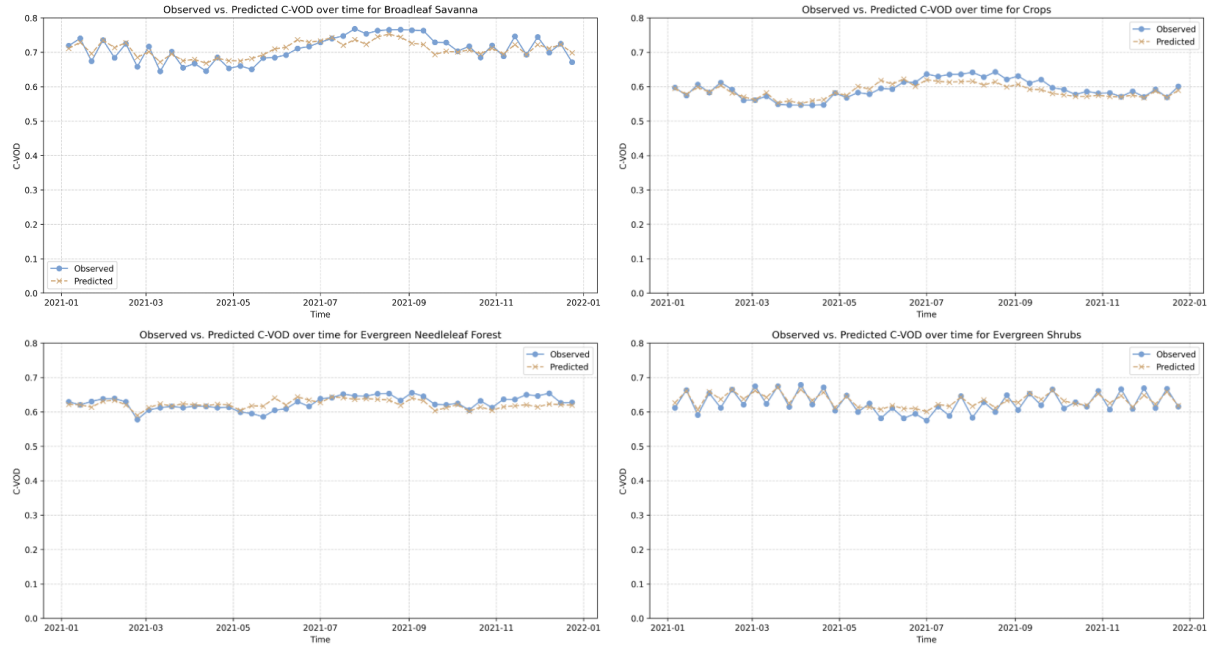

Figure A.9: Assessment of XGBoost performance for C-VOD predictions per vegetation type. Observed (solid blue) and predicted (dashed orange) C-VOD values are compared over time for the year 2021.

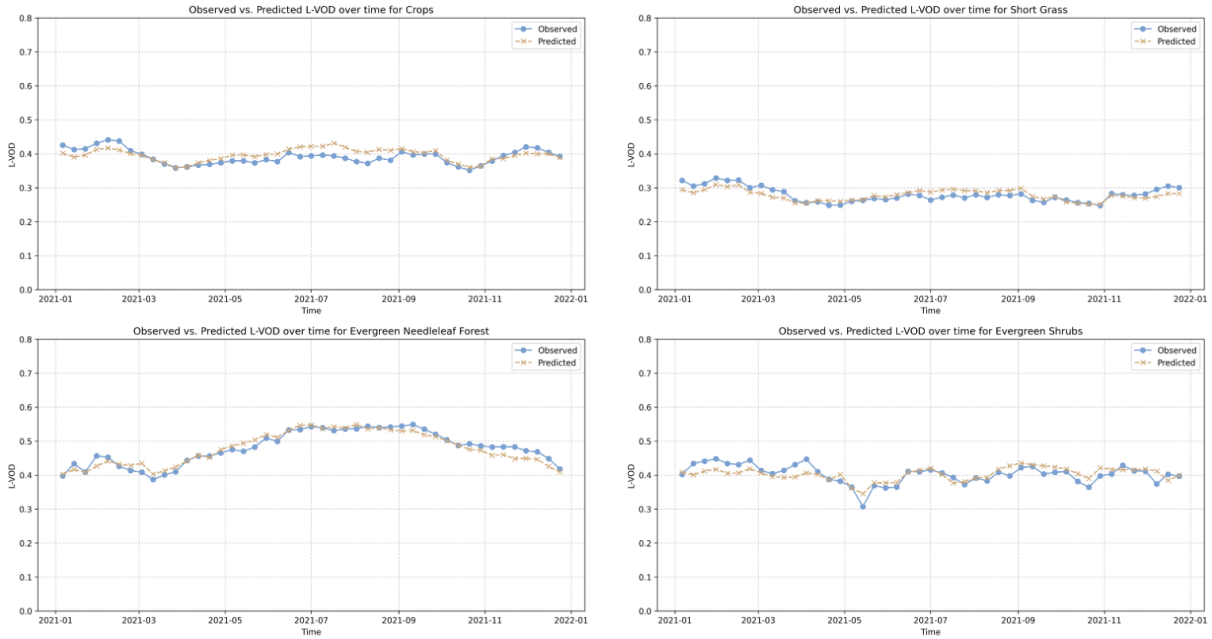

Figure A.10: Assessment of XGBoost performance for L-VOD predictions per vegetation type. Observed (solid blue) and predicted (dashed orange) L-VOD values are compared over time for the year 2021.

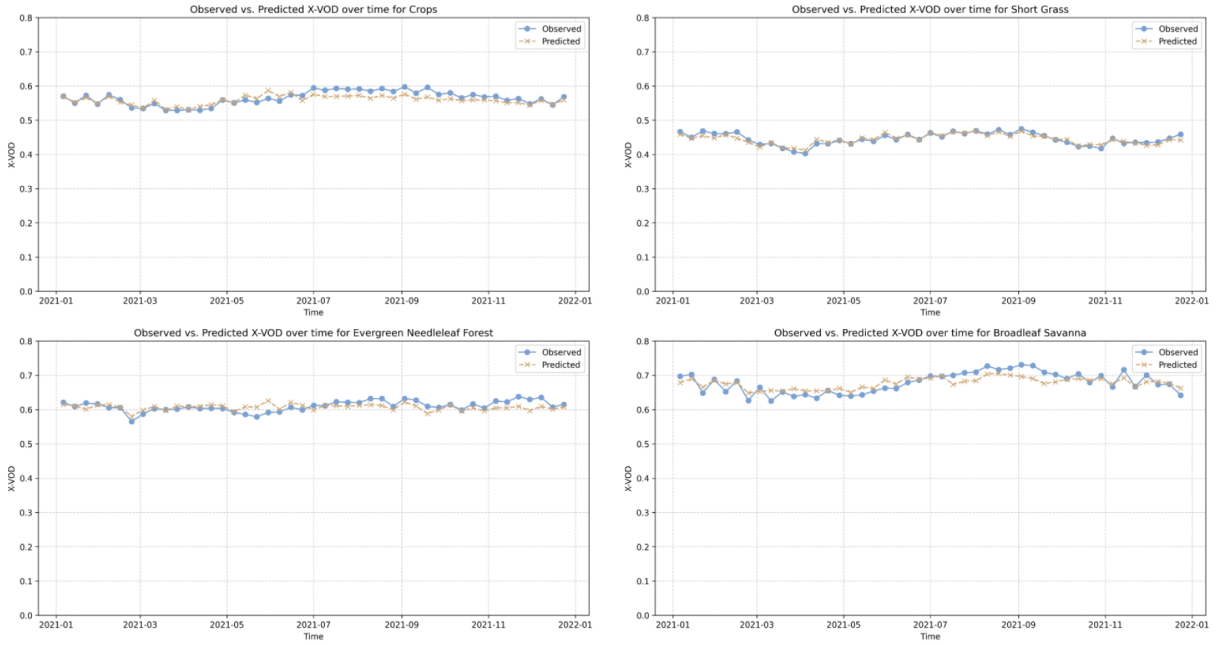

Figure A.11: Assessment of XGBoost performance for X-VOD predictions per vegetation type. Observed (solid blue) and predicted (dashed orange) X-VOD values are compared over time for the year 2021.

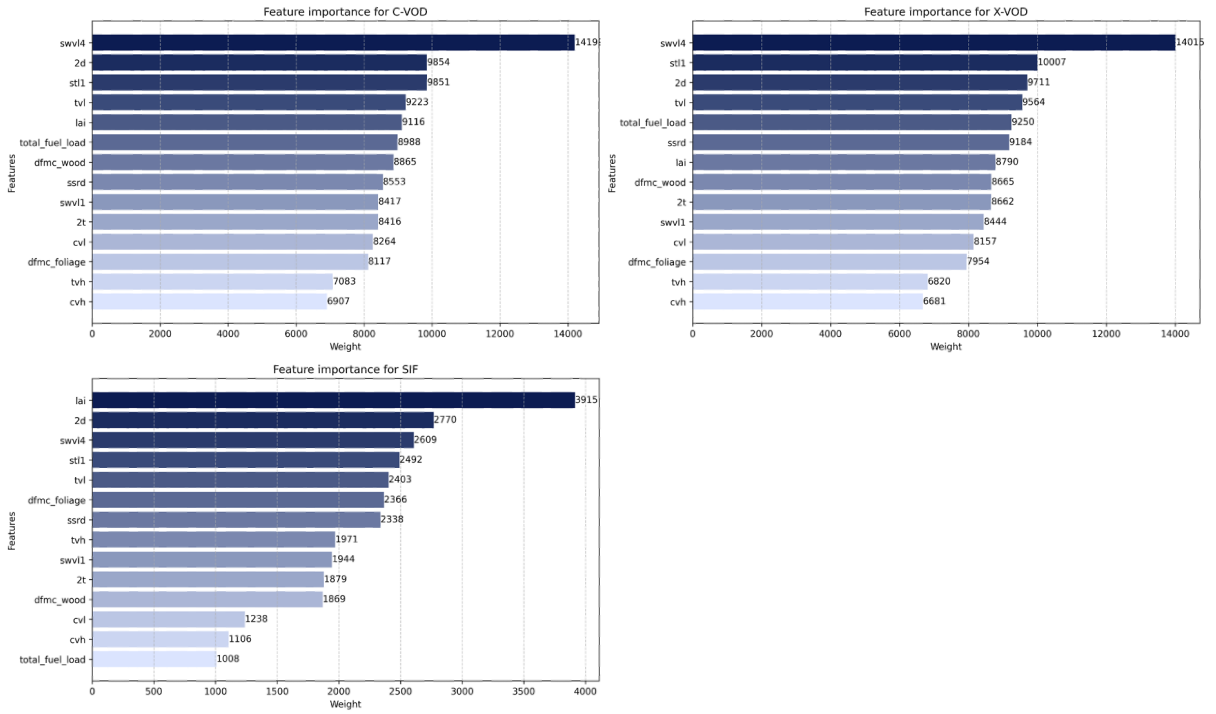

Figure A.12: Feature importance derived from XGBoost model for C-VOD, X-VOD, and SIF.

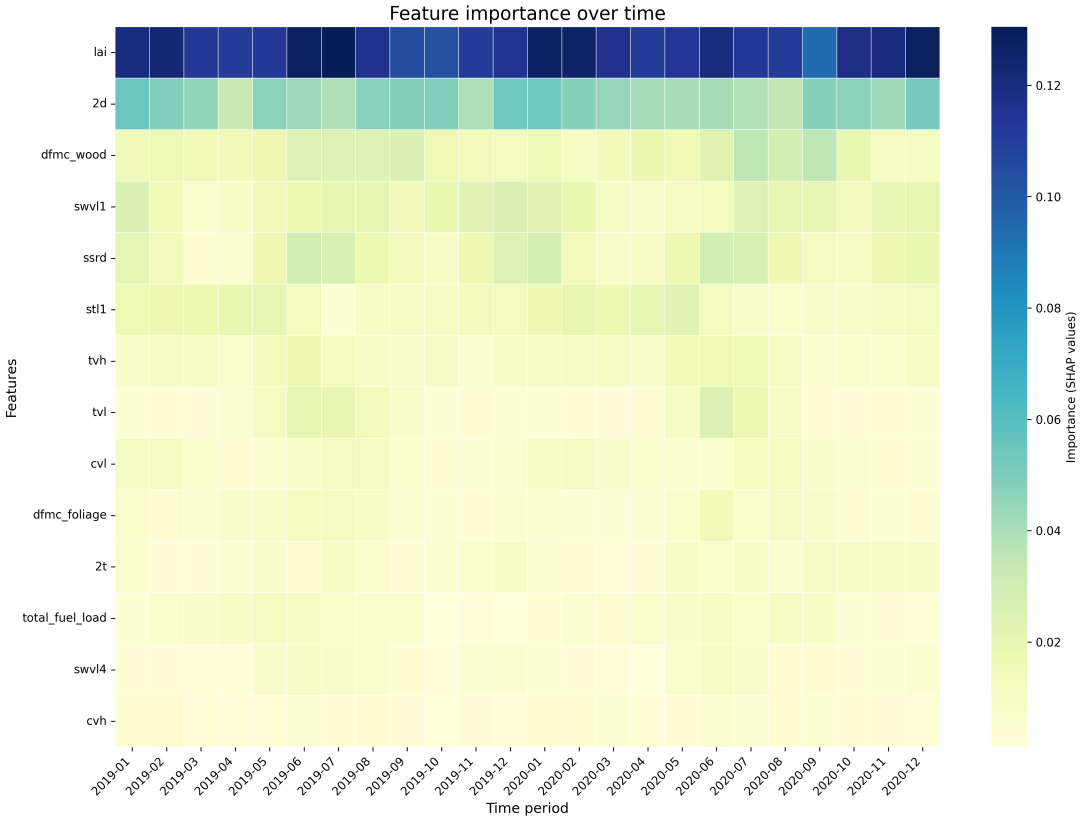

Figure A.13: Feature importance over time derived using SHAP values for an XGBoost model built separately for each month to predict SIF. The analysis spans the training period from 2019 to 2020, using monthly satellite observations. The heatmap highlights seasonal patterns. Darker blue shades represent features with greater influence on model predictions.

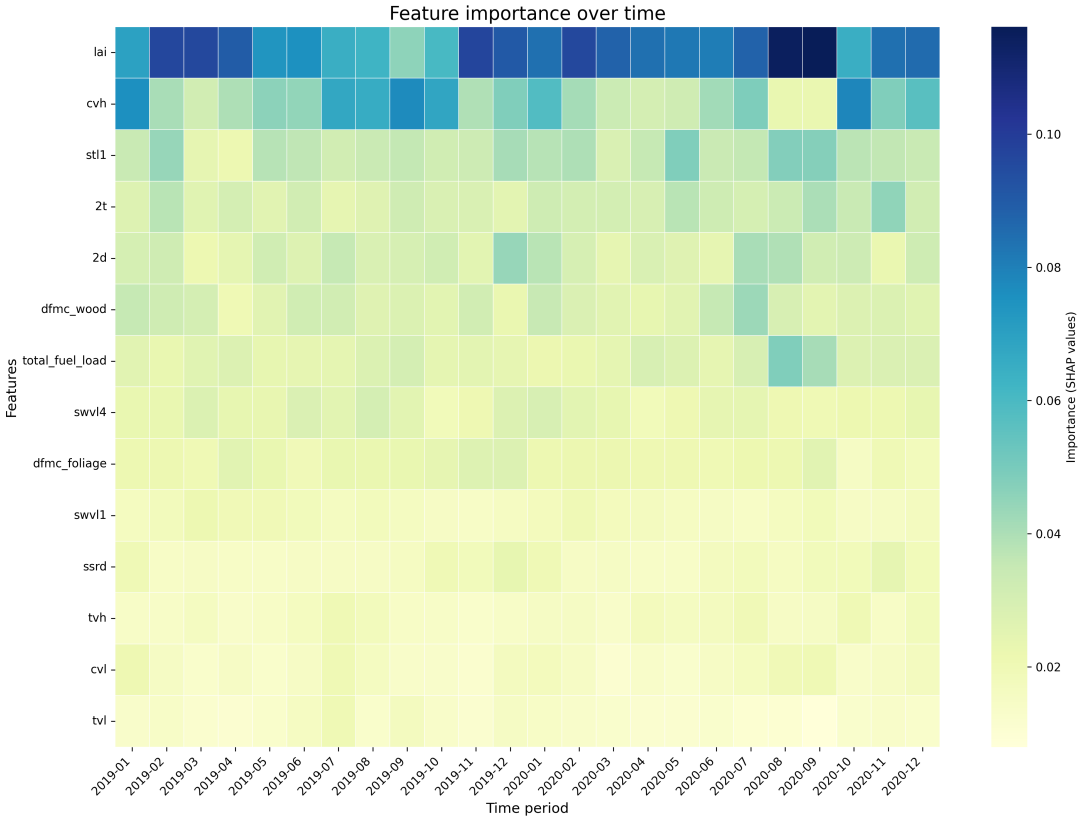

Figure A.14: Feature importance over time derived using SHAP values for an XGBoost model built separately for each month to predict C-VOD. The analysis spans the training period from 2019 to 2020, using monthly satellite observations. The heatmap highlights seasonal patterns. Darker blue shades represent features with greater influence on model predictions.

References

Agustí-Panareda, A., Massart, S., Chevallier, F., Balsamo, G., Boussetta, S., Dutra, E., and Beljaars, A.: A biogenic CO_2 flux adjustment scheme for the mitigation of large-scale biases in global atmospheric CO_2 analyses and forecasts, Atmospheric Chemistry and Physics, 2016.

Agustí-Panareda, A., Diamantakis, M., Massart, S., Chevallier, F., Muñoz Sabater, J., Barré, J., Curcoll, R., Engelen, R., Langerock, B., Law, R. M., Loh, Z., Morguí, J. A., Parrington, M., Peuch, V.-H., Ramonet, M., Roehl, C., Vermeulen, A. T., Warneke, T., and Wunch, D.: Modelling CO_2 weather – why horizontal resolution matters, Atmospheric Chemistry and Physics, 2019.

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

Boussetta, S., Balsamo, G., Beljaars, A., Kral, T., and Jarlan, L.: Impact of a satellite-derived Leaf Area Index monthly climatology in a global Numerical Weather Prediction model, International Journal of Remote Sensing, 34, 3520–3542, 2013.

Boussetta, S., Balsamo, G., Arduini, G., Dutra, E., McNorton, J., Choulga, M., Agustí-Panareda, A., Beljaars, A., Wedi, N., Munõz-Sabater, J., de Rosnay, P., Sandu, I., Hadade, I., Carver, G., Mazzetti, C., Prudhomme, C., Yamazaki, D., and Zsoter, E.: ECLand: The ECMWF Land Surface Modelling System, Atmosphere, 2021.

Burton, C., Lampe, S., Kelley, D. I., Thiery, W., Hantson, S., Christidis, N., Gudmundsson, L., Forrest, M., Burke, E., Chang, J., Huang, H., Ito, A., Kou-Giesbrecht, S., Lasslop, G., Li, W., Nieradzik, L., Li, F., Chen, Y., Randerson, J., Reyer, C. P. O., and Mengel, M.: Global burned area increasingly explained by climate change, Nature Climate Change, 14, 1186–1192, 2024.

Calvet, J.-C., Bacour, C., Bonan, B., Corchia, T., Garrigues, S., Kaminski, T., Knorr, W., Maignan, F., Peylin, P., de Rosnay, P., Scholze, M., Tartaglione, V., Vanderbecken, P., Voßbeck, M., and Vural, J.: Final review and improvement of land forward operator for SIF and MW data, Deliverable D4.2, CORSO HE project, December 2024.

Carlson, J. D., Bradshaw, L. S., Nelson, R. M., Bensch, R. R., and Jabrzemski, R.: Application of the Nelson model to four timelag fuel classes using Oklahoma field observations: model evaluation and comparison with National Fire Danger Rating System algorithms, International Journal of Wildland Fire, 16, 204–216, https://doi.org/10.1071/WF06073, 2007.

Carmona-Moreno, C., Belward, A., Malingreau, J.-P., Hartley, A., Garcia-Alegre, M., Antonovskiy, M., Buchshtaber, V., and Pivovarov, V.: Characterizing interannual variations in global fire calendar using data from Earth observing satellites, Global Change Biology, 11, 2005.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 785–794, ACM, https://doi.org/10.1145/2939672.2939785, URL http://dx.doi.org/10.1145/2939672.2939785, 2016.

Chevallier, F., Ciais, P., Conway, T. J., Aalto, T., Anderson, B. E., Bousquet, P., Brunke, E. G., Ciattaglia, L., Esaki, Y., Fröhlich, M., Gomez, A., Gomez-Pelaez, A. J., Haszpra, L., Krummel, P. B., Langenfelds, R. L., Leuenberger, M., Machida, T., Maignan, F., Matsueda, H., Morguí, J. A., Mukai, H.,

Nakazawa, T., Peylin, P., Ramonet, M., Rivier, L., Sawa, Y., Schmidt, M., Steele, L. P., Vay, S. A., Vermeulen, A. T., Wofsy, S., and Worthy, D.: Journal of Geophysical Research: Atmospheres, Journal of Geophysical Research: Atmospheres, 2010.

Courtier, P., Thépaut, J.-N., and Hollingsworth, A.: A strategy for operational implementation of 4D-Var, using an incremental approach, Quarterly Journal of the Royal Meteorological Society, https://doi.org/10.1002/qj.49712051912, 1994.

Di Giuseppe, F., Pappenberger, F., Wetterhall, F., Krzeminski, B., Camia, A., Libertá, G., and San Miguel, J.: The potential predictability of fire danger provided by numerical weather prediction, Journal of Applied Meteorology and Climatology, 55, 2469–2491, 2016.

Di Giuseppe, F., Benedetti, A., Coughlan, R., Vitolo, C., and Vuckovic, M.: A Global Bottom-Up Approach to Estimate Fuel Consumed by Fires Using Above Ground Biomass Observations, Geophysical Research Letters, 48, e2021GL095 452, 2021.

El Garroussi, S., Di Giuseppe, F., Barnard, C., and Wetterhall, F.: Europe faces up to tenfold increase in extreme fires in a warming climate, npj Climate and Atmospheric Science, 7, 30, https://doi.org/10.1038/s41612-024-00575-8, 2024.

Enquist, B. J., Brown, J. H., and West, G. B.: Allometric scaling of plant energetics and population density, Nature, 395, 163–165, https://doi.org/10.1038/25977, URL https://doi.org/10.1038/25977, 1998.

Fleming, L., Bean, J., Kirkpatrick, B., Cheng, Y., Pierce, R., Naar, J., Nierenberg, K., Backer, L., Wanner, A., Reich, A., Zhou, Y., Watkins, S., Henry, M., Zaias, J., Abraham, W., Benson, J., Cassedy, A., Hollenbeck, J., Kirkpatrick, G., Clarke, T., and Baden, D.: Exposure and effect assessment of aerosolized red tide toxins (brevetoxins) and asthma., Environmental Health Perspectives, https://doi.org/10.17226/13115, 2009.

Fuster, B., Sánchez-Zapero, J., Camacho de Coca, F., Garcia-Santos, V., Verger, A., Lacaze, R., Weiss, M., Frederic, B., and Smets, B.: Quality Assessment of PROBA-V LAI, fAPAR and fCOVER Collection 300 m Products of Copernicus Global Land Service, Remote Sensing, 12, 2020.

Guanter, L., Bacour, C., Schneider, A., Aben, I., van Kempen, T. A., Maignan, F., Retscher, C., Köhler, P., Frankenberg, C., Joiner, J., and Zhang, Y.: The TROPOSIF global sun-induced fluorescence dataset from the Sentinel-5P TROPOMI mission, Earth System Science Data, 13, 5423–5440, 2021.

Guglielmetti, M., Schwank, M., Mätzler, C., Oberdörster, C., Vanderborght, J., and Flühler, H.: Measured microwave radiative transfer properties of a deciduous forest canopy, Remote Sensing of Environment, 109, 523–532, https://doi.org/10.1016/j.rse.2007.02.003, 2007.

Guo, R., Zhao, Z., Wang, T., Liu, G., Zhao, J., and Gao, D.: Degradation state recognition of piston pump based on ICEEMDAN and XGBoost, Applied Sciences, 2020.

Hantson, S., Andela, N., Goulden, M. L., and Randerson, J. T.: Human-ignited fires result in more extreme fire behavior and ecosystem impacts, Nature Communications, 13, 2022.

Harper, A. B., Cox, P. M., Friedlingstein, P., Wiltshire, A. J., Jones, C. D., Sitch, S., Mercado, L. M., Groenendijk, M., Robertson, E., Kattge, J., Bönisch, G., Atkin, O. K., Bahn, M., Cornelissen, J., Niinemets, Ü., Onipchenko, V., Peñuelas, J., Poorter, L., Reich, P. B., Soudzilovskaia, N. A., and Bodegom, P. V.: Improved representation of plant functional types and physiology in the Joint UK

Land Environment Simulator (JULES v4.2) using plant trait information, Geosci. Model Dev., 9, 2415–2440, https://doi.org/10.5194/gmd-9-2415-2016, 2016.

Harper, A. B., Wiltshire, A. J., Cox, P. M., Friedlingstein, P., Jones, C. D., Mercado, L. M., Sitch, S., Williams, K., and Duran Rojas, C.: Vegetation distribution and terrestrial carbon cycle in a carbon cycle configuration of JULES4.6 with new plant functional types, Geosci. Model Dev., 11, 2857–2873, https://doi.org/10.5194/gmd-11-2857-2018, 2018.

Hood, S. M., Varner, J. M., Jain, T. B., and Kane, J. M.: A framework for quantifying forest wildfire hazard and fuel treatment effectiveness from stands to landscapes, Fire Ecology, 18, 2022.

McNorton, J. R. and Di Giuseppe, F.: A global fuel characteristic model and dataset for wildfire prediction, Biogeosciences, https://doi.org/10.5194/bg-21-279-2024, 2024.

McNorton, J. R., Di Giuseppe, F., Pinnington, E., Chantry, M., and Barnard, C.: A Global Probability-Of-Fire (PoF) Forecast, Geophysical Research Letters, 51, 2024.

Moesinger, L., Dorigo, W., de Jeu, R., van der Schalie, R., Scanlon, T., Teubner, I., and Forkel, M.: The global long-term microwave Vegetation Optical Depth Climate Archive (VODCA), Earth System Science Data, 12, 177–196, 2020.

Muñoz Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, 2021.

Nelson, Ralph M., J.: Prediction of diurnal change in 10-h fuel stick moisture content, Canadian Journal of Forest Research, 30, 1071–1087, 2000.

Rodríguez-Fernández, N. J., Mialon, A., Mermoz, S., Bouvet, A., Richaume, P., Al Bitar, A., Al-Yaari, A., Brandt, M., Kaminski, T., Le Toan, T., Kerr, Y. H., and Wigneron, J.-P.: An evaluation of SMOS L-band vegetation optical depth (L-VOD) data sets: high sensitivity of L-VOD to above-ground biomass in Africa, Biogeosciences, 15, 4627–4645, 2018.

Rosnay, P., Drusch, M., Vasiljevic, D., Balsamo, G., Albergel, C., and Isaksen, L.: A simplified Extended Kalman Filter for the global operational soil moisture analysis at ECMWF, Quarterly Journal of the Royal Meteorological Society, 139, 2013.

Santoro, M., Cartus, O., Carvalhais, N., Rozendaal, D. M. A., Avitabile, V., Araza, A., de Bruin, S., Herold, M., Quegan, S., Rodríguez-Veiga, P., Balzter, H., Carreiras, J., Schepaschenko, D., Korets, M., Shimada, M., Itoh, T., Moreno Martínez, A., Cavlovic, J., Cazzolla Gatti, R., da Conceição Bispo, P., Dewnath, N., Labrière, N., Liang, J., Lindsell, J., Mitchard, E. T. A., Morel, A., Pacheco Pascagaza, A. M., Ryan, C. M., Slik, F., Vaglio Laurin, G., Verbeeck, H., Wijaya, A., and Willcock, S.: The global forest above-ground biomass pool for 2010 estimated from high-resolution satellite observations, Earth System Science Data, 13, 2021.

Ulaby, F. T., Moore, R. K., and Fung, A. K.: Microwave remote sensing: Active and passive. volume 1-microwave remote sensing fundamentals and radiometry, Earth Resources And Remote Sensing, NASA, 1981.

Yebra, M., Scortechini, G., Badi, A., Beget, M. E., Boer, M. M., Bradstock, R., Chuvieco, E., Danson, F. M., Dennison, P., Resco de Dios, V., Di Bella, C. M., Forsyth, G., Frost, P., Garcia, M., Hamdi, A.,

He, B., Jolly, M., Kraaij, T., Martín, M. P., Mouillot, F., Newnham, G., Nolan, R. H., Pellizzaro, G., Qi, Y., Quan, X., Riaño, D., Roberts, D., Sow, M., and Ustin, S.: Globe-LFMC, a global plant water status database for vegetation ecophysiology and wildfire applications, Scientific Data, 2019.