CO2MVS RESEARCH ON SUPPLEMENTARY OBSERVATIONS

D4.1: first review and improvement of land forward operator for SIF and MW data

Due date of deliverable	December 2023
Submission date	January 2024
File Name	CORSO-D4.1-V1.4
Work Package /Task	WP4
Organisation Responsible of Deliverable	Meteo-France
Author name(s)	Jean-Christophe Calvet, Cédric Bacour, Vladislav Bastrikov, Bertrand Bonan, Timothée Corchia, Sébastien Garrigues, Fabienne Maignan, Philippe Peylin, Patricia de Rosnay, Pierre Vanderbecken, Jasmin Vural
Revision number	V1.4
Status	Issued
Dissemination Level / location	PUBLIC www.corso-project.eu

The CORSO project (grant agreement No 101082194) is funded by the European Union.

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. Three land surface models are used to provide predictors for training the observation operators: ISBA, ECLand and ORCHIDEE (MF, ECMWF and CEA respectively). This report is an intermediate document presenting preliminary results. Machine learning was used to simulate ASCAT backscatter coefficients and SIF. One of the issues concerns the optimal temporal frequency of 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. The feasibility of applying the methodology established for ASCAT to passive microwave data (SMAP, SMOS, AMSR-2) still needs to be demonstrated. For SIF, a process-based description of leaf fluorescence and its integration at canopy level, taking into account canopy structure, was used in the ORCHIDEE 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 CO2 Human Emissions), and CoCO2 (Copernicus CO2 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 first 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.

A consolidated version of this document (D4.2 - Final review and improvement of land surface forward operators for SIF and low frequency MW data) will be issued in December 2024.

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 SMAP. SMOS L-band observations will be considered at a later stage. 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, preliminary results are presented for each model. A comparison between the several approaches will be presented in the consolidated version of this document (D4.2).

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 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 (https://s5p-troposif.noveltis.fr/data-access/) 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).

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.

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 main parameters of ORCHIDEE related to photosynthesis and phenology were calibrated using a co-assimilation of space-borne estimates of SIF from Sentinel-5p observations and *in situ* Gross Primary Productivity (GPP) data. The observation operator for SIF followed a process-based description of the leaf fluorescence and its integration at canopy level accounting for the canopy structure (see initial description in Bacour et al. 2019). The optimized parameters were then applied to perform a simulation of GPP at a global scale, which were compared to those obtained with the standard parameter values and to a reference GPP product (FLUXSAT, Joiner et al. 2018). The differences between the prior and optimized simulations, and with the FLUXSAT 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).

4.1.2 Data assimilation approach

We employed 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 used a Genetic Algorithm (Goldberg, 1989), to iteratively minimize the misfit function (Bastrikov et al., 2018).

The assimilations were conducted on a PFT-basis, against GPP data (site scale estimates or FLUXSAT data for three PFTs for which no *in situ* data are available) and TROPOMI SIF retrievals for a collection of selected sites and grid cells. The co-assimilation of these two variables helps prevent parameter overfitting. Two assimilation experiments are conducted depending on the cloud fraction threshold to select the SIF observations (see below).

We used the daily averaged SIF retrievals of the TROPOSIF product (Guanter et al., 2021), over the period 2018-2020. Only observations passing the quality flag and associated with view zenith angles below 40° and two cloud fraction thresholds (below 0.2 and 0.5) were considered. The data were binned at 8-day/0.25° resolution. We selected fifteen grid cells for each of the 14 vegetation PFTs, with the highest thematic homogeneity and ensuring a correct sampling of the global distribution. For most PFTs, we assimilated daily *in situ* GPP estimates from FLUXNET (Baldocchi et al., 2001; Pastorello et al., 2020), while we used FLUXSAT-GPP (Joiner et al., 2018) data for three PFTs (TrDBF, BoDNF, TrC3GRA) without *in situ* GPP estimates. 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.

4.1.3 Justification of the use of weekly means for SIF

We used 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 ISBA modelling framework

MF worked on SIF, ASCAT and SMAP. At this stage, tests were made over southwest France and over the European CAMS domain, before going global in a next stage.

ASCAT data are assimilated 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 the ASCAT backscatter signal. The locally trained NN-based observation operators (one per grid cell) are implemented in LDAS-Monde, which allows the sequential assimilation of backscatter observations (Corchia et al. 2023).

As far as SIF is concerned, before working at a global scale, MF re-gridded the daily TROPOSIF product over the CAMS European domain on a regular grid at a spatial resolution of 0.1 x 0.1°. First tests were made to simulate the daily SIF product using a machine-learning method similar to the one used for ASCAT.

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 model, such as soil moisture and soil temperature, can be used to train neural networks (NNs) able to simulate satellite observations such as radiances, 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 are used during the NN training phase rather than modelled LAI. NN observation operators for radiances, 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.

4.3 ECLand modelling framework

The work of ECMWF was dedicated to the design of machine learning-based observation operators to assimilate passive multi-frequency microwave data (AMSR-2), active microwave data (ASCAT backscatter) and SIF in the IFS.

4.3.1 Design of a new training database for land processes

While existing training databases were used for ASCAT and AMSR-2, a new training database for land surface processes was developed for the CORSO project. The model fields were derived from 13 years of ECLand offline simulations (2010-2022) at a resolution of 25km and 1-h time step, using ERA-5 climate forcing. The observations include SIF (Caltech and Troposif), GPP (fluxcom and VODCA2GPP) and LAI (CGLS) satellite-based variables. Integration of microwave observations is in progress and will be completed in the next months. Zarr and Dask technologies are exploited to ensure efficient data access and archiving of the data. The training database relies on accurate collocation between the model fields and the satellite observations in the observation space. The designed framework is generic and reproducible to facilitate the update of the training database with new model or observation versions. This new database will be used to design the SIF and the microwave level-1 forward operators.

4.3.2 AMSR-2 information content analysis

An existing AMSR-2 training database at ECMWF (credit: Alan Geer, ECMWF), which is shared with the CERISE project, was produced 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 AMSR-2 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).

4.3.3 ASCAT

A four-year training database (credit: Aires et al., 2021), which relates ASCAT backscatter at 40° to ERA-5 model variables (LAI, soil moisture and soil temperature in first 3 layers, soil type and vegetation type), was used. Several architectures of feedforward neural network (NN) and gradient boosted trees (xgboost package) 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. The current work is dedicated to the implementation of the NN in the IFS which requires code adjustments to assimilate level-1 backscatter ASCAT data in place of soil moisture retrieval.

4.3.4 SIF

Two global datasets derived from TROPOMI observations in the 735-758 nm window were acquired and pre-processed. The ungridded Caltech dataset (Koehler et al., 2018) was regridded at 0.1° at 8-day and 1-day temporal frequencies. The gridded Troposif dataset (Guanter et a.l, 2021) was produced at 0.1° spatial resolution and 8-day temporal frequency by the LSCE. The pre-processing of both datasets include daily correction factor to obtain daily estimate of SIF along with cloud and unfavourable geometries (view and solar zenith

angle) filtering. These SIF datasets were compared with GPP (fluxcom: Jung et al., 2021; VODCA2GPP: Wild et al., 2022) and LAI (CGLS dataset) satellite-based observations. The evaluations were conducted at continental scale (latitude transects) and site scale to understand (1) how SIF correlates with GPP and LAI observations and (2) identify possible discrepancies between the Caltech and Troposif SIF datasets.

5 Results

5.1 ORCHIDEE modelling framework

Figure 1 below illustrates the improvement of the model prediction after the assimilation of both GPP and SIF data.

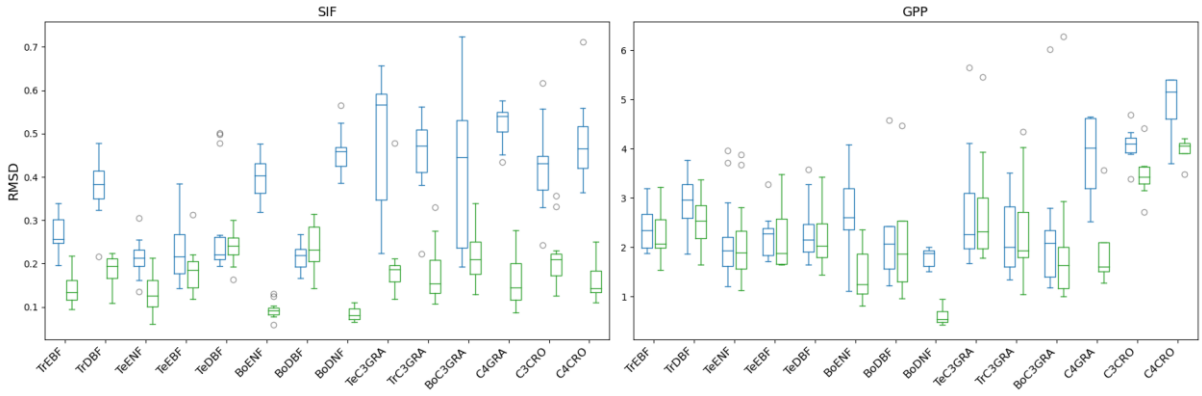

Figure 1: Comparison of the RMSD between data and model simulations before (blue) and after (green) assimilation, for SIF (left) and GPP (right), over ORCHIDEE's PFTs. The assimilations considered here are conducted on SIF data selected using a threshold of 0.5 on cloud fraction.

The improvement is revealed by comparing the prior and posterior RMSDs for GPP and SIF respectively, calculated over all pixels considered for each PFT optimization, with a cloud fraction threshold of 0.5. Except for grasses and crops, the prior GPPs simulated by ORCHIDEE agree well with the *in situ* data, with a RMSD typically lower than 3 gC m^{-2} d^{-1}. Grasses and crops show a larger inter-sites/grid-cells variability and a higher model-data mismatch. A model improvement with respect to GPP following the assimilation is observed for all PFTs. The prior SIF simulations largely overestimate the TROPOMI SIF data. The RMSD is generally largely decreased after the assimilation, except for TeDBF and BoDBF. Except for a few PFTs (TeDBF and BoDBF mainly), the threshold on cloud fraction (CF=0.2 and 0.5) used to select the SIF data that are assimilated (binned at 0.25°/8-day resolutions) has a marginal impact on the model improvement (not shown).

The optimized values of the model parameters were then used for a global scale simulation with ORCHIDEE (ORCHIDEE-opt for CF=0.2, and ORCHIDEE-opt2 for CF=0.5). The global scale simulations were performed at 0.5°/monthly resolutions, based on the CRUJRA meteorological forcing data (Harris et al., 2020; Kobayashi et al., 2015).

Figure 2 shows the yearly GPP maps over the period 2018-2020 for ORCHIDEE simulations, prior and posterior to the data assimilation, as well as for the FLUXSAT reference product.

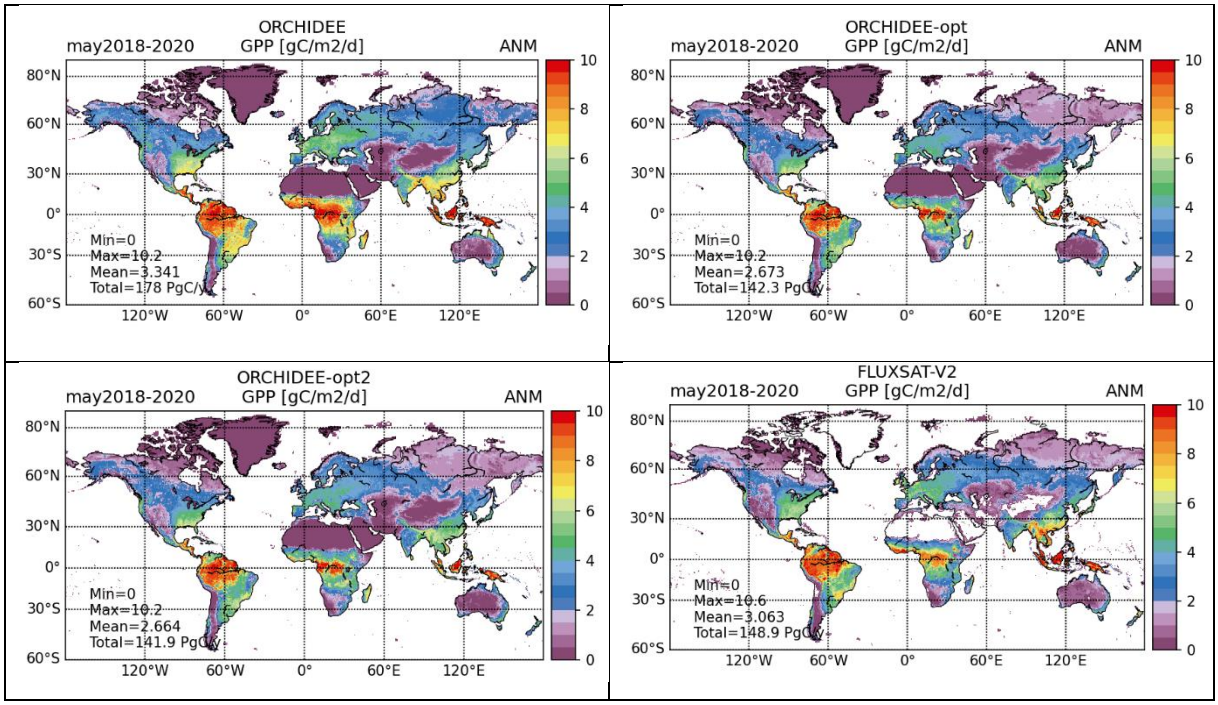

Figure 2: Yearly mean map over the period 2018 (from May) – 2020, for the simulations performed with the ORCHIDEE land surface model prior and posterior ("opt" for assimilations conducted on SIF data selected using a threshold of 0.2 on cloud fraction; "opt2" for a threshold of 0.5 on CF) to data assimilation, and the FLUXSAT reference product. The global minimum, maximum, and mean values are provided (gC m^{-2} d^{-1}), as well as the global budget (in PgC m^{-2} yr^{-1}).

The simulations performed with the prior parameter values resulted in a mean global GPP budget of 178 GtC $yr^{-1,}$ which falls within the upper range of typical GPP estimates. This is a feature specific to this new 2-flux version of ORCHIDEE (which distinguishes between direct vs diffuse light) for which the model parameters were not initially calibrated. The co-assimilation of SIF and GPP data decreased the global budget by about 35 GtC yr^{-1}, resulting in a closer agreement with that of FLUXSAT (149 GtC yr^{-1}). The spatial distribution of the optimized GPP over the tropics is closer to that of FLUXSAT than the prior distribution. This can be partly explained by the fact that the constraint on the optimized model parameters relied on FLUXSAT estimates for the TrDBF (tropical deciduous broadleaf forest) PFT because no *in situ* data were available. However, in ORCHIDEE, this PFT is mostly dominant in the Northern and Southern parts of the African tropical forest, as well as in Northern Australia, and not over the Amazon basin (mostly tropical evergreen broadleaf forests). For the temperate and high latitude regions, we also a convergence between ORCHIDEE and FLUXSAT after the assimilation. The improved agreement in ORCHIDEE simulations against reference GPP data (FLUXSAT as well as GOSIF (Li and Xiao (2019), and LSM simulations performed in the context of the TRENDY v11 exercise (https://sites.exeter.ac.uk/trendy))), achieved through data assimilation, can also be evaluated in Figure 3 with respect to the averaged time series at global scale.

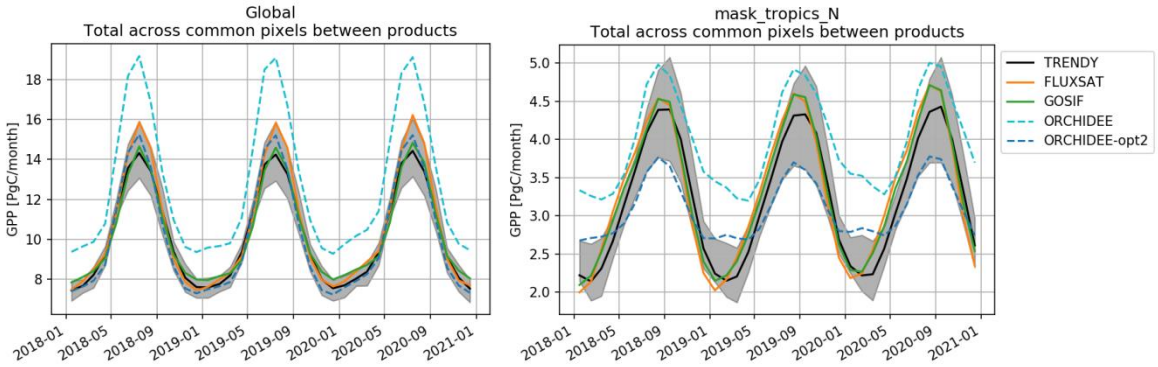

Figure 3: Comparison of the average GPP time series for the ORCHIDEE simulations prior and posterior to data assimilation ("opt2" using a threshold of 0.5 on CF), and against the mean TRENDY simulations (black, standard deviation is shown in grey), FLUXSAT and GOSIF data. The comparison are shown for the global scale (left) and over the tropics - between 30°N and 30°S (right) accounting only for the common pixels.

The global seasonal amplitude of the GPP, too large in the prior, is close to the one of FLUXSAT and GOSIF products after optimization. However, discrepancies are observed in different regions / biomes, as highlighted here over the tropics (30°N to 30°S), where the limited availability of instrumented sites increases the uncertainty in FLUXSAT and GOSIF GPP estimates. In this tropical regions, ORCHIDEE posterior estimates has a much lower seasonal amplitude than FLUXSAT AND GOSIF.

5.2 ISBA modelling framework

5.2.1 Observation operator for ASCAT sigma0

Figure 4 shows the statistical and spatial distributions of the RMSD of the simulated ASCAT sigma0 over southwestern France after training the NNs (one per grid cell).

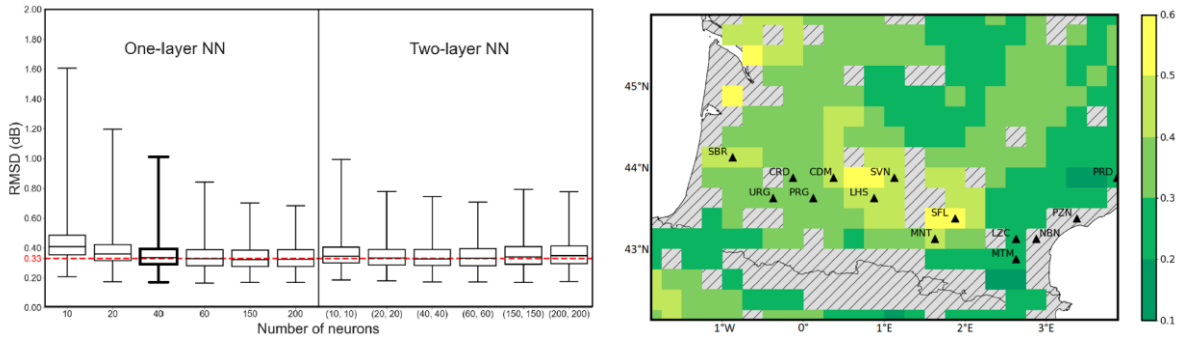

Figure 4: Predicted ASCAT sigma0 RMSD over southwestern France: (left) box plots for 12 NN configurations for the 2007–2014 training period for one-layer and two-layer NN configurations, (right) map for 1-layer, 40-neuron local NNs using ISBA surface soil moisture and soil temperature simulations as input together with PROBA-V LAI observations, for the 2015–2018 test period. Adapted from Corchia et al. 2023.

This figure shows that a single layer consisting of 40 neurons is sufficient to achieve a median RMSD value comparable to the observation error of ASCAT (0.33 dB). Further increasing the number of neurons only slightly decreases the RMSD, while adding hidden layers does not improve the sigma0 predictions. Consequently, a single layer with 40 neurons is determined to be the optimal choice. The RMSD of the simulated sigma0 is often in the range of 0.3 to 0.4 dB (for about 45% of the grid cells). This is in agreement with the mean ASCAT observational error of 0.33 dB.

5.2.2 Observation operator for SMAP

Figure 5 shows the statistical distribution of simulated and observed H-pol SMAP TBs over Europe, from May 2018 to January 2020.

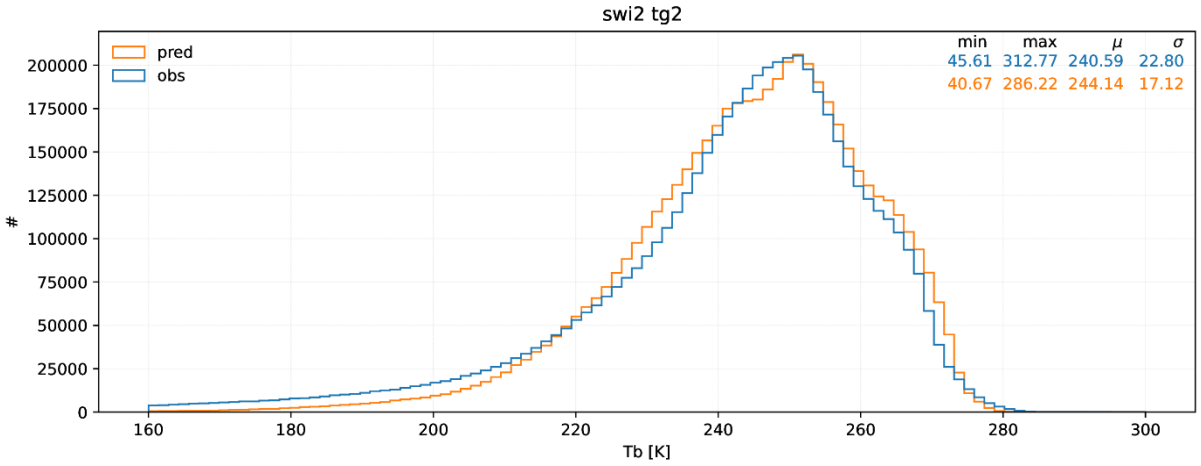

Figure 5: Statistical distribution of predicted and observed SMAP TB at H polarization over Europe, from May 2018 to January 2020. The predicted TB is produced by a global 3-layer NN including 193 neurons, using ISBA surface soil moisture and soil temperature simulations as input.

In this case a single NN is used for all 0.1° x 0.1° grid-cells over Europe and the ISBA surface soil moisture and surface soil temperature simulations are used as predictors to train the NN.

5.2.3 Observation operator for SIF

Figure 6 shows the statistical distribution of simulated and observed SIF (daily TROPOSIF product) over Europe, from June 2019 to May 2020, using a single NN for all 0.1° x 0.1° grid-cells over Europe.

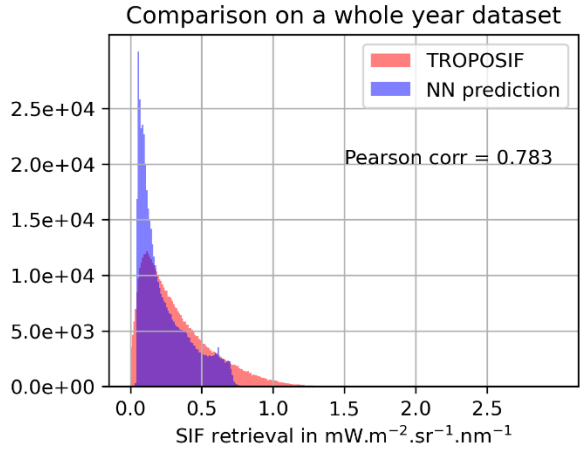

Figure 6: Statistical distribution of predicted and observed SIF (daily TROPOSIF product) over Europe, from 1 June 2019 to 31 May 2020. The predicted SIF is produced by a global 3-layer NN including 193 neurons, using latitude, altitude, plant functional type, ISBA surface soil moisture, soil temperature, GPP, and PROBA-V LAI observations as input.

At this stage of the study, the NN is not able to represent the largest SIF values but the SIF predictions correlate well (r = 0.78) with the observations.

5.3 ECLand modelling framework

5.3.1 AMSR-2 information content analysis

Figure 7 shows the correlations between each model field with the polarization index (PI) in selected AMSR-2 channels.

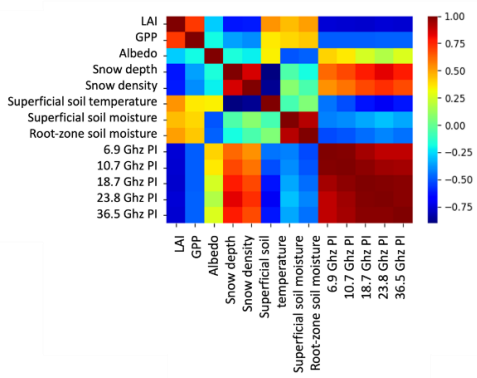

Figure 7: Correlation map of IFS model fields (vegetation, albedo, snow, soil temperature, soil moisture) with polarization index in selected AMSR-2 bands.

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 AMSR-2 brightness temperature of the PI do not show strong dependency with microwave frequency.

5.3.2 ASCAT machine-learning based forward operator

The comparison of ensemble trees xgboost method and feedforward NN showed that a NN with 4 hidden layers, 60 neurons provides the most accurate predictions of ASCAT backscatter at global scale. Figure 8 shows that the spatial distribution of backscatter and its pattern as a function of soil moisture and LAI are accurately reproduced by the NN.

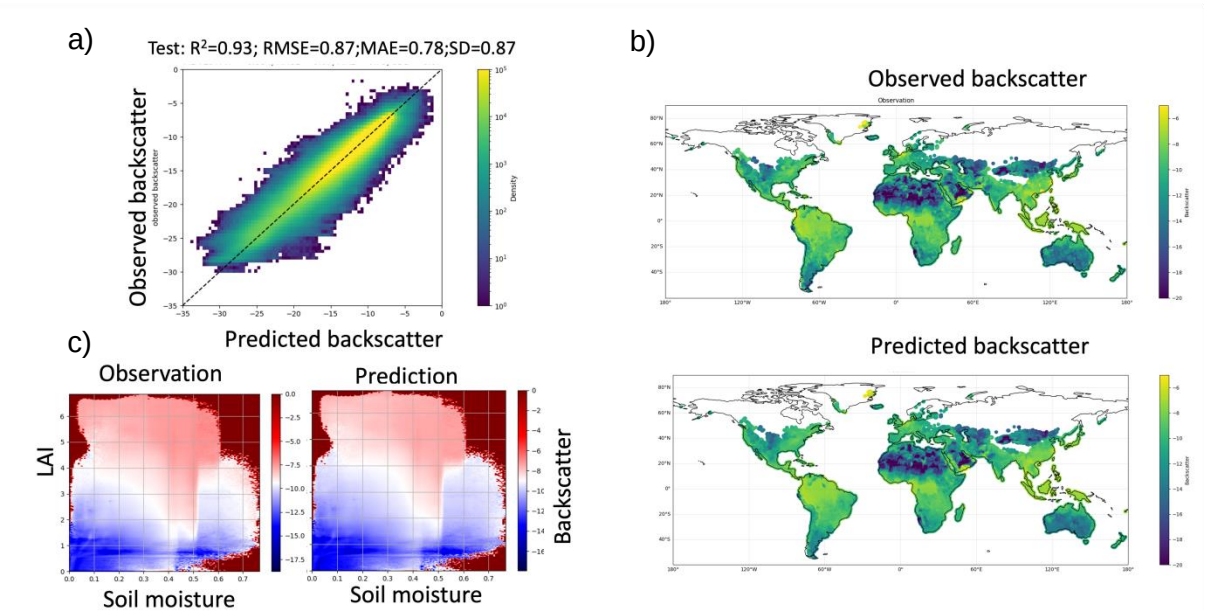

Figure 8: Evaluation of the ASCAT feedforward neural network for year 2019. a): Observation versus NN prediction scatterplot; b): Global maps of observed and predicted backscatter; c) Comparison of predicted and observed backscatter patterns as a function of modelled LAI and surface soil moisture.

The MAE obtained at global scale is within the expected error of the backscatter at 40° product.

5.3.3 SIF data analysis

Both Caltech and Troposif datasets show very consistent spatial (Figure 9) and temporal (Figure 10) distribution.

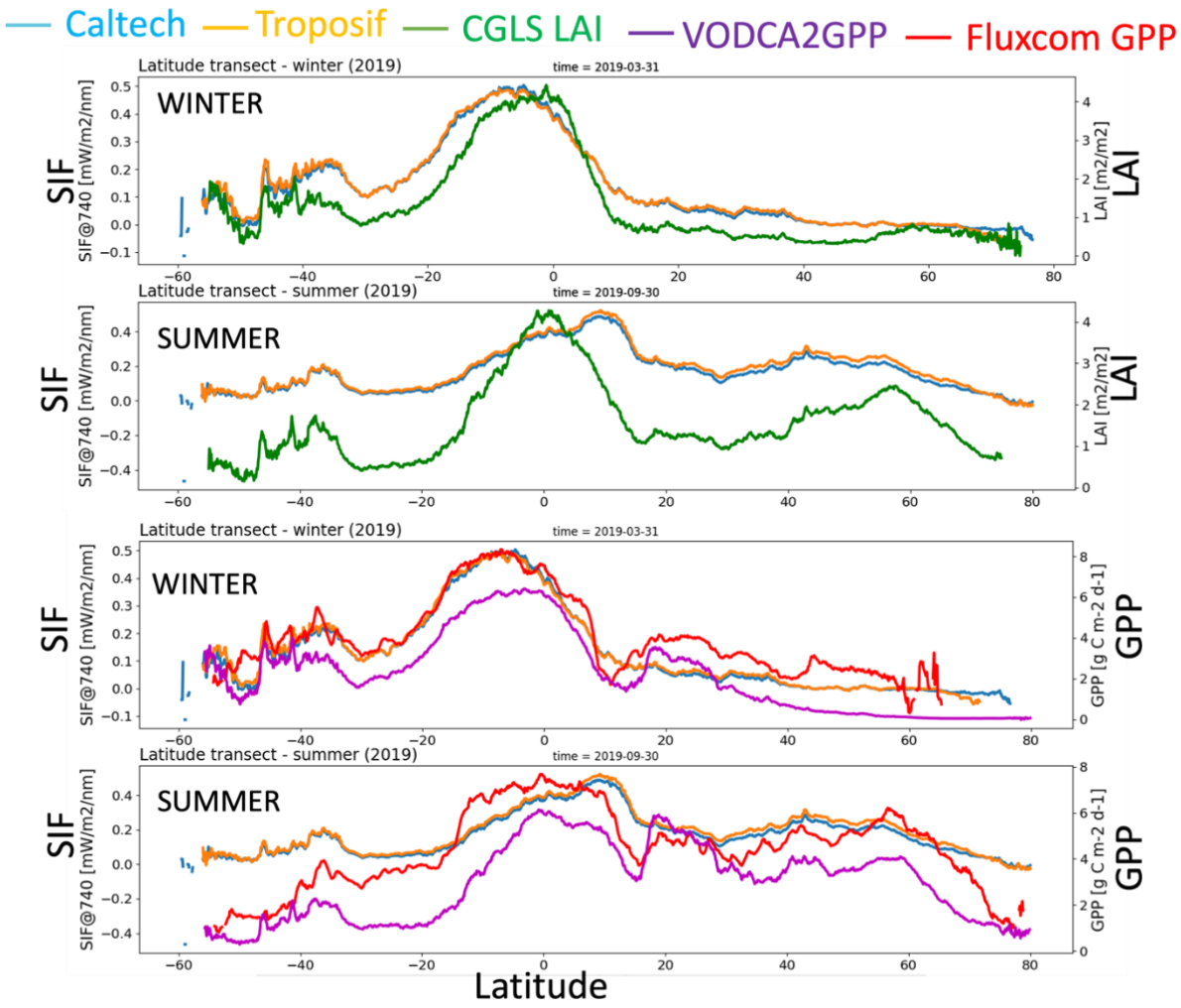

Figure 9: Latitude transects of seasonal mean (summer and winter) of Caltech and Troposif SIF, CGLS LAI and fluxcom and VODCA2GPP GPP satellite-based observations.

The temporal evolutions of both SIF datasets and satellite LAI (CGLS) are very consistent particularly over cropland (Figure 10).

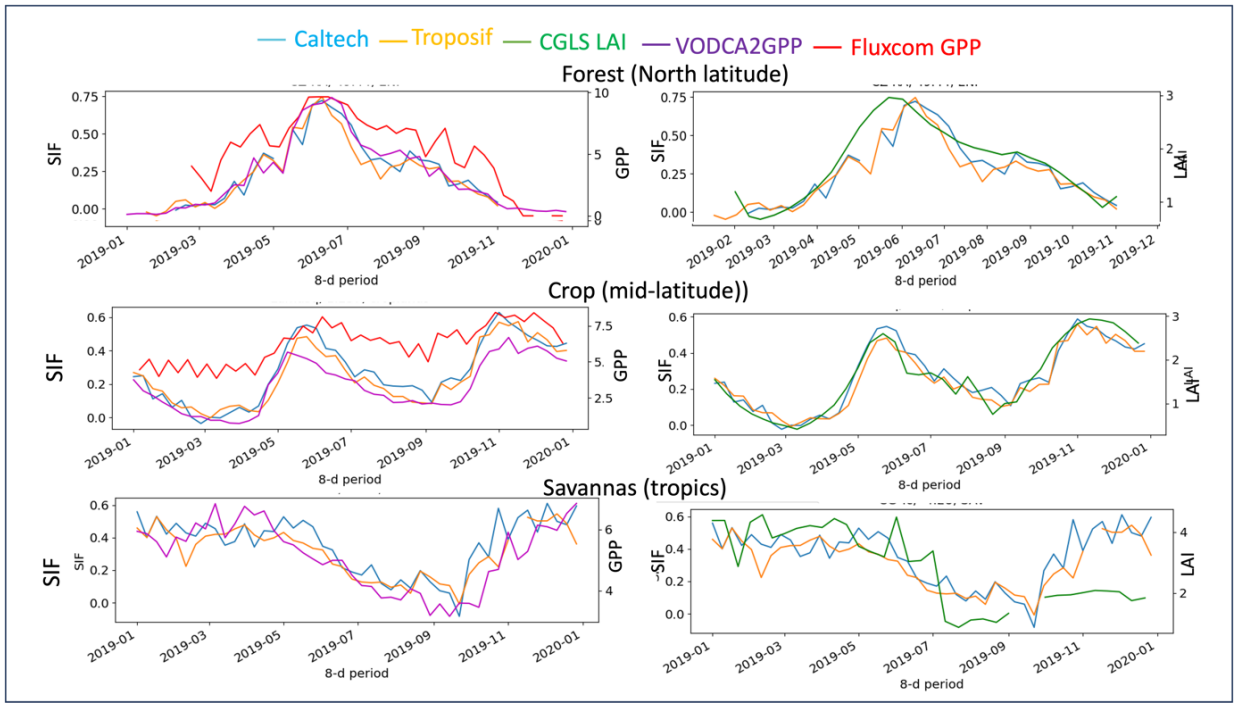

Figure 10: Temporal evolutions of Caltech and Troposif SIF, CGLS LAI and fluxcom and VODCA2GPP GPP satellite-based observations at the site level.

The shift in vegetation peak between LAI and SIF observed at the North latitude forest site is likely related to differences in temporal sampling between the satellite products at high latitude due to differences in geometry or/and cloud contamination. While SIF theoretically relates to photosynthetic activity at the leaf level the SIF signal retrieved from satellite can be strongly influenced by canopy architecture and thus LAI. This explains the large correlation observed between the SIF and LAI datasets. The more erratic evolutions of SIF and LAI observed over the tropical savanna site can be due to residual cloud screening. Both Tropomi and Caltech SIFdatasets show higher spatial and temporal correlations with fluxcom GPP than VOD2CAGPP GPP. This latter shows higher values and more erratic temporal and spatial distributions than fluxcom GPP. Besides, it exhibits spurious temporal evolutions over the cropland site. This illustrates the large uncertainties that are associated with GPP datasets. However, the good correlation between SIF and fluxcom GPP reinforces the idea of using SIF to analyse GPP in land data assimilation system. The next step will consist in evaluating the relationship between SIF and the model GPP.

6 Conclusion

This report presents preliminary results from the development of observation operators for novel satellite observations.

In ECLand and ISBA, machine learning was used to simulate ASCAT sigma0 and SIF.

The next steps for the ECLand ASCAT forward operator will be its implementation in IFS to analyse soil moisture and LAI simultaneously. The impact on carbon, water fluxes and NWP results will be evaluated.

For SIF, the new land surface processes database will be used to test different NN architectures. One of the issues concerns the optimal temporal frequency of SIF to properly represent the temporal variations of GPP.

The next stage will include the following tests:

(1) 1-day and 8-day frequencies in the training of the SIF NN, the latter being first tested in the offline ECLand model and then implemented in the IFS,

(2) the possibility of updating GPP and/or LAI.

Finally, the methodology established for ASCAT will be applied to other passive microwave satellites (SMAP, SMOS, AMSR-2) using the new land surface process training database.

Prior to the global deployment in ISBA, initial NN training tests were carried out over south-west France and Europe for ASCAT sigma0, SMAP TBs and SIF. The next step will be to extend the training to a global scale for these data and to evaluate the observation operators in a data assimilation context. Other data sources such as AMSR2 and SMOS will be considered.

In ORCHIDEE a process-based description of leaf fluorescence was used as an observation operator for SIF. It is able to represent fluorescence integration at canopy level, taking into account canopy structure. The integration of this information into the ORCHIDEE model has to be done together with the integration of in situ GPP observations in order to efficiently optimise the model parameter values.

The different technological choices made between the three modelling frameworks (physical approach in ORCHIDEE, empirical in ECLand and ISBA) will be evaluated in a next step.

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, doi: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, doi: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, doi: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, doi:10.5194/gmd-11-4739-2018, 2018.

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

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.

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.

Harris, I., Osborn, T. J., Jones, P. and Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, Sci. Data, 7(1), 1–18, doi:10.1038/s41597-020-0453-3, 2020.

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, doi:10.1016/j.rse.2014.06.022, 2014.

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

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K. and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, J. Meteorol. Soc. Japan. Ser. II, 93(1), 5–48, doi:10.2151/jmsj.2015-001, 2015.

Koehler, P., Frankenberg, C.,Magney, T. S., Guanter, L., Joiner, J.,& Landgraf, J. (2018). Global retrievals of solar-induced chlorophyllfluorescence with TROPOMI: Firstresults and intersensor comparison to OCO-2. Geophysical ResearchLetters, 45, 10,456–10,463. https://doi.org/10.1029/2018GL079031

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, doi:10.1029/2003GB002199, 2005.

Li, X., Xiao, J. (2019b) Mapping photosynthesis solely from solar-induced chlorophyll fluorescence: A global, fine-resolution dataset of gross primary production derived from OCO-2. Remote Sensing, 11(21), 2563; https://doi.org/10.3390/rs11212563.

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, doi:10.1029/2021GB007177, 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, doi:10.1038/s41597-020-0534-3, 2020.

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, doi:10.5194/bg-6-3109-2009, 2009.

Wild, B., Teubner, I., Moesinger, L., Zotta, R.-M., Forkel, M., van der Schalie, R., Sitch, S., and Dorigo, W.: VODCA2GPP – a new, global, long-term (1988–2020) gross primary production dataset from microwave remote sensing, Earth Syst. Sci. Data, 14, 1063–1085, https://doi.org/10.5194/essd-14-1063-2022, 2022.

Zhang, Y., Bastos, A., Maignan, F., Goll, D., Boucher, O., Li, L., Cescatti, A., Vuichard, N., Chen, X., Ammann, C., Arain, A., Black, T. A., Chojnicki, B., Kato, T., Mammarella, I., Montagnani, L., Roupsard, O., Sanz, M., 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. Discuss., 1–35, doi:10.5194/gmd-2020-96, 2020.

Document History

Version	Author(s)	Date	Changes
1.0	Jean-Christophe Calvet	14.11.2023	Initial version
1.1	Sébastien Garrigues, Cédric Bacour, Pierre Vanderbecken, Jasmin Vural	28.11.2023	Inputs from co-authors
1.2	Cédric Bacour	05.12.2023	Corrections
1.3	Jean-Christophe Calvet	20.12.2023	Revised version 1.3
1.4	Jean-Christophe Calvet	05.01.2024	Revised after internal reviews

Internal Review History

Internal Reviewers	Date	Comments
Frederic Chevallier (LSCE), Lukas Wacker (ETHZ)	Dec 23 and Jan 24	