CO2MVS RESEARCH ON SUPPLEMENTARY OBSERVATIONS

D4.2: final review and improvement of land forward operator for SIF and MW data

Due date of deliverable	December 2024
Submission date	December 2024
File Name	CORSO-D4-2-V1.1
Work Package /Task	WP4
Organisation Responsible of Deliverable	Meteo-France
Author name(s)	Jean-Christophe Calvet, Cédric Bacour, Bertrand Bonan, Timothée Corchia, Sébastien Garrigues, Thomas Kaminski, Wolfgang Knorr, Fabienne Maignan, Philippe Peylin, Patricia de Rosnay, Marko Scholze, Vincent Tartaglione, Pierre Vanderbecken, Michael Voßbeck, Jasmin Vural
Revision number	V1.1
Status	Issued
Dissemination Level / location	PUBLIC www.corso-project.eu

Funded by the European Union

The CORSO project (grant agreement No 101082194) is funded by the European Union.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

1 Executive Summary

The objective of this work is to build observation operators for the assimilation of radiance satellite observations: low frequency microwave brightness temperatures and backscatter coefficients, and solar induced fluorescence (SIF). Neural networks are used for the microwave observations. For SIF, both neural networks and physically based observational operators are considered. Four land surface models are used to provide predictors for training the observation operators: ISBA, ECLand, ORCHIDEE and D&B (MF, ECMWF, CEA, and ULund/iLab, respectively). ECLand is the land surface component of the IFS. This report presents the consolidated results. Machine learning (ML) was used in ISBA and ECLand to simulate ASCAT backscatter coefficients, SMOS, SMAP, AMSR2 brightness temperatures and SIF. For SIF, a process-based description of leaf fluorescence and its integration at the canopy level, taking into account the canopy structure, was used in the ORCHIDEE and D&B models. For SMOS L-VOD, an empirically derived observation operator based on hydrological status and vegetation carbon pools was used in D&B. One of the issues is the optimal temporal frequency of the SIF to properly represent the temporal variations of GPP. 1-day and 8-day frequencies were considered in the training of the SIF NN. The latter was tested in the offline ECLand model and the former in the ISBA model.

Table of Contents

2 Introduction

2.1 Background

To enable the European Union (EU) to move towards a low-carbon economy and implement its commitments under the Paris Agreement, a binding target was set to cut emissions in the EU by at least 40% below 1990 levels by 2030. European Commission (EC) President von der Leyen committed to deepen this target to at least 55% reduction by 2030. This was further consolidated with the release of the Commission's European Green Deal on the 11th of December 2019, setting the targets for the European environment, economy, and society to reach zero net emissions of greenhouse gases in 2050, outlining all needed technological and societal transformations that are aiming at combining prosperity and sustainability. To support EU countries in achieving the targets, the EU and European Commission (EC) recognised the need for an objective way to monitor anthropogenic CO_2 emissions and their evolution over time.

Such a monitoring capacity will deliver consistent and reliable information to support informed policy- and decision-making processes, both at national and European level. To maintain independence in this domain, it is seen as critical that the EU establishes an observation-based operational anthropogenic CO_2 emissions Monitoring and Verification Support (MVS) (CO2MVS) capacity as part of its Copernicus Earth Observation programme.

The CORSO research and innovation project will build on and complement the work of previous projects such as CHE (the CO_2 Human Emissions), and CoCO2 (Copernicus CO_2 service) projects, both led by ECMWF. These projects have already started the ramping-up of the CO2MVS prototype systems, so it can be implemented within the Copernicus Atmosphere Monitoring Service (CAMS) with the aim to be operational by 2026. The CORSO project will further support establishing the new CO2MVS addressing specific research & development questions.

The main objectives of CORSO are to deliver further research activities and outcomes with a focus on the use of supplementary observations, i.e., of co-emitted species as well as the use of auxiliary observations to better separate fossil fuel emissions from the other sources of atmospheric CO_2. CORSO will deliver improved estimates of emission factors/ratios and their uncertainties as well as the capabilities at global and local scale to optimally use observations of co-emitted species to better estimate anthropogenic CO_2 emissions. CORSO will also provide clear recommendations to CAMS, ICOS, and WMO about the potential added-value of high-temporal resolution $^{14}CO_2$ and APO observations as tracers for anthropogenic emissions in both global and regional scale inversions and develop coupled land-atmosphere data assimilation in the global CO2MVS system constraining carbon cycle variables with satellite observations of soil moisture, Leaf Area Index (LAI), Solar Induced Fluorescence (SIF), and vegetation biomass. Finally, CORSO will provide specific recommendations for the topics above for the operational implementation of the CO2MVS within the Copernicus programme.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverables

This deliverable aims to summarise the results of Task 4.1, which is dedicated to the design of forward operators for multi-satellite data assimilation for the analysis of land surface variables controlling carbon fluxes.

It is a consolidated version of deliverable D4.1 (First review and improvement of land surface forward operators for SIF and low frequency MW data) that was issued in December 2023.

2.2.2 Work performed in this deliverable

In this task we acquired and pre-processed SIF observations from Sentinel-5p/TROPOMI and low frequency microwave C- and X-band observations from ASCAT, AMSR2 and L-band observations from SMOS and SMAP. In parallel, observation operators for these observations were developed using neural network (NN) techniques and tested against physically based forward models using three different land surface models (ECLand, ISBA, ORCHIDEE). In this document, results are presented for each model and a comparison between the several approaches is presented.

2.2.3 Deviations and counter measures

2.3 Task 4.1 partners

Partners	
EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS	ECMWF
COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES	CEA
METEO-FRANCE	MF

3 Data

3.1 Background

The IFS-based global component of the CO2MVS assimilates the same observations as are used for Numerical Weather Prediction (NWP), such as SMOS and ASCAT. The aim of this work is to extend the use of those observations to constrain additional model variables that are relevant for the land carbon fluxes, and to develop the assimilation of existing observations that are not yet used, such as Solar Induced Fluorescence (SIF) observations.

3.2 Solar Induced Fluorescence (SIF) observations from Sentinel-5p/TROPOMI

The ESA TROPOSIF product is derived from Sentinel 5-P TROPOMI observations in the 743-758 nm near-infrared window (Guanter et al., 2021). The associated retrieval error is typically 0.5 W m^{-2} sr^{-1} m^{-2} μm^{-1}, raising a relative uncertainty on the order of 30%. Daily estimates are used (SIF_Corr_743). They are based on a time and day-length correction factor following Frankenberg et al. (2011). The products generated in the context of an ESA funded project cover the period 2018-2021 and are available from https://s5p-troposif.noveltis.fr/data-access/. Since then, the retrieval scheme has been implemented on the ESA S5P-PAL data portal which generates pre-operational L2 and L2B products on a daily basis (https://data-portal.s5p-pal.com/products/troposif.html). Gridded spatio-temporal binned (0.1°/8-day) estimates of these L2B TROPOSIF retrievals (SIF and vegetation indices) have been generated on a regular basis from 2018 onwards at LSCE (https://doi.org/10.14768/b391bda9-fdfb-40cb-9deb-59b121a18cfb).

3.3 C-band microwave observations from ASCAT

The ASCAT data consist of C-band radar backscatters (sigma0). The ASCAT sigma0 at an incidence angle of 40 degrees is available from the EUMETSAT HSAF service. Digital Object Identifier (DOI) is: https://doi.org/10.15770/EUM_SAF_H_0009

3.4 C-band and X-band microwave observations from AMSR2

The AMSR2 data consist of C-band and X-band brightness temperatures (TB). Data at higher microwave frequencies are also available but they are less sensitive to land surface variables. DOI for original L1B-TB GCOM-W/AMSR2 L1B JAXA data is:

https://doi.org/10.57746/EO.01gs73ans548qghaknzdjyxd2h

3.5 L-band microwave observations from SMAP

The SMAP data consist of L-band brightness temperatures (TB). The original L1C data can be accessed from https://nsidc.org/data/spl1ctb/versions/5.

3.6 L-band microwave observations from SMOS

The SMOS data consist of L-band brightness temperatures (TB). In this work, the L3TB brightness temperature product from the Centre Aval de Traitement des Données SMOS (CATDS) is used (https://doi.org/10.12770/6294e08c-baec-4282-a251-33fee22ec67f).

4 Methods

4.1 ORCHIDEE modelling framework

CEA worked on assessing the potential of space-borne SIF data to improve the space-time distribution of GPP simulated by the ORCHIDEE (Organizing Carbon and Hydrology In Dynamic Ecosystems) land surface model. The observation operator for SIF follows a process-based description of the leaf fluorescence and its integration at canopy level accounting for the canopy structure. A revised 2-flux version of SIF and photosynthesis modelling, in comparison to the one described in Bacour et al. (2019) is used. The main parameters of ORCHIDEE related to photosynthesis and phenology are calibrated using the assimilation of space-borne estimates of SIF from Sentinel-5p observations (TROPOSIF product generated by LSCE at 0.1°/8-day) and *in situ* Gross Primary Productivity (GPP) data. In order to assess the informational constraint brought by satellite SIF data on the model parameter, three assimilation experiments are conducted: in the first experiment, SIF data are assimilated alone over an handful of selected 0.1° pixels for each plant fuctional type (PFT) and the impact on GPP simulations is assessed by comparing them with independent data-driven GPP products (FLUXCOM-X-BASE, Nelson et al., 2024; FluxSatLUXSAT, Joiner et al. 2014) over independent pixels as well as with in situ GPP estimates from eddy-covariance instrumented sites (FLUXNET - Baldocchi et al., 2001); in the second experiment, only *in situ* GPP data are assimilated; in the last experiment, satellite SIF and *in situ* GPP data are co-assimilated. The evaluation of the optimized simulations is performed using pixels/sites not used in assimilation. For the various sites, a novel characterization in terms of PFT fractions has been conducted specifically in order to improve the consistency of ORCHIDEE simulations with the tower footprints (Tartaglione et al., in prep).

The optimized parameters are finally applied to perform simulations of GPP at a regional scale and at a global scale, which are compared to those obtained with the standard parameter values and to reference GPP products (FLUXCOM-X-BASE and FluxSat). The differences between the prior and optimized simulations, and with the reference data, highlight the combined constraint brought by GPP and SIF data to improve the model prediction.

4.1.1 Land surface model

ORCHIDEE is a mechanistic land surface model (LSM) designed to simulate the fluxes of carbon, water, and energy between the biosphere and atmosphere (Krinner et al., 2005). It is a component of the Earth System Model developed by Institut Pierre-Simon Laplace IPSL-CM. The model operates from local to global scale, representing the spatial distribution of vegetation using fractions of plant functional types (PFTs) for each grid cell. Currently 14 PFTs are used: https://orchidas.lsce.ipsl.fr/dev/lccci/orchidee_pfts.php. Recent developments were made for this study with both photosynthesis and fluorescence modules that now account for the partition between sun and shaded leaves within the canopy (Zhang et al. 2020). The fluorescence module, now following a 2-flux radiative transfer scheme, differs from that described in Bacour et al. (2019), which was based on a parametric emulator of the SCOPE model (van der Tol et al., 2009). The calculation of chlorophyll fluorescence emission at the leaf level follows the Fluor-MODleaf concepts (Pedrós et al., 2010) and the integration of SIF at the canopy level follows a SAIL-like two-stream scheme (based on Yang et al., 2017).

4.1.2 Data assimilation approach

We use the ORCHIDAS Data Assimilation tool (https://orchidas.lsce.ipsl.fr/) (MacBean et al., 2022; Bacour et al., 2023). The assimilation relies on a Bayesian framework with a global misfit function between model simulations and observational data, considering error

covariance matrices and prior information. We use a Genetic Algorithm optimization approach (Goldberg, 1989), to iteratively minimize the misfit function (Bastrikov et al., 2018).

The assimilations are conducted on a PFT-basis, against GPP data (site scale estimates or FluxSat data for the PFTs for which no *in situ* data are available) and TROPOMI SIF retrievals for a collection of selected homogeneous grid cells. Three assimilation experiments are performed: one in which space-borne SIF data (0.1°/8-day resolutions) are assimilated alone, one in which only *in* situ GPP data are assimilated, and one in which SIF and GPP data are co-assimilated. The co-assimilation of these two variables is expected to prevent parameter overfitting and help better constraining parameters mostly related to the SIF observation operator vs those impacting both SIF and GPP.

We have rebinned at 8-day/0.1° the daily averaged SIF retrievals of the TROPOSIF product (Guanter et al., 2021), over the period 2019-2022 (https://doi.org/10.14768/b391bda9-fdfb-40cb-9deb-59b121a18cfb). Only observations passing the quality flag and associated with view zenith angles below 40° and cloud fraction below 0.5 are considered. We selected twenty grid cells for each of the 14 vegetation PFTs, with the highest thematic homogeneity and ensuring a correct sampling of the global distribution. Daily *in situ* GPP estimates from FLUXNET (Baldocchi et al., 2001; Pastorello et al., 2020) are assimilated.

The diagonal of the error covariance matrix on observations is populated by the root mean square difference (RMSD) between observations and model simulations using prior standard parameter values (MacBean et al., 2022; Bacour et al., 2023). We then balanced the misfit functions associated respectively to SIF and GPP at the first iteration to account for the larger number of GPP observations. We optimized parameters related to photosynthesis, phenology, SIF and hydrology. About 15 parameters are optimized when SIF data are assimilated and 10 parameters when *in situ* GPP data are assimilated. All parameters are optimized in the co-assimilation experiment.

4.1.3 Justification of the use of weekly means for SIF

We use TROPOSIF weekly means in order to decrease the relatively high random error associated to individual retrievals, and to smooth directional effects, which are usually not modelled in land surface models. Using instantaneous values would also have meant managing the time of the acquisition in the model to get the correct corresponding time step for GPP. Regarding data assimilation in the ORCHIDEE land surface model, the minimization algorithms used to optimize model parameter values usually compute squared differences between model and observations, and they would be very sensitive to instantaneous large errors. This would require specifying variable observation/model errors (R matrix) with larger errors for “outliers”, which is still a difficult task. The linearity of the relationship between SIF and GPP usually breaks down at high spatial/high temporal resolution. Incorrect parameterizations of their respective temporal dynamics in the model may introduce some estimation bias if instantaneous data are assimilated. In addition, accounting for instantaneous data is associated with higher computational burdens (increased frequency of inputs/outputs, memory, etc.) which may become limiting when considering observations over many pixels. This is another incentive to work with weekly means.

4.2 ML-based observation operators for ISBA

MF worked on SIF, ASCAT, SMAP, SMOS, and X-band AMSR2 data over agricultural areas, at a global scale. The objective is to assimilate these observations in the ISBA land surface model using MF’s global Land Data Assimilation System (LDAS-Monde) tool. Observation operators based on neural networks (NNs) are trained with ISBA simulations and LAI observations from the PROBA-V satellite to predict SIF and the microwave signal. The globally trained NN-based observation operators (one NN for all grid cells) is implemented in LDAS-

Monde, which allows the sequential assimilation of backscatter observations (Corchia et al. 2023).

The daily SIF product is simulated at a global scale using a machine-learning method similar to the one used for microwave observations.

Table 1: ISBA global cropland forward operator training configurations for SIF, ASCAT sigma0, SMAP, SMOS, AMSR2 TB. All microwave data are in V polarisation. Structural predictors such as longitude, latitude, day of year (DOY) and soil texture (sand and clay fractions) are used together with biophysical predictors derived from either satellite observations (LAI from CLMS) or ISBA model simulations (surface soil moisture, surface soil wetness index, surface temperature, WG2, SWI2, TG2, respectively).

	SIF	**ASCAT sigma0**	**SMAP TB**	**SMOS TB**	**AMSR2 TB**
Hidden layers	2	2	3	2	3
Neurons	128 / 128	64 / 64	128 / 32 / 8	64 / 64	128 / 32 / 8
Train/val/test [%]	40 / 10 / 50	50 / 10 / 40	60 / 20 / 20	50 / 10 / 40	60 / 20 / 20
Learning rate	adaptive	0.01	0.01 (adaptive)	0.01	0.01 (adaptive)
Activation function	ReLU	ReLU	ReLU	ReLU	ReLU
Loss function	Huber loss	mean-squared error	mean-squared error	mean-squared error	mean-squared error
Predictors	LAI	LAI, WG2, TG2	LAI, SWI2, TG2	LAI, WG2, TG2	LAI, SWI2, TG2
Predictors (structural)	lon, lat, DOY	lon, lat, SAND, CLAY	lon, lat, DOY, altitude, topo complexity	lon, lat, SAND, CLAY	lon, lat, DOY, altitude, topo complexity

4.2.1 Land surface model

The version of the model that is used for this study is capable of representing soil moisture, soil temperature, photosynthesis, plant growth and senescence. Phenology is driven entirely by photosynthesis, using a simple allocation scheme. Net leaf CO_2 assimilation is used to represent the incoming carbon flux for leaf biomass growth. A photosynthesis-dependent leaf mortality rate is calculated. The balance between the leaf carbon uptake and the leaf mortality rate results in an increase or a decrease in leaf biomass. Leaf biomass is converted to LAI using a fixed value of specific leaf area (SLA) per plant functional type.

4.2.2 Observation operators

The simulated LAI is flexible and LAI observations can easily be used to correct the simulated LAI using a simple Kalman filter in the LDAS-Monde sequential data assimilation framework. Variables simulated by the ISBA model, such as soil moisture and soil temperature, can be

used to train neural networks (NNs) able to simulate satellite observations such as SIF, brightness temperatures (TB) and radar backscatter coefficients (sigma0). Since the simulated LAI may be affected by strong biases due to the lack of representation of anthropogenic processes (e.g. crop rotation), satellite LAI observations from the Copernicus Land Monitoring Service (CLMS) are used during the NN training phase rather than modelled LAI. NN observation operators for SIF, TB and sigma0, need to be constructed before implementing the sequential assimilation of these quantities. Checking the ability of the sequential assimilation to improve the simulation of the observations is one way of ensuring that major model biases are not introduced into the observation operator. The ISBA training configurations for SIF, ASCAT, SMAP, SMOS, and AMSR2 are summarized in Table 1.

4.3 ML-based observation operators for the IFS

The work of ECMWF was dedicated to the design of machine learning-based observation operators to assimilate passive multi-frequency microwave data (AMSR2), active microwave data (ASCAT backscatter normalized at 40°) and SIF (TROPOMI) in the ECMWF Integrated Forecasting System (IFS). The implementation of the assimilation of these observations consists of four steps: (1) development of the training database including quality control and filtering out unfavourable surfaces (snow, frozen soil, orographic regions, water bodies); (2) analysis of the information content using process-based knowledge and sensitivity analysis to identify the most influent predictors on the satellite signal; (3) training, hyperparameter tuning and evaluation on independent dataset of the ML-model (here Gradient boosted trees (XGBOOST) and feedforward neural network (NN)); (4) implementation and test of the assimilation of the observations in coupled or offline data assimilation experiments. The challenge is to design a ML-based observation operator which is accurate enough to predict the model-counterpart of the observation and which has enough sensitivity to the variable that will be updated by the data assimilation system. The work presented for AMSR2 concerns step 1 and 2. For ASCAT and SIF, step 1, 2, 3 are presented here and step 4 will be addressed in separate reports.

4.3.1 AMSR2 information content analysis

An existing AMSR2 training database, collocating observations and model information at ECMWF (credit: Alan Geer, ECMWF), which is shared with the CERISE project, was produced in collaboration between CERISE and CORSO, using the IFS Cycle 47r1 and the all-sky observation framework of IFS cycle 47r3, using a N256 reduced Gaussian grid, over a 15-month period (2020/07/01-2021/09/30). The database includes the brightness temperatures from the 14 AMSR2 channels in both vertical and horizontal polarizations. The training database has been modified for its use in CORSO with the introduction of vegetation and carbon flux variables, soil and vegetation types. A preliminary work has focused on the evaluation of the correlations between the brightness temperatures in C, X, Ku and Ka bands and the IFS model fields (vegetation parameters, soil moisture, soil temperature, albedo among others).

4.3.2 ASCAT

A four-year training database (2016-2019), which relates ASCAT backscatter at 40° to ERA-5 reanalysis variables was used (Aires et al., 2021). The ERA-5 model fields were interpolated at the time and location of the ASCAT observations. The spatial and temporal sampling of the training database is that of ASCAT (25 km and daily frequency over most locations). The IFS model fields used to predict ASCAT backscatter at 40° include soil moisture and soil temperature in the first 3 soil layers (up to 1m depth) and Leaf Area Index (LAI). Frozen soil, water bodies, snow area and mountain area were excluded from the training database. The ASCAT ML model is trained over 2016-2018 period and tested over 2019.

Several architectures of NN and XGB (Chen et al., 2016) models were tested to simulate ASCAT backscatter normalized at 40° at global scale from the IFS model fields. ML models

were developed in the observation space, at global scale, with the use of latitude and longitude as additional features to represent local observation conditions. The NN model was developed using the PYTORCH ML package.

4.3.3 SIF

For SIF, the ML model is trained over 2019-2020 and tested over 2021. The database used to develop the SIF ML-based observation operator models consists of the ECLand (land surface component of the IFS) simulated fields collocated with the Copernicus S5p TROPOMI satellite observations provided by LSCE at 0.1° spatial resolution and 8-d time frequency (see Section 4.1.3). Cloud filtering and unfavourable satellite geometry screening, namely excluding samples with view and solar zenith angles above 60° and 70°, respectively, have been applied to the TROPOSIF dataset. Orographic area, snow, frozen soil area and water bodies samples, for which the SIF signal is too uncertain, were removed from the database. The period 2019-2020 is used for training which is sufficient to capture the main spatial and seasonal patterns of SIF at 0.1° spatial resolution and 8-d time frequency. 2021 is used for the validation step which consists in refining the set of predictors and tuning the ML model hyperparameters. The candidate models were evaluated over 2022.

Table 2: IFS global forward operator training configurations for SIF. SM is the soil moisture of the first soil layer, SM-1m is the root-zone soil moisture within 1m of soil, ST is the soil temperature of the first soil layer, T2M and D2M are the 2m temperature and dewpoint, CVH and CVL are the fractions of high and low vegetation respectively, SWDOWN is the short-wave downwelling radiation

Model	Vegetation	Atmospheric forcing	Surface conditions	Localization in space and time
M1	LAI, CVH, CVL,	SWDOWN, T2M, D2M	SM, SM-1m, ST,	No
M2	LAI	SWDOWN, T2M, D2M	SM, SM-1m, ST,	No
M3	Satellite LAI	None	None	Time, latitude, longitude

The predictors were selected from process-based knowledge of the SIF drivers at canopy scale which are i) the vegetation structure ii) the photosynthetic active radiation represented by the short-wave downwelling radiation (SWDOWN), iii) the vegetation physiology processes (GPP), the environmental conditions (soil moisture, soil temperature, 2m air humidity and temperature) and vegetation characteristics (high (CVH) and low (CVL) vegetation fractions). Several combinations of predictors were tested. Table 2 presents the 3 models that will be discussed in this report. M1 and M2 rely on selected ECLand/IFS physical predictors while M3 was trained from the Copernicus land satellite LAI combined with spatial (latitude, longitude) and temporal (week of the year) localization variables. M2 is a reduced version of M1 in which the low impacts predictors, CVH and CVL (Figure 15) were removed.

XGB was chosen because of its easy implementation and its fast training and hyperparameter tuning. Besides, its regularization and tree pruning capability minimize risk of overfitting. For

each set of hyperparameters, the value which leads to the minimum RMSE between the predicted and observed SIF over the validation dataset are selected. The number of trees and the learning rate, which are the main drivers of XGB performances, are first chosen, followed by the maximum depth and the min child weight which control model complexity. The default values were used for the gamma and lambda regularization parameters which had little impacts on the prediction performances.

4.4 D&B modelling framework

ULUND and iLab used observation operators for SIF and SMOS Vegetation Optical Depth (L-VOD) that are coupled to the D&B terrestrial ecosystem community model (Knorr et al. 2024). The D&B model is based upon three interconnected sub-model components: (i) photosynthesis and autotrophic respiration, (ii) energy and water balance, and (iii) carbon allocation and cycling, including heterotrophic respiration. The first component includes a process-based description of uptake of CO_2 via plant photosynthetic activity (gross primary production, GPP), regulated by temperature, light absorption across the canopy, and stomatal control, and of carbon loss from the respiration of live vegetation (RA, autotrophic respiration). The remainder, net primary production (NPP = GPP - RA), is then passed over to the Carbon Allocation and Cycling component and distributed among the various carbon pools. The Energy and Water Balance component regulates the energy input to and output from the canopy in the form of radiative, latent and sensible heat exchange with the atmosphere, taking into account the hydrological status of the canopy and soil, as well as the plant transpiration. Components (i) and (ii) are based on BETHY (Knorr, 2000), and component (iii) on DALEC (Williams et al., 2005). D&B is set up for assimilating remotely sensed soil moisture and FAPAR besides SIF and L-VOD. Both soil moisture and FAPAR are internally calculated in D&B.

4.4.1 SIF observation operator

The SIF observation operator has been described in detail in Knorr et al. (2024). Basically, we use the formulation of Gu et al. (2019), which is motivated by the direct link to the photosynthesis routines and its modular implementation fitting the overall D&B modelling strategy. The hourly canopy layer SIF, S_n, is calculated as a function of mainly the electron transport in canopy layer n calculated by D&B's photosynthesis component, and of the photon escape probability from the canopy which in D&B is calculated explicitly by the layered 2-stream model in the energy and water balance component (Quaife, 2024). As an extension to the model by Gu et al. (2019) in view of the anticipated calibration in a data assimilation scheme, we further introduce a scaling factor s_{SIF}. This scaling factor compensates for large uncertainties in some of the constants needed to calculate S_n and in the spectral conversion from $mol\ m^{-2}s^{-1}$ (total flux of photons into the hemisphere above the canopy for all wavelengths) as calculated by the model to $Wm^{-2}s^{-1}nm^{-1}sr^{-1}$ (energy flux units per steradian, per nano-metre of the SIF spectra), that is usually used for satellite measurements and in situ observations. For the conversion we use a SIF emission spectrum observed at the Hyytiälä site in Finland (Magney et al., 2019). The SIF spectrum was measured for four Scots pine trees at light level of 1200 $\mu mol\ m^{-2}s^{-1}$ and then averaged.

4.4.2 SMOS L-VOD observation operator

VOD describes the attenuation of microwave radiation at a given wavelength caused by the presence of vegetation and mainly depends on the dielectric properties (regulated by water content, temperature and chemical composition) and the canopy structure. Due to the relatively static nature of structure, dynamics of VOD are generally attributed to changes in above ground biomass and water content (Konings et al., 2019). To link the model state to the L-band VOD measurements from SMOS a semi-empirical relationship is used. L-VOD is calculated as a function of the leaf and woody biomass pools as well as of the fractional plant-available soil water content, f_{soil}, and the ratio of actual to potential transpiration, f_E. f_{soil}

describes slow changes in the plant's hydrological status whereas f_E fast changes. Following Schwank et al. (2021) an explicit temperature dependency is introduced to approximate the theoretically derived behaviour of VOD around the freezing point. Multiplicative factors are introduced in the empirical relationship to account for that VOD will be zero if no biomass is present and for positive VOD even if vegetation water stress is at its maximum. The details and exact formulation of the empirical relationship together with prior values for the parameters relating L-VOD to the quantities described above are given in Knorr et al. (2024).

5 Results

5.1 ORCHIDEE modelling framework

We evaluate the model's initial performance (i.e. using the prior parameter values) through statistical comparisons between simulations and observations for each vegetation PFT. The assessment is performed 1) over the selected homogeneous grid cells at 0.1°resolution against SIF retrievals from TROPOSIF (weekly, over 2019-2022) and GPP estimates from the FLUXCOM-X-BASE and FluxSat data-driven approaches (daily, over 2001-2021), and 2) over a collection of eddy-covariance sites against daily *in situ* GPP data.

Figures 1 and 2 present the boxplot distributions over the 14 model PFTs of three metrics for both SIF and GPP - RMSD, bias, and coefficient of determination (R^2) - computed between the prior model simulations and the evaluation datasets over the selected 0.1° pixels. A similar evaluation for GPP is performed for a set of eddy-covariance sites in Figure 3. A large variability in terms of model-data agreement is seen across PFTs, with lower performance over crops and C4 grass. For SIF, the model data error is usually lower than the error associated to the daily SIF retrievals (0.5 mW m^{-2} sr^{-1} nm^{-1}); note that the latter has been strongly reduced by the space-time binning of the TROPOSIF data. We observe a generally consistent (mis)match between model and data across the various PFTs for both SIF and GPP variables for the pixel scale evaluation (i.e. higher/lower errors in SIF simulation associated with higher/lower errors in modelled GPP). This result suggests that adjusting one of these variables (e.g. SIF) has potential to have a positive impact on the other (e.g. GPP). However, this is not the case for some PFTs (e.g., tropical evergreen broadleaf forest or C4 grass), indicating where co-assimilation should be even more relevant. Regarding GPP, ORCHIDEE seems generally in better agreement with the data-driven products than with the *in situ* data; the lower R^2 values seen for the eddy-flux estimates (compared to the data-driven GPP products) highlight some difficulty of the model in capturing the day-to-day variability of the measurements at higher spatial resolution.

These model-data mismatches highlight the strong potential of data assimilation to improve model performance with respect to both SIF and GPP.

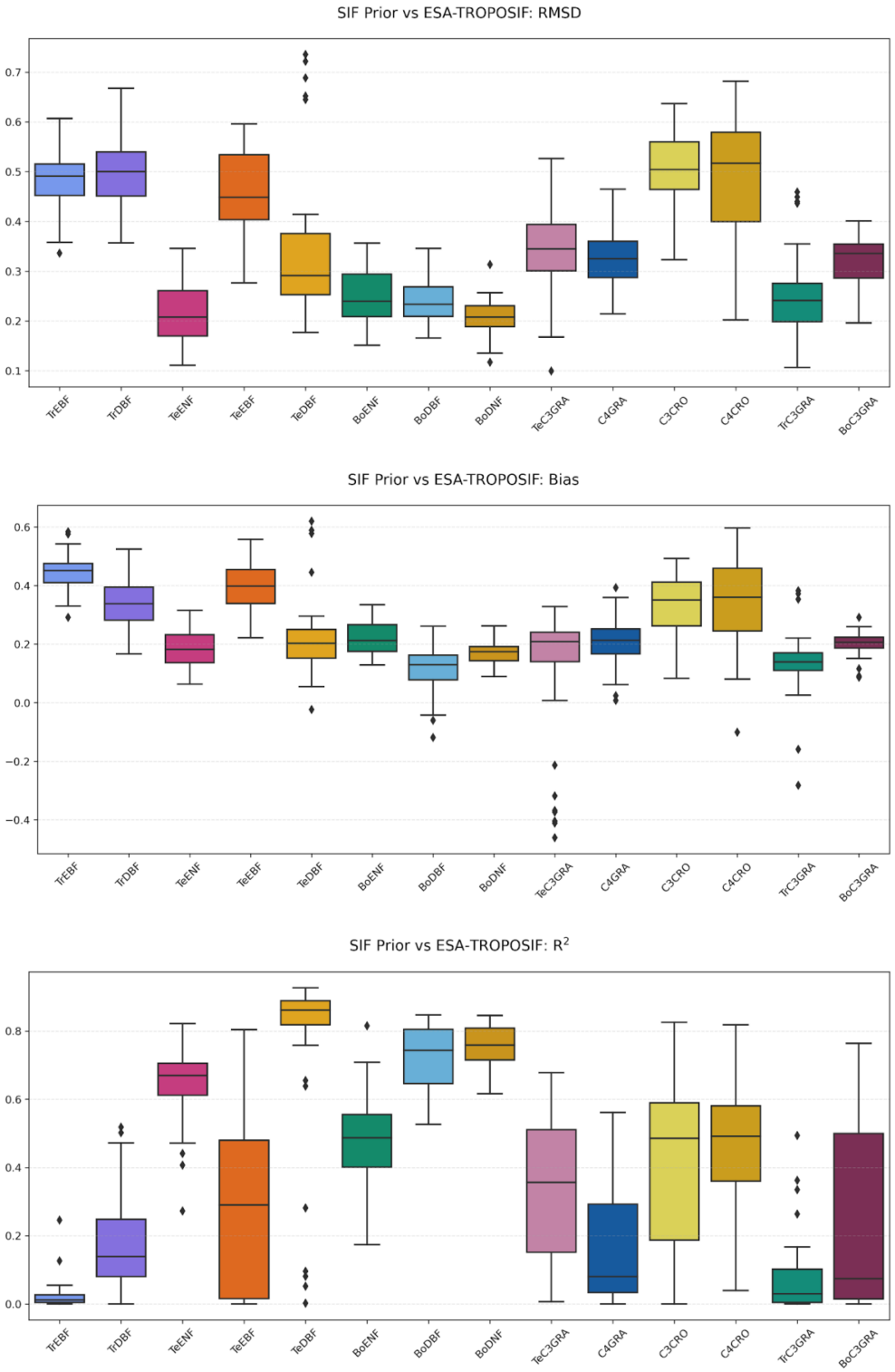

Figure 1: Boxplot of (top) Root Mean Squared Differences (RMSD), (middle) bias and (bottom) coefficient of determination (R^2), for prior SIF (in $mW.m^{-2}.sr^{-1}.nm^{-1}$) vs. TROPOSIF observations over the period 2019-2022 (weekly), over an ensemble of homogeneous pixels (0.1°) of (from left to right) tropical evergreen broadleaf forest, tropical deciduous broadleaf forest, temperate evergreen needleleaf forest, temperate evergreen broadleaf forest, temperate deciduous broadleaf forest, boreal evergreen needleleaf forest, boreal deciduous broadleaf forest, boreal deciduous needleleaf forest, temperate C3 grass, C4 grass, C3 crop, C4 crop, tropical C3 grass, boreal C3 grass.

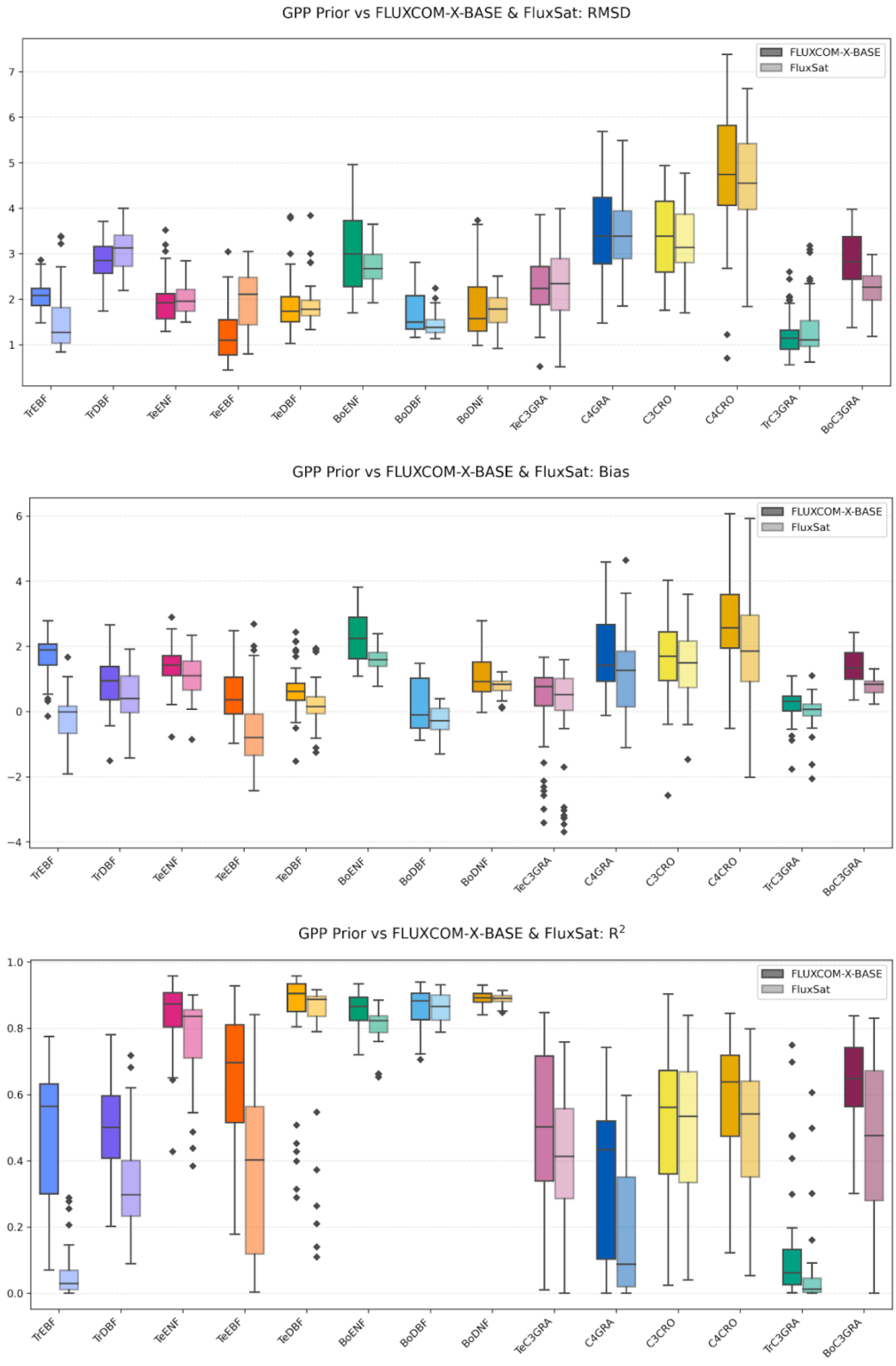

Figure 2: As in Figure 1, except for prior GPP (gC.m^{-2}.d^{-1}) vs FLUXCOM-X-BASE / FluxSat estimations over the period 2001-2021 (daily).

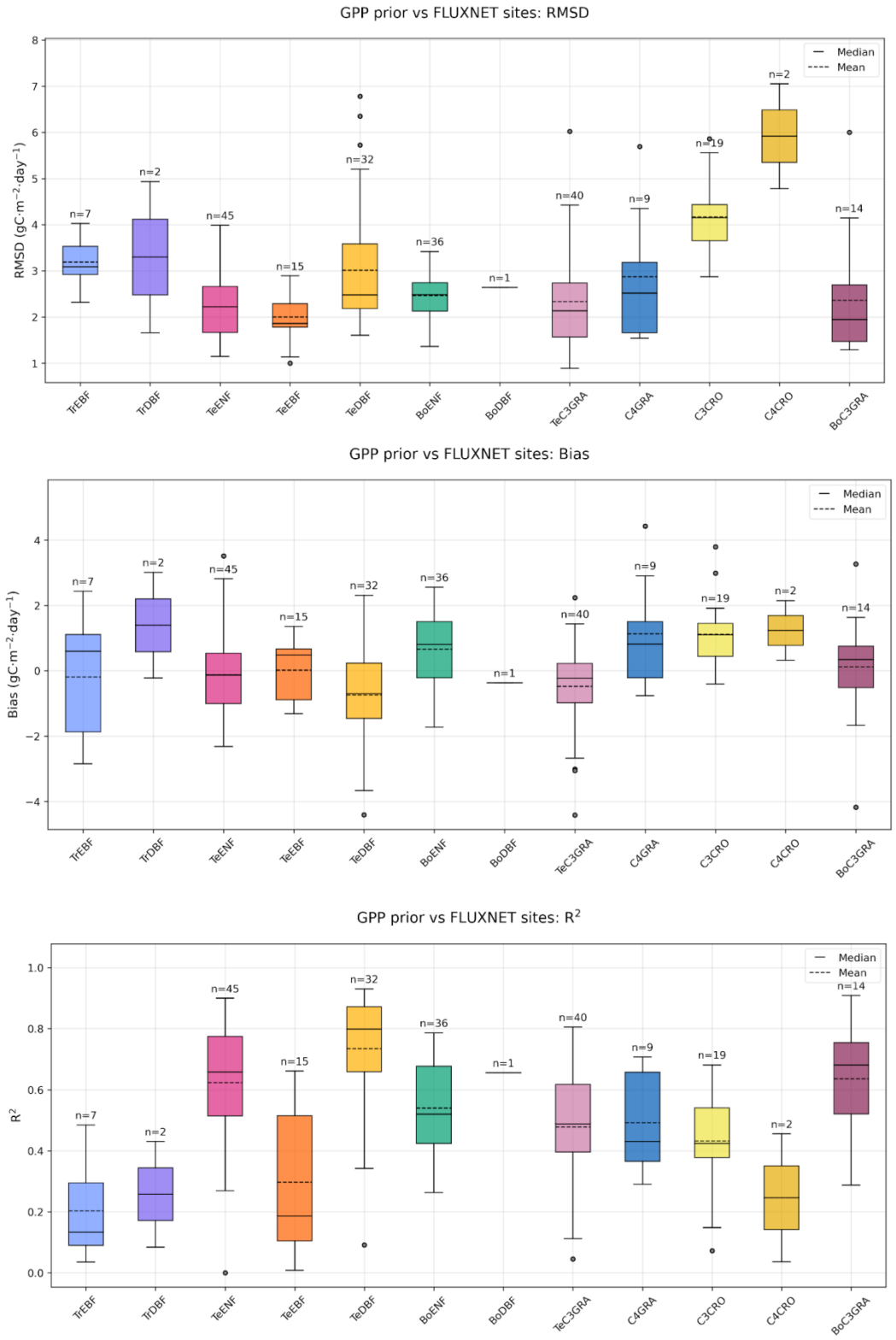

Figure 3: Same as in Figure 2, except for the evaluation of the simulated GPP at the daily resolution over eddy covariance sites.

5.2 ISBA modelling framework

From the land surface variables that can be simulated by the ISBA land surface model, machine learning was used to simulate SIF, ASCAT sigma0 and SMAP, SMOS and AMSR2 TB. The ISBA training configurations are summarized in Table 1.

Global maps of the RMSE and Pearson correlation (*R*) values are shown for cropland in Figures 4 and 5 for SIF and microwave data, respectively. Table 3 summarises the obtained mean scores (RMSE and *R*) and their standard deviation (SD). The relative standard deviation of the scores (RSD), the ratio of the SD to the mean, is also given.

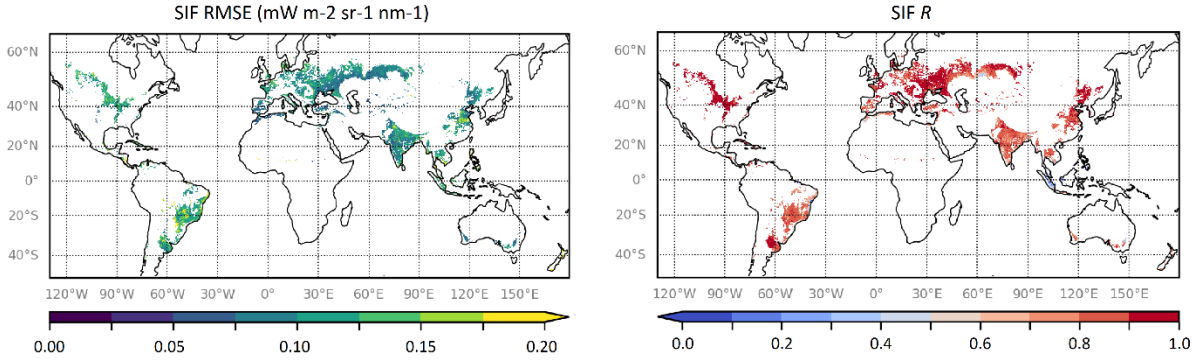

Figure 4: SImulated SIF scores over global cropland during testing period, from May 2018 to April 2019: (left) RMSE, (right) *R*.

Table 3: Assessment of ISBA global cropland forward operator during test periods for SIF and for all microwave data. Mean values of per grid-cell RMSE and *R* values are given together with the spatial standard deviation (SD) of the score values, the relative SD (RSD), and the mean number of observations used in the score calculation. The length of the test period is indicated. All microwave data are with V-polarisation. Small and large RSD and high and low *R* values are in bold, white and dark, respectively.

Instrument (observation)	RMSE			*R*			Mean number per grid-cell per month
	Mean	SD	RSD	Mean	SD	RSD	
TROPOMI (SIF)	0.11 (mw m^{-2} sr^{-1} nm^{-1})	0.03	**30** (%)	**0.83**	0.15	**18** (%)	19 (over 12 months)
ASCAT (C-band sigma0)	0.68 (dB)	0.29	43 (%)	**0.63**	0.17	27 (%)	19 (over 48 months)
SMAP (L-band TB)	9.7 (K)	7.3	**75** (%)	0.76	0.18	24 (%)	16 (over 12 months)
SMOS (L-band TB)	12.3 (K)	9.5	**77** (%)	**0.67**	0.28	**42** (%)	6 (over 48 months)
AMSR2 (X-band TB)	3.5 (K)	1.7	49 (%)	**0.92**	0.09	**10** (%)	34 (over 6 months)

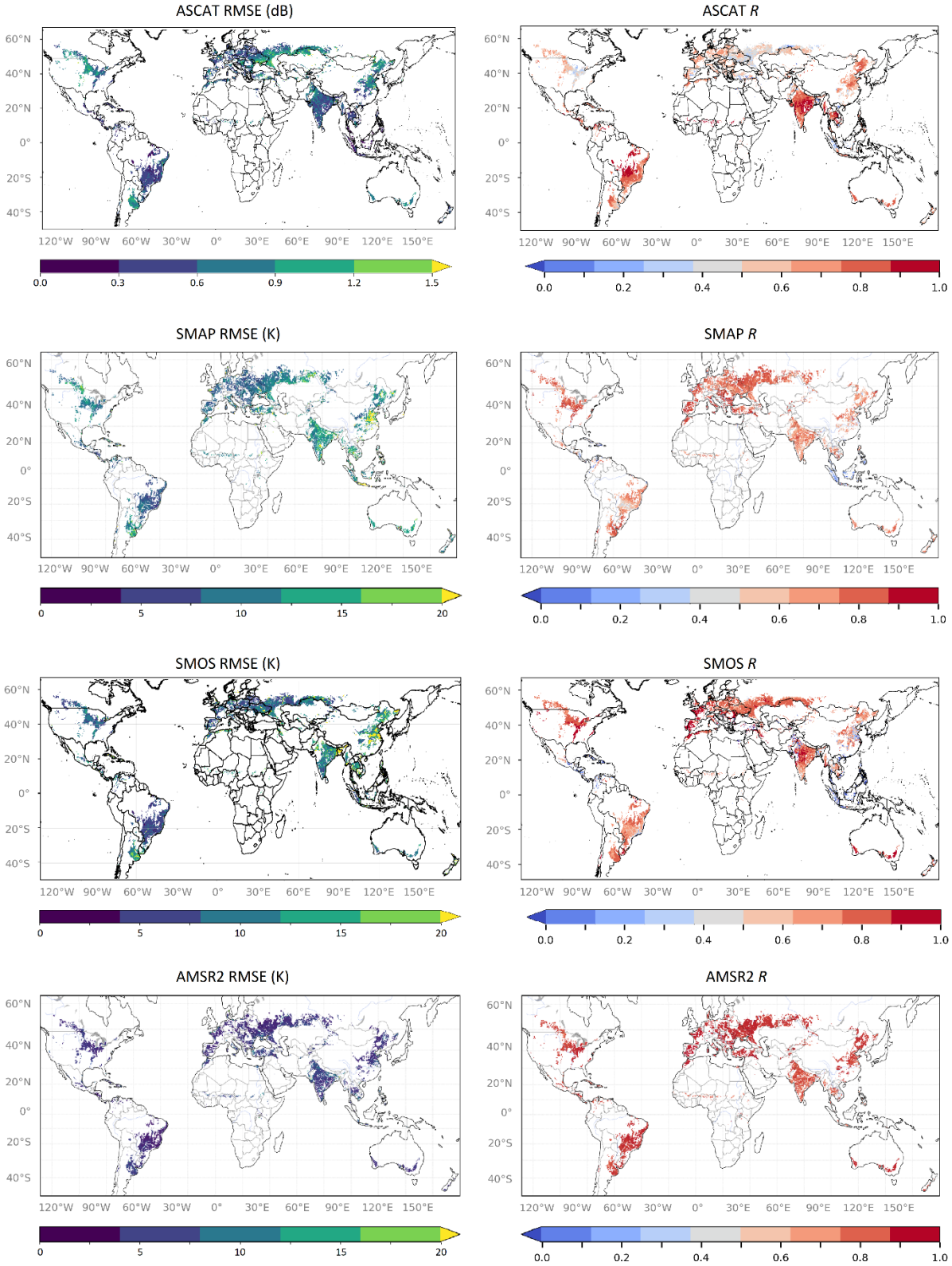

Figure 5: Simulated microwave data scores over global cropland during testing period: (left) RMSE, (right) *R*, (from top to bottom) ASCAT sigma0 from January 2016 to December 2019, SMAP TB from January to December 2021, SMOS TB from January 2016 to December 2019, and X-band AMSR2 TB from July to December 2020.

5.2.1 Observation operator for SIF

The TROPOMI SIF data used are daily and cover global cropland from May 2018 to April 2020. Data from May 2018 to April 2019 were used to train the neural network, and data from May 2019 to April 2020 were used to test the NN. No overfitting problem was found by comparing the losses between the training and validation datasets. The evaluation of the feature importance, which can be compared with the evaluation of the sensitivity to the inputs, shows that LAI and DOY (day of year) are the most important features. Figure 6 shows the 2D histogram of the comparison between the NN predictions and the TROPOMI SIF values from May 2018 to April 2019. The Pearson correlation reaches a value of 0.86, which is satisfactory, and an RMSE of 0.12 mW m^{-2} sr^{-1} nm^{-1}. However, it shows some limitations in accurately predicting high or very low TROPOMI SIF values. Figure 4 and Table 3 show that SIF score values are more evenly distributed around the globe compared to microwave instruments. The SIF has the highest mean *R* and low RSD values for both RMSE and *R*. The *R* values are more evenly distributed than the RMSE values, with RSDs of 18% and 30% respectively. This can be explained by the fact that the annual peaks of the SIF values can vary from one region to another and that the RMSE can be related to the annual peak.

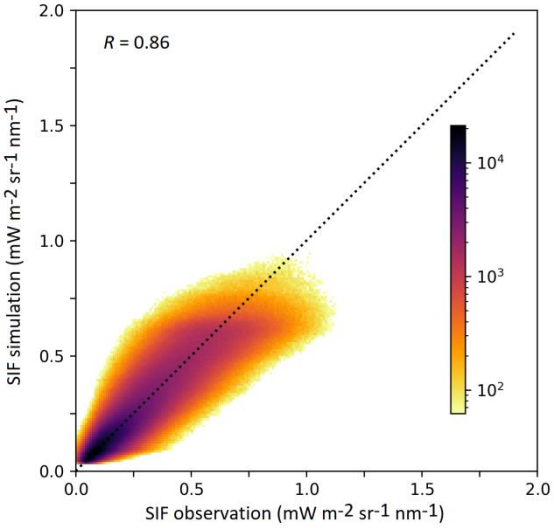

Figure 6: Predicted vs. observed TROPOMI SIF over global cropland.

5.2.2 Observation operator for ASCAT sigma0

Initial tests over grid cells in southwest France showed that simple NNs can predict ASCAT sigma0 with the RMSE of the simulated sigma0 often in the range of 0.3 to 0.4 dB, close to the mean ASCAT observation error of 0.33 dB (Corchia et al. 2023). In the global cropland NN configuration described in Table 1, Figure 5 and Table 3 show that both RMSE and *R* can vary considerably from one region to another. Correlations can reach 0.8 or more in South America, India, South West Asia, China and Australia. Lower values are obtained in Europe and the United States, with correlations around 0.6 in most of Western Europe and the western part of the United States. The lowest values are found in Eastern Europe and the eastern part of the United States, with correlations below 0.5. For all pooled data, from January 2016 to December 2019, *R* = 0.89 (Figure 7). The lowest RMSE values are found in Brazil, India and South East Asia with RMSEs around 0.5 dB and in some areas even lower with values closer to 0.25 dB. The higher RMSEs are found in Eastern Europe, with values around 1 dB in Eastern Ukraine and Russia. Overall, LAI is the most important predictor of sigma0, followed by surface soil moisture.

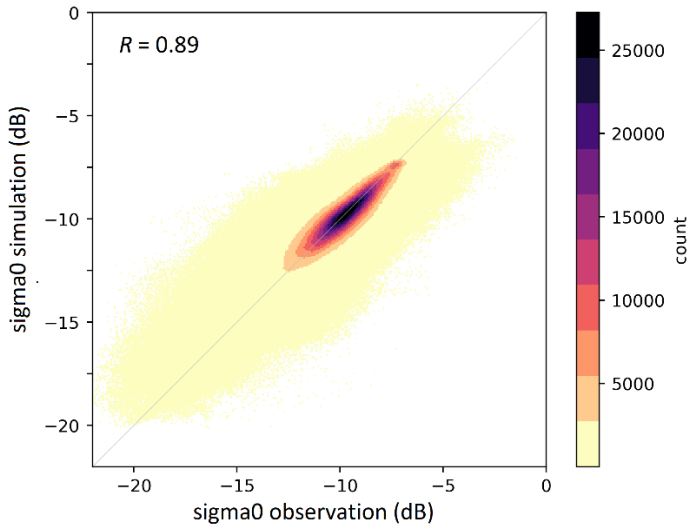

Figure 7: Predicted vs. observed ASCAT sigma0 over global cropland.

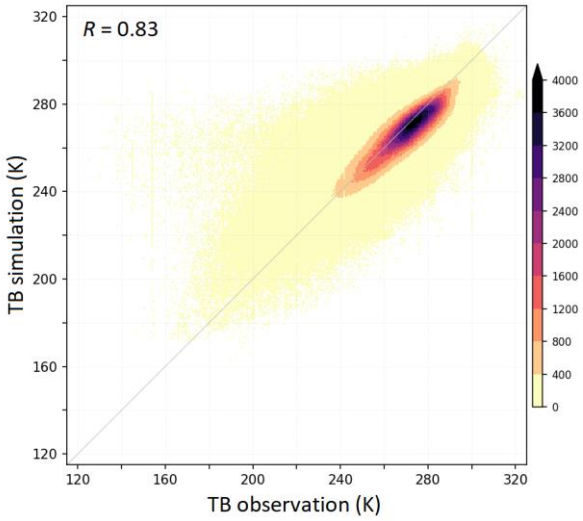

Figure 8: Predicted vs. observed SMAP TB (vertical polarization) over global cropland.

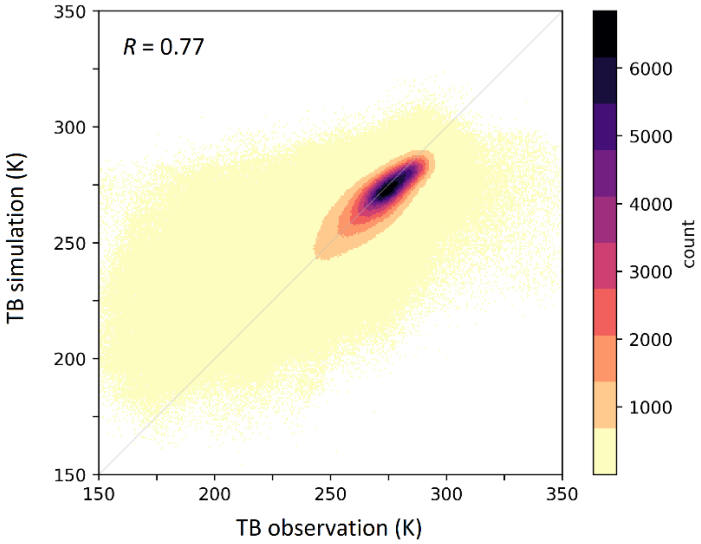

Figure 9: Predicted vs. observed SMOS TB (vertical polarization) over global cropland.

5.2.3 Observation operator for SMAP

We use the L1C product of SMAP, which contains gridded TBs for the two polarization H and V. After initial testing of the hyperparameter space, we found that the configuration given in Table 1 (among others) gives the best results. The exact architecture of the neural network itself did not turn out to be of great importance. In contrast, the preparation of the dataset was essential to obtain good skill metrics, especially the consistent masking of nonphysical and possibly RFI-contaminated pixels. In addition to applying the internal quality flags, we extracted data from the physiographic dataset (ECOCLIMAP-SG) used by the ISBA model to mask pixels with a high proportion of water, urban/coastal areas or frozen ground. In addition, we removed non-physical peaks at well-defined values in the histograms, as well as grid points with very high standard deviations. As the global model dataset provides 3-hourly data, the observations were also averaged over 3 hours. The NN was trained and validated with data from 2017 to 2020 and tested for the year 2021. In general, the performance for the H polarization was more sensitive to changes in the NN configuration than for the V polarization. Including H polarization in the training did not add significant value compared to training on V polarization only. For the sake of simplicity, the training results are described for V polarization only. Figure 8 shows the generally good correlation between predictions and observations from January to December 2021 (R = 0.83), but with a slight underestimation for high values and a large scatter for lower values. Figure 5 shows a low correlation around Indonesia and Central America, while the RMSE shows higher values in eastern China, where poor data coverage may be responsible, but also in other parts of the world. Figure 5 and Table 3 show that the SMAP TB RMSE values are not uniformly distributed around the globe. The RMSE values have an RSD of 75 %. In general, it can be said that the observation operator works sufficiently with different configurations of the neural network (Fig. 8), while the preparation of the dataset remains the main challenge.

5.2.4 Observation operator for SMOS

RFI filtering is a critical step before training the NN on the SMOS TB data. Data affected by the RFIs could introduce a bias during the training of the NN and degrade the quality of the predictions. To filter the affected data, we apply a method developed by the CATDS and described in detail by Mahmoodi et al. (2022). To perform the SMOS TB, we used the same NN architecture as for ASCAT sigma0 (Section 5.2.1). Figure 5 shows that the pixels with the highest correlation values are located in Western Europe with values above 0.9, in some regions of France, Spain and Portugal. High correlations can also be seen in the United States, with values above 0.8 in most of the pixels present in the country. In South America, lower correlations are found in south-eastern Brazil, where the values fall below 0.5 near the coasts. The lowest correlation values are found in Asia, especially in China and Indonesia, where correlations can drop below 0.4 in certain areas. The grid cells with the lowest RMSEs are located in Brazil with areas with values below 5 K. The highest RMSE values are located in Argentina, India and China with RMSE values above 20 K. Table 3 shows that, similar to SMAP, the SMOS TB RMSE values are not uniformly distributed around the globe. The RMSE values have an RSD of 77%. In addition, SMOS has a particularly large RSD of 42% for correlation. This is related to the greater difficulty in filtering out RFI than for SMAP. Figure 9 shows the fair correlation between predictions and observations (R = 0.77), from January 2016 to December 2019.

5.2.5 Observation operator for AMSR2

We used the settings found for the SMAP observation operator (Section 5.2.3) to train on X-band AMSR2 observations (10 GHz of the L2B product). The masking strategy developed for SMAP was adapted using the internal quality flags provided by the product and not flagging data with high variability. In addition, we worked with data from May 2018 to December 2020 and tested the trained NN from July to December 2020. The training performs well over most of the globe with slightly lower correlations in South East Asia and higher RMSE additionally around the Black Sea. We analysed the performance of the NN in the same way as described

in section 5.2.3. Figure 5 and Table 3 show that the AMSR2 TB *R* values are more uniformly distributed around the globe than for the other products. The *R* values have an RSD of 10%. Figure 10 illustrates the good overall correlation with a small overestimation for lower brightness temperatures.

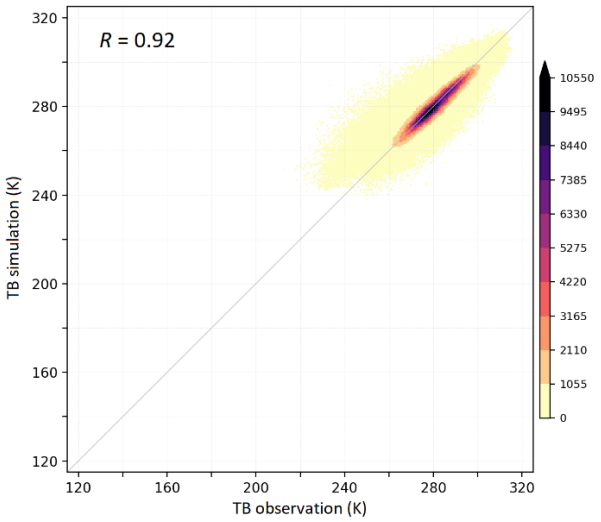

Figure 10: Predicted vs. observed X-band AMSR2 TB (vertical polarization) over global cropland.

5.3 ECLand modelling framework

5.3.1 AMSR2 training database information content analysis

Figure 11 shows the correlations between each model field with the polarization index (PI) in selected AMSR2 channels.

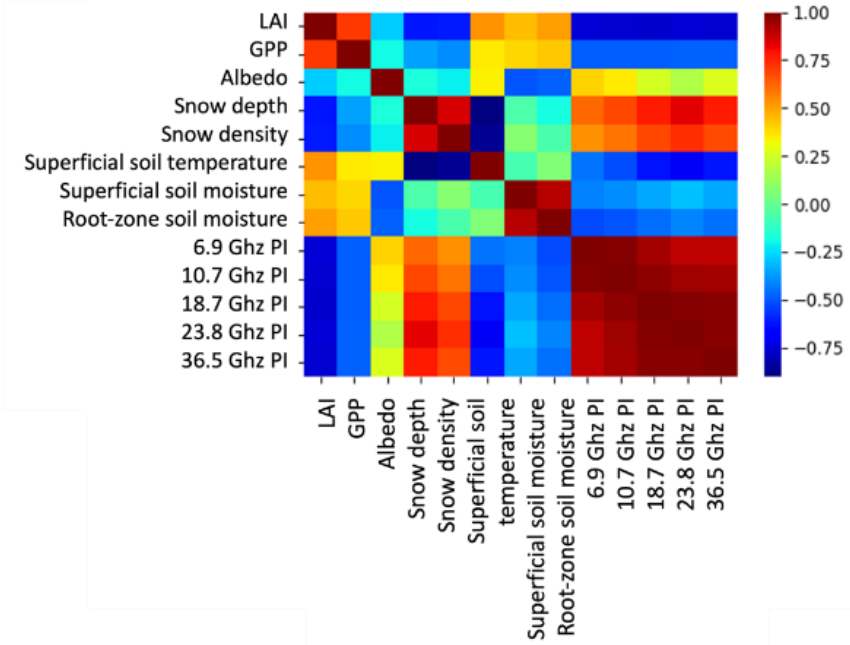

Figure 11: Correlation map of IFS model fields (vegetation, albedo, snow, soil temperature, soil moisture) with polarization index in selected AMSR2 bands.

PI is the ratio of the difference and the sum of the brightness temperature in vertical (V) and horizontal (H) polarization ($PI = \frac{V-H}{V+H}$). The correlations are negative with vegetation variables, positive with snow, negative with soil temperature and negative with soil moisture. The relationships between the model fields and the AMSR2 brightness temperature of the PI do not show strong dependency with microwave frequency.

5.3.2 ASCAT ML-based forward operator

The analysis of the information content of the ASCAT training database shows that the spatial distribution of the backscatter normalized at 40° is mainly driven by the vegetation spatial pattern while the temporal evolution of the signal relates to the temporal evolution of soil moisture. The most influent physical predictors of the satellite backscatter signal are LAI and soil moisture (Figure 12). The introduction of latitude and longitude substantially improves the performance of the ML model by providing local surface characteristics which are not resolved by the IFS model and were found to be essential to accurately predict the backscatter signal.

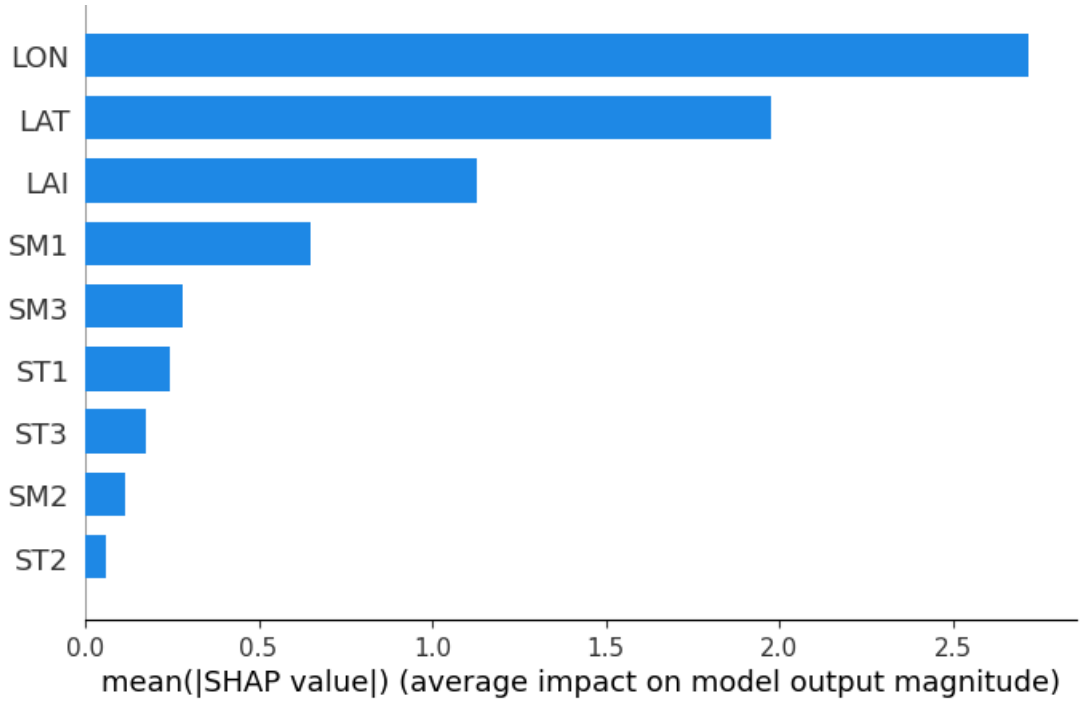

Figure 12: Feature importance for the ASCAT observation operator. SMx and STx are soil moisture and soil temperature IFS model fields of the x soil layer.

The comparison of XGB and NN showed that a NN with 4 hidden layers, 60 neurons provides the most accurate predictions of ASCAT backscatter at global scale. Figure 13(a,b) highlights the overall good performances of the ML model. The mean absolute error (MAE) of 0.78 dB obtained at global scale is within the expected error of the backscatter at 40° product. The spatial distribution of the backscatter and its pattern as a function of soil moisture and LAI are accurately reproduced by the NN (Figure 13(c)).

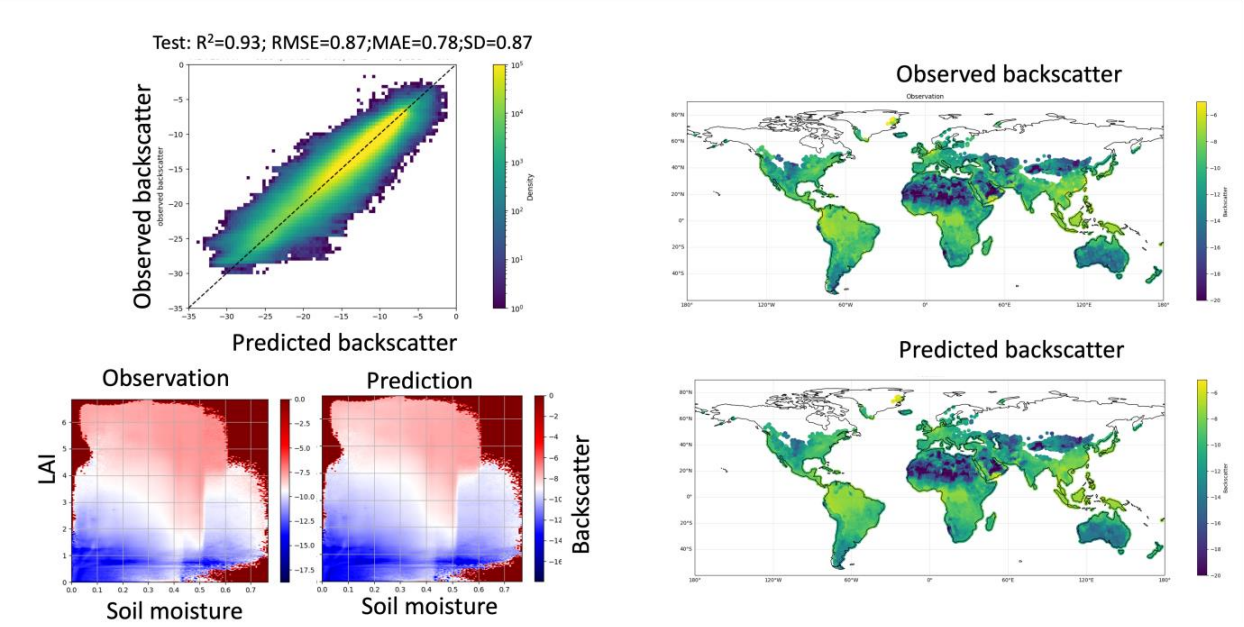

Figure 13: Evaluation of the ASCAT feedforward neural network for year 2019. a): Observation versus NN prediction scatterplot; b): Global maps of observed and predicted backscatter; c) Comparison of predicted and observed backscatter patterns as a function of modelled LAI and surface soil moisture.

Figure 14 shows high correlation and low RMSE over most regions. Lower performances are observed in arid regions (e.g. central Australia) and over part of tropical rainforests (Amazon, Central Africa).

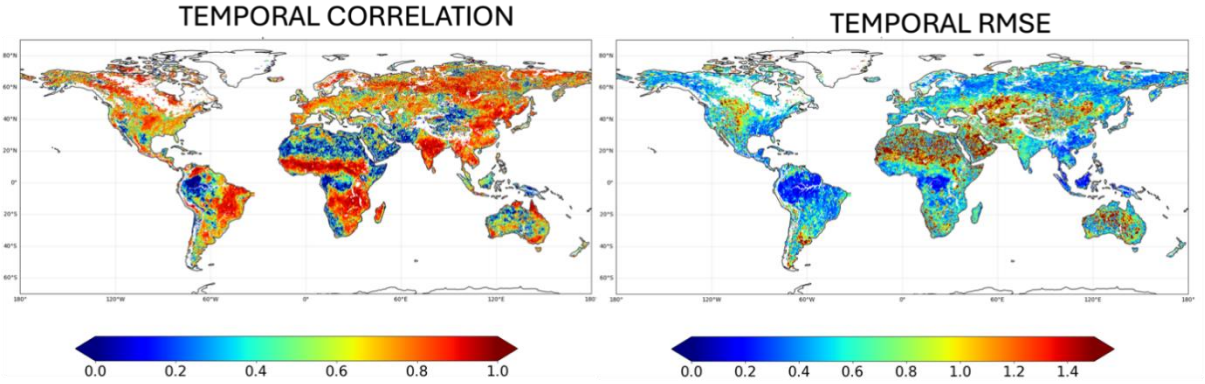

Figure 14: Temporal correlation and RMSE between observed and predicted ASCAT backscatter at 40° evaluated over 2019.

5.3.3 SIF ML-based forward operator

Figure 15 highlights that the most influent IFS predictors of SIF are the shortwave downwelling radiation and LAI. This is consistent with the process-based knowledge of the SIF drivers at canopy scale (van der Tol et al., 2009). The moisture and thermal characteristics of the surface layer (soil moisture, soil temperature), the water content available for the vegetation (root-zone soil moisture) and the thermal and moisture conditions of low-level atmosphere, which are key drivers of photosynthesis, bring essential information to predict SIF. GPP was not selected as

predictor since it does not significantly improve the performances of the ML model which confirms that the SIF signal at 0.1° spatial resolution scale is mainly driven by LAI for most vegetation types (Dechant et al., 2020).

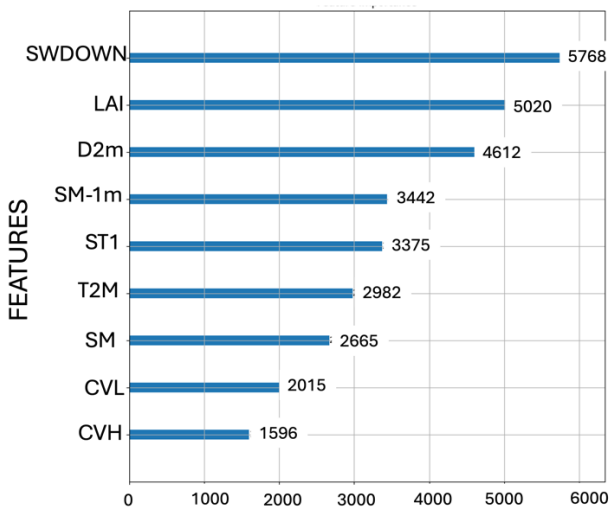

Figure 15: SIF feature importance diagram from model M1.

The ML models M1, M2, M3 show overall good performances at global scale (Figure 16). They underestimate high SIF satellite values and they saturate above 1.25 mW m^{-2} nm^{-1} sr^{-1}. Performances are, however, lower compared to the one derived for ASCAT backscatter, indicating larger uncertainties and lack of information content (e.g. leaf physiology) in the IFS model fields to accurately predict the SIF satellite signal at global scale. Figure 17 indicates lower correlation between predicted and observed SIF for tropical rainforest (Amazon, Central Africa) and semi-arid regions (Australia). The use of vegetation characteristics in M1 slightly increases the model performances over North America and Europe. Stronger correlation and lower RMSD are obtained for M3, which is trained on the satellite LAI.

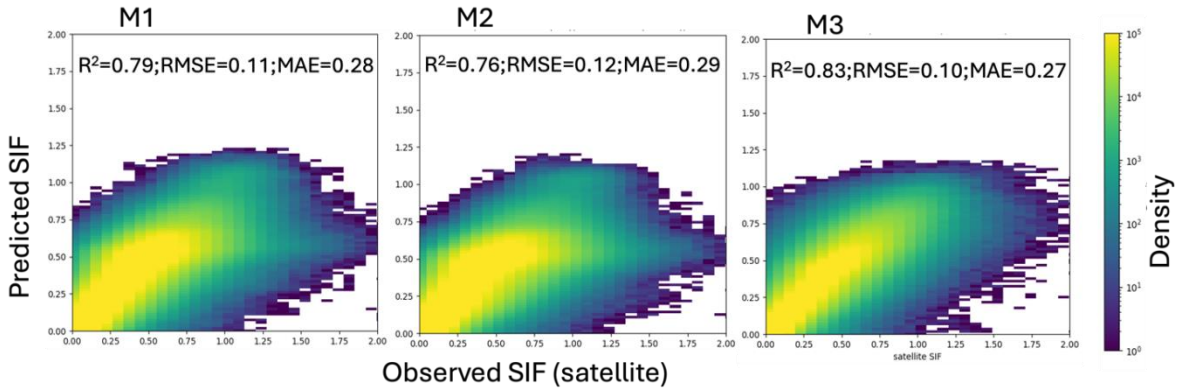

Figure 16: Scatterplots of predicted SIF versus SIF satellite observation (year 2022 using all samples in space and time, SIF unit: mW m^{-2} $nm^{-1}sr^{-1}$).

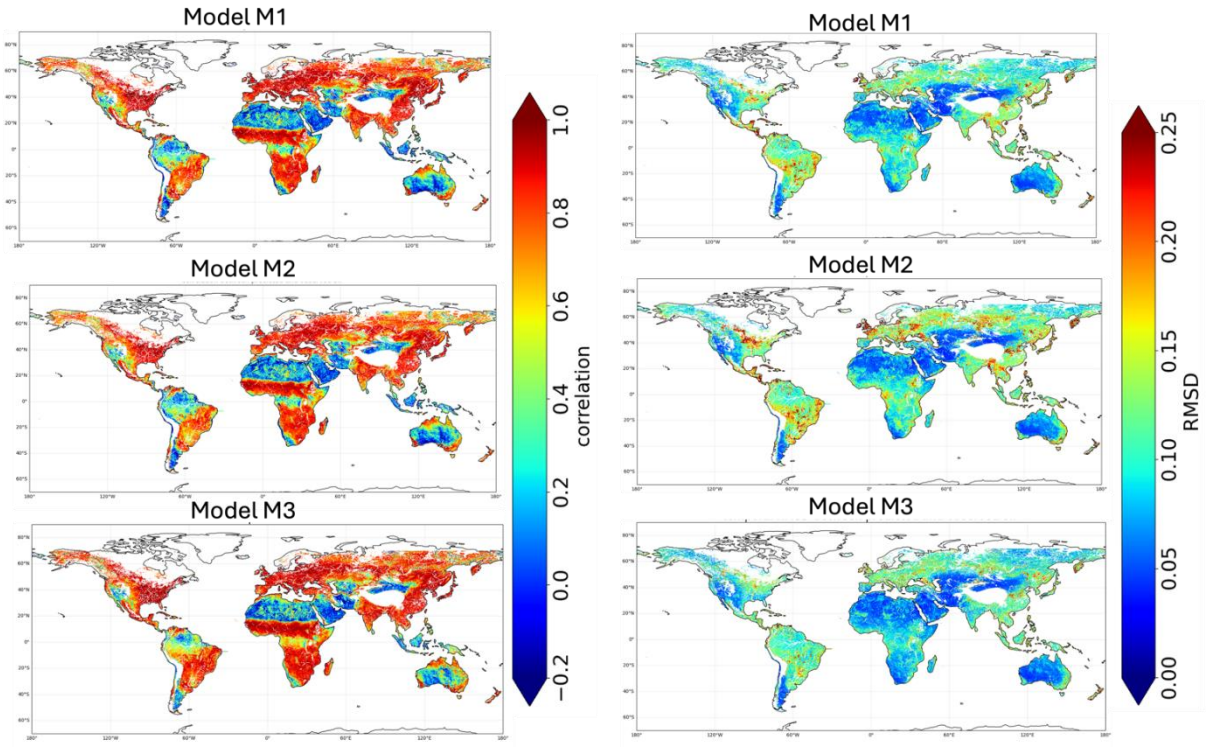

Figure 17: Temporal correlation (left column) and root mean square difference (RMSD, right column) maps between observed and predicted SIF for the models M1, M2, M3 for the year 2022.

All the models accurately reproduce the spatial distribution of SIF (Figure 18) and the seasonal evolution of SIF for different land cover types (Figure 19). The prediction is generally more accurate over middle-latitute cropland and grassland for which the LAI and the solar radiation are the main driver of SIF at the canopy scale. Noisier SIF observations and less accurate predictions are observed over evergreen broadleaf forest (tropical rainforest) which is related to (1) more frequent cloud contamination and (2) a higher sensitivity of SIF to the varability in light use efficiency which is not properly represented by the IFS predictors. The interannual variability of SIF can be poorly captured by the models M1 and M2 which are trained on the LAI climatology compared to M3 trained on satellite LAI. The lack of vegetation characteristics (fraction of low and high vegetation) in M2 can lead to underestimation of vegetation peaks as shown on the grassland and mixed forest sites.

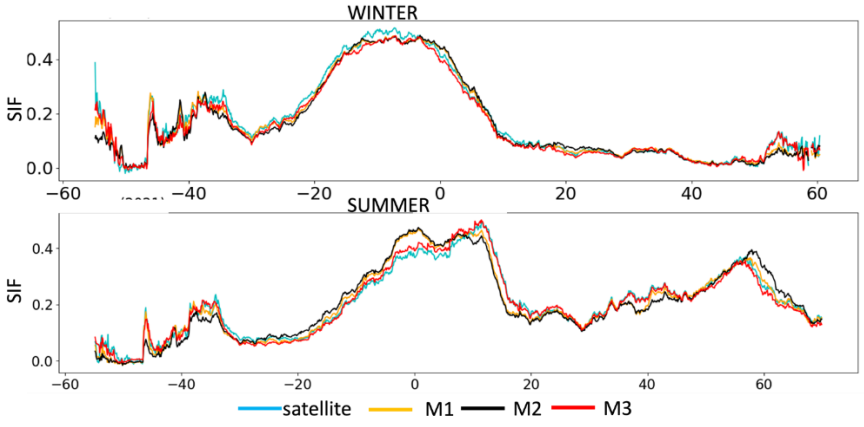

Figure 18: Latitude transects of observed SIF and multi-model predictions of SIF for winter and summer 2022.

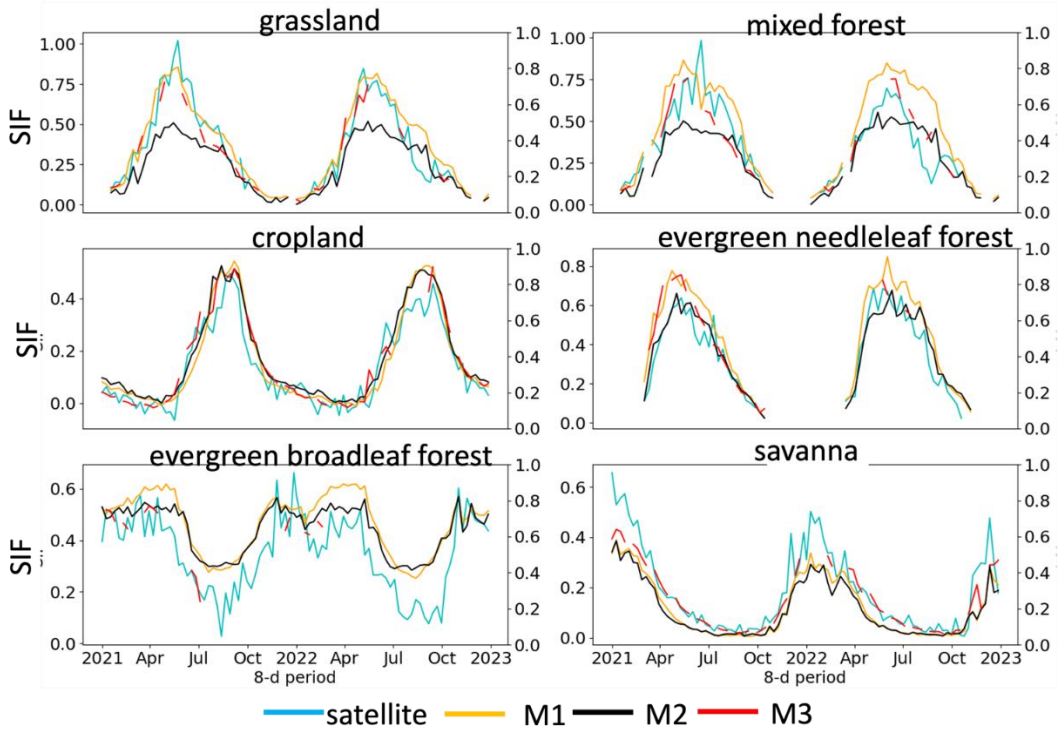

Figure 19: Seasonal evolution of observed SIF and multi-model predictions of SIF for selected sites for the 2021-2022 period.

While M3 provides the best agreement with the observed SIF, the latitude and longitude predictors dominate the feature importance diagram (not shown here) which may be a sign of overfitting and may mask the physical variability of LAI. The reduction of the model complexity by removing low-influential features such as the fractions of low and high vegetation does not significantly affect the performances of M2 compared to M1. The next step to select the ML model will be to implement M1, M2, M3 in land data assimilation experiments to evaluate the LAI increments produced by the data assimilation system and assess the impacts on low-level meteorological variables and carbon flux forecasts.

5.4 D&B modelling framework

5.4.1 Observation operator for SIF

Figure 20 shows a comparison of the simulated SIF against site level measurements for Sodankylä, a site in Northern Finland representative for boreal forests and dominated by evergreen pine trees. Simulated SIF values shown here are multiplied with a factor of 10, i.e. with the scaling factor sSIF in the observation operator set to 10. This reflects the high uncertainty regarding the absolute magnitude of the measured SIF. Observations are shown for two methods of retrieving SIF from the actual site-scale measurements using a FloX-Box, namely Frauenhofer line discrimination and spectral fitting.

The difference in magnitude between the modelled and observed SIF is likely due to the choice of prior parameters for the SIF model, taken from Gu et al. (2019), and the specific spectral conversion used. Although it has not been done here, there is scope within D&B to adjust these parameters in the assimilation. We believe, however, that it is more important, in the first instance, that we have a model that can track the seasonal and diurnal cycle of the observations. At the Sodankylä site, the simulations are able to track both the diurnal and

seasonal cycles of the observations reasonably well (Figure 20). A more rigorous evaluation of the D&B model performance and the SIF observation operator is given in Knorr et al. (2024).

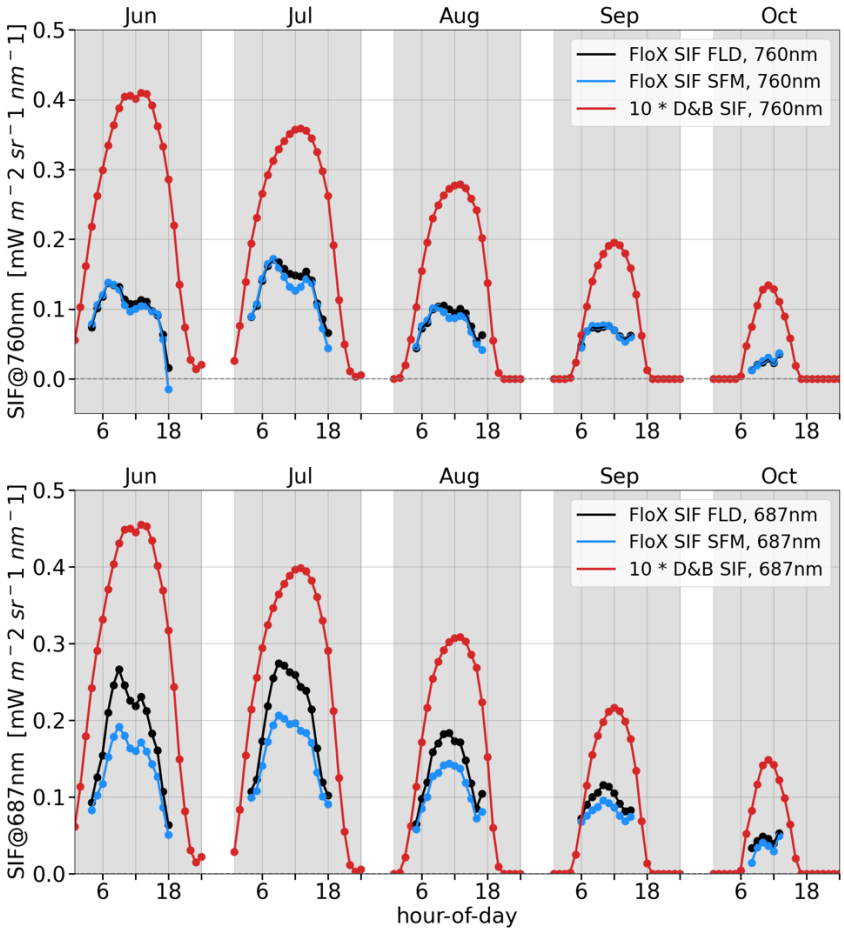

Figure 20: Average diurnal cycle by month of far-red (upper panel) and red SIF (lower panel) at Sodankylä for months June to October in 2021. D&B simulations (red) against measurements with the FloX-Box: retrievals made with the Frauenhofer line discrimination (black) and retrieval made with the spectral fitting method (blue).

5.4.2 Observation operator for L-VOD

Figure 21 shows a comparison between simulated and observed L-VOD for the Sodankylä site. Observations are made with an Elbara II radiometer at different elevation angles and only include the pine trees at the site (and no understory vegetation). Therefore simulated L-band VOD is for the tree plant functional type only. The temporal variations of the measurements are well captured by the simulation and match both temporal variations and magnitude of the locally measured L-VOD rather well suggesting a very satisfactory performance of the empirical L-VOD observation operator together with D&B. A more rigorous evaluation of the D&B model performance and the L-VOD observation operator is given in Knorr et al. (2024).

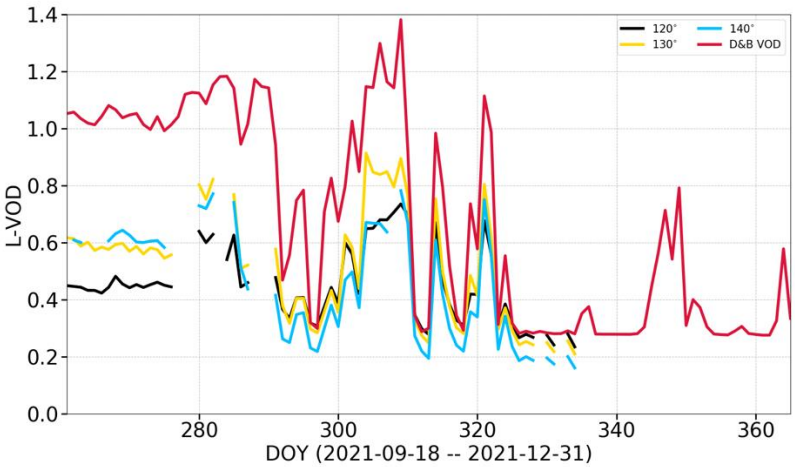

Figure 21: L-band VOD from Elbara II over a pine stand for different elevation angles compared to D&B simulated L-band VOD, for boreal evergreen trees only. Time axis starts on 18 September 2021.

6 Conclusion

In this work, observation operators have been constructed that will be used for the assimilation of radiance satellite observations: solar induced fluorescence (SIF), low frequency microwave brightness temperatures and backscatter coefficients.

Both neural networks and physically based observation operators were considered for SIF. The ORCHIDEE and D&B modelling frameworks focused on a physically based observation operator for SIF. With the uncalibrated version of ORCHIDEE, good correlation results of SIF simulations (R^2 > 0.6) are obtained for temperate and boreal forests, except for temperate evergreen broadleaf forests. On the other hand, simulations for crops and grasslands do not show good agreement with SIF observations. The model also tends to overestimate SIF, with median bias values greater than 0.1 mW m^{-2} sr^{-1} nm^{-1} for all vegetation types. Similar results are found for GPP. It is expected that the assimilation of SIF in ORCHIDEE will improve the model performance for both SIF and GPP. The uncalibrated D&B model tends to underestimate SIF over a site in Northern Finland, but is able to follow the seasonal and diurnal cycle of the observations.
As with ORCHIDEE, it is expected that assimilation of SIF in D&B will bring modelled SIF values in line with observations and, by propagating information through the model's process parameters, also improve model performance for GPP. ML-based observation operators for SIF in the ISBA LSM and in the IFS show overall good performance on a global scale, with a mean RMSE of about 0.1 mW m^{-2} sr^{-1} nm^{-1}. They show some limitations in accurately predicting high or very low SIF values (e.g. in the Amazon forest or in semi-arid areas).
The best results are obtained with a short list of predictors. In this approach, LAI is the most important biophysical predictor of SIF. Simple structural predictors such as latitude, longitude and DOY are sufficient to represent the solar radiation driving conditions for SIF.
LAI can vary rapidly in vegetation growth and senescence conditions and is strongly related to GPP and factors influencing GPP such as drought. The observed relationship between LAI and SIF is good news because it means that SIF observations are another source of information for analysing LAI and soil moisture.

Neural networks were used for the microwave observations. ML-based observation operators give good results for ASCAT sigma0 and SMAP, SMOS and AMSR2 TB. The RMSE for SMAP and SMOS L-band sensors and the Pearson correlation for SMOS are less uniformly distributed over the globe than for ASCAT and AMSR2. This could be related to the difficulty of filtering out RFI in some regions. LAI is an important predictor for microwave observations, but surface soil moisture and surface temperature are more important than LAI for L-band TB.

Although different modelling frameworks were used (IFS coupled model and ISBA off-line surface model), similar results were obtained in ML-based training of observational operators. It is shown that it is important to identify a parsimonious set of predictors that ensures a sufficiently accurate prediction of the observations while providing sufficient sensitivity to the analysed variables in the data assimilation system. It is useful to use (1) latitude and longitude as localisation variables to compensate for the lack of information on static local surface conditions, (2) LAI satellite observations, in the training database. These works illustrate the potential of ML to implement the assimilation of new observations.

7 References

Bacour, C., MacBean, N., Chevallier, F., Léonard, S., Koffi, E. N. and Peylin, P.: Assimilation of multiple datasets results in large differences in regional- to global-scale NEE and GPP budgets simulated by a terrestrial biosphere model, Biogeosciences, 20(6), 1089–1111, https://doi.org/10.5194/BG-20-1089-2023, 2023.

Bacour, C., Maignan, F., MacBean, N., Porcar-Castell, A., Flexas, J., Frankenberg, C., Peylin, P., Chevallier, F., Vuichard, N. and Bastrikov, V.: Improving Estimates of Gross Primary Productivity by Assimilating Solar-Induced Fluorescence Satellite Retrievals in a Terrestrial Biosphere Model Using a Process-Based SIF Model, J. Geophys. Res. Biogeosciences, 124(11), 3281–3306, https://doi.org/10.1029/2019JG005040, 2019.

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. and Wofsy, S.: FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities, Bull. Am. Meteorol. Soc., 82(11), 2415–2434, https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2, 2001.

Bastrikov, V., Macbean, N., Bacour, C., Santaren, D., Kuppel, S. and Peylin, P.: Land surface model parameter optimisation using in situ flux data: Comparison of gradient-based versus random search algorithms (a case study using ORCHIDEE v1.9.5.2), Geosci. Model Dev., 11(12), 4739–4754, https://doi.org/10.5194/gmd-11-4739-2018, 2018.

Chen , T., C. Guestrin, "XGBoost: A Scalable Tree Boosting System." In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA, ACM, 785–794. (KDD '16), https://doi.org/10.1145/2939672.2939785, 2016.

Corchia, T., Bonan, B., Rodríguez-Fernández, N., Colas, G., and Calvet, J.-C.: Assimilation of ASCAT Radar Backscatter Coefficients over Southwestern France. Remote Sens. 2023, 15, 4258. https://doi.org/10.3390/rs15174258

Dechant, B., Ryu, Y., Badgley, G., Zeng, Y., Berry, J.A., Zhang, Y., Goulas, Y., Li, Z., Zhang, Q., Kang, M., Li, J., Moya, I.: Canopy structure explains the relationship between photosynthesis and sun-induced chlorophyll fluorescence in crops. Remote Sens. Environ. 241, 111733. https://doi.org/10.1016/j.rse.2020.111733, 2020.

Frankenberg, C., Fisher, J. B., Worden, J., Badgley, G., Saatchi, S. S., Lee, J. E., Toon, G. C., Butz, A., Jung, M., Kuze, A. and Yokota, T.: New global observations of the terrestrial carbon cycle from GOSAT: Patterns of plant fluorescence with gross primary productivity, Geophys. Res. Lett., 38(17), 1–6, https://doi.org/10.1029/2011GL048738, 2011.

Goldberg, D. E.: Genetic algorithms in search, optimization, and machine learning, Addison-Wesley Longman Publishing Co., Inc., MA, United States., 1989.

Gu, L., Han, J., Wood, J. D., Chang, C. Y.-Y., and Sun, Y.: Sun-induced Chl fluorescence and its importance for biophysical modeling of photosynthesis based on light reactions, New Phytologist, 223, 1179–1191, https://doi.org/10.1111/nph.15796, 2019.

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 Syst. Sci. Data, 13(11), 5423–5440, https://doi.org/10.5194/essd-13-5423-2021, 2021.

Joiner, J., Yoshida, Y., Vasilkov, A. P., Schaefer, K., Jung, M., Guanter, L., Zhang, Y., Garrity, S., Middleton, E. M., Huemmrich, K. F., Gu, L. and Belelli Marchesini, L.: The seasonal cycle of satellite chlorophyll fluorescence observations and its relationship to vegetation phenology

and ecosystem atmosphere carbon exchange, Remote Sens. Environ., 152, 375–391, https://doi.org/10.1016/j.rse.2014.06.022, 2014.

Knorr, W.: Annual and interannual CO_2 exchanges of the terrestrial biosphere: process-based simulations and uncertainties, Glob. Ecol. Biogeogr., 9, 225–252, https://doi.org/10.1046/j.1365-2699.2000.00159.x, 2000.

Knorr, W., Williams, M., Thum, T., Kaminski, T., Voßbeck, M., Scholze, M., Quaife, T., Smallmann, L., Steele-Dunne, S., Vreugdenhil, M., Green, T., Zähle, S., Aurela, M., Bouvet, A., Bueechi, E., Dorigo, W., El-Madany, T., Migliavacca, M., Honkanen, M., Kerr, Y., Kontu, A., Lemmetyinen, J., Lindqvist, H., Mialon, A., Miinalainen, T., Pique, G., Ojasalo, A., Quegan, S., Rayner, P., Reyes-Muñoz, P., Rodríguez-Fernández, N., Schwank, M., Verrelst, J., Zhu, S., Schüttemeyer, D., and Drusch, M.: A comprehensive land surface vegetation model for multi-stream data assimilation, D&B v1.0, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-1534, 2024.

Konings, A. G., Rao, K., and Steele-Dunne, S. C.: Macro to micro: microwave remote sensing of plant water con- tent for physiology and ecology, New Phytologist, 223, 1166–1172, https://doi.org/10.1111/nph.15808, 2019.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S. and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biogeochem. Cycles, 19(1), 1–33, https://doi.org/10.1145/2939672.293978510.1029/2003GB002199, 2005.

MacBean, N., Bacour, C., Raoult, N., Bastrikov, V., Koffi, E. N., Kuppel, S., Maignan, F., Ottlé, C., Peaucelle, M., Santaren, D. and Peylin, P.: Quantifying and Reducing Uncertainty in Global Carbon Cycle Predictions: Lessons and Perspectives From 15 Years of Data Assimilation Studies With the ORCHIDEE Terrestrial Biosphere Model, Global Biogeochem. Cycles, 36(7), e2021GB007177, https://doi.org/10.1029/2021GB007177, 2022.

Magney, T. S., Frankenberg, C., Köhler, P., North, G., Davis, T. S., Dold, C., Dutta, D., Fisher, J. B., Grossmann, K., Harring- ton, A., Hatfield, J., Stutz, J., Sun, Y., and Porcar-Castell, A.: Disentangling changes in the spectral shape of chlorophyll fluorescence: implications for remote sensing of photosynthesis, J. Geophys. Res.: Biogeosciences, 124, 1491–1507, https://doi.org/https://doi.org/10.1029/2019JG005029, 2019.

Nelson, J. A., Walther, S., Gans, F., Kraft, B., Weber, U., Novick, K., Buchmann, N., Migliavacca, M., Wohlfahrt, G., Šigut, L., Ibrom, A., Papale, D., Göckede, M., Duveiller, G., Knohl, A., Hörtnagl, L., Scott, R. L., Zhang, W., Hamdi, Z. M., Reichstein, M., Aranda-Barranco, S., Ardö, J., Op de Beeck, M., Billesbach, D., Bowling, D., Bracho, R., Brümmer, C., Camps-Valls, G., Chen, S., Cleverly, J. R., Desai, A., Dong, G., El-Madany, T. S., Euskirchen, E. S., Feigenwinter, I., Galvagno, M., Gerosa, G. A., Gielen, B., Goded, I., Goslee, S., Gough, C. M., Heinesch, B., Ichii, K., Jackowicz-Korczynski, M. A., Klosterhalfen, A., Knox, S., Kobayashi, H., Kohonen, K.-M., Korkiakoski, M., Mammarella, I., Gharun, M., Marzuoli, R., Matamala, R., Metzger, S., Montagnani, L., Nicolini, G., O'Halloran, T., Ourcival, J.-M., Peichl, M., Pendall, E., Ruiz Reverter, B., Roland, M., Sabbatini, S., Sachs, T., Schmidt, M., Schwalm, C. R., Shekhar, A., Silberstein, R., Silveira, M. L., Spano, D., Tagesson, T., Tramontana, G., Trotta, C., Turco, F., Vesala, T., Vincke, C., Vitale, D., Vivoni, E. R., Wang, Y., Woodgate, W., Yepez, E. A., Zhang, J., Zona, D., and Jung, M.: 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, 2024.

Mahmoodi. A. et al.: SMOS Level 3 Brightness Temperature (TB) Users' manual and useful tips Quality assessment flags, https://www.catds.fr/content/download/170793/file/SO-TN-CB-GS-0097_v2_SMOS_BTs_filtering.pdf, 2022.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y. W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van Ingen, C.,

Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J. M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Sci. data, 7(1), 225, https://doi.org/10.1145/2939672.293978510.1038/s41597-020-0534-3, 2020.

Pedrós, R., Goulas, Y., Jacquemoud, S., Louis, J., and Moya, I.: FluorMODleaf: A new leaf fluorescence emission model based on the PROSPECT model. Remote Sens. Environ., 114(1), 155-167, https://doi.org/10.1016/j.rse.2009.08.019, 2010.

Quaife, T. L.: A two stream radiative transfer model for vertically inhomogeneous vegetation canopies including internal emission Journal of Advances in Modeling Earth Systems, under review, 2024.

Schwank, M., Kontu, A., Mialon, A., Naderpour, R., Houtz, D., Lemmetyinen, J., Rautiainen, K., Li, Q., Richaume, P., Kerr, Y., and Mätzler, C.: Temperature effects on L-band vegetation optical depth of a boreal forest, Remote Sensing of Environment, 263, 112 542, https://doi.org/https://doi.org/10.1016/j.rse.2021.112542, 2021.

van der Tol, C., Verhoef, W., Timmermans, J., Verhoef, a. and Su, Z.: An integrated model of soil-canopy spectral radiances, photosynthesis, fluorescence, temperature and energy balance, Biogeosciences, 6(12), 3109–3129, https://doi.org/10.5194/bg-6-3109-2009, 2009.

Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., and Kurpius, M. R.: An improved analysis of forest carbon dynamics using data assimilation, Global Change Biology, 11, 89–105, https://doi.org/10.1111/j.1365-2486.2004.00891.x, 2005.

Yang, P., Verhoef, W., and van der Tol, C.: The mSCOPE model: A simple adaptation to the SCOPE model to describe reflectance, fluorescence and photosynthesis of vertically heterogeneous canopies. Remote Sens. Environ., 201, 1-11, https://doi.org/10.1016/j.rse.2017.08.029, 2017.

Zhang, Y., Bastos, A., Maignan, F., Goll, D., Boucher, O., Li, L., Cescatti, A., Vuichard, N., Chen, X., Ammann, C., Arain, M. A., Black, T. A., Chojnicki, B., Kato, T., Mammarella, I., Montagnani, L., Roupsard, O., Sanz, M. J., Siebicke, L., Urbaniak, M., Vaccari, F. P., Wohlfahrt, G., Woodgate, W., and Ciais, P.: Modeling the impacts of diffuse light fraction on photosynthesis in ORCHIDEE (v5453) land surface model, Geosci. Model Dev., 13, 5401–5423, https://doi.org/10.5194/gmd-13-5401-2020, 2020.

Document History

Version	Author(s)	Date	Changes
0.1	Jean-Christophe Calvet	26.09.2024	
0.1	Cédric Bacour, Bertrand Bonan, Timothée Corchia, Sébastien Garrigues, Thomas Kaminski, Wolfgang Knorr, Fabienne Maignan, Philippe Peylin, Patricia de Rosnay, Marko Scholze, Vincent Tartaglione, Pierre Vanderbecken, Michael Voßbeck, Jasmin Vural	15.11.2024	Inputs from WP4 partners
0.2	Jean-Christophe Calvet	03.12.2024	Consolidated version
1.0			First issued version

Internal Review History

Internal Reviewers	Date	Comments
Richard Engelen ECMWF	16 Dec 2024	In text comments