D4.3: Report on SIF data assimilation method and preliminary testing in the IFS

Due date of deliverable	December 2024
Submission date	December 2024
File Name	CORSO-D4.3-V1.5
Work Package /Task	WP4
Organisation Responsible of Deliverable	ECMWF
Author name(s)	Patricia de Rosnay, Cédric Bacour, Bertrand Bonan, Jean-Christophe Calvet, 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	V1.5
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 investigate Solar Induced Fluorescence (SIF) data assimilation to consistently analyse soil moisture and vegetation variables to constrain NWP and the CO2MVS carbon fluxes in the ECMWF IFS. This work relies on observation operators and land data assimilation systems developments that were conducted in the IFS ECLand Land Data Assimilation System (ECMWF). It is also supported by developments conducted in ISBA (MF), ORCHIDEE (CEA), and D&B (iLab/ULund), allowing to explore different methodologies with different levels of complexity to exploit SIF observations. Data assimilation experiments were conducted in these four land surface models: ECLand, ISBA, ORCHIDEE, and D&B. This report presents the data assimilation approaches and preliminary results obtained in each system, guiding further developments of SIF data assimilation for potential application for the CO2MVS.

Table of Contents

1 Executive Summary 2
2 Introduction 5
2.1 Background 5
2.2 Scope of this deliverable 5
2.2.1 Objectives of this deliverables 5
2.2.2 Work performed in this deliverable 6
2.2.3 Deviations and counter measures 6
2.3 Task 4.2 partners 6
3 Data 6
4 Methods 7
4.1 ORCHIDEE modelling framework 7
4.1.1 Land surface model 7
4.1.2 Data assimilation approach 7
4.1.3 Data 8
in situ GPP data 8
Gridded SIF and GPP data 9
4.1.4 Optimized parameters 10
4.1.5 SIF observations and observation operator 11
4.1.6 Numerical experiments 11
4.2 DALEC & BETHY and TCCAS 12
4.2.1 Land surface model 12
4.2.2 SIF observations and observation operator 13
4.2.3 TCCAS 13
4.2.4 Numerical experiments 13
4.3 ISBA modelling and LDAS-Monde data assimilation framework 13
4.3.1 Land surface model 14
4.3.2 SIF observations and observation operators 14
4.3.3 Simplified Extended Kalman filter (SEKF) 14
4.3.4 Numerical experiments 15
4.4 ECLand modelling and data assimilation framework 16
4.4.1 Land surface model 17
4.4.2 SIF observations and observation operator 18
4.4.3 ECMWF Simplified Extended Kalman Filter 18
4.4.4 ECMWF global LDAS Numerical experiments 19
4.4.5 ECMWF Coupled NWP experiments 20
5 Results 20
5.1 Impact of SIF data assimilation in ORCHIDEE 20

5.1 SIF data assimilation into D&B 21
5.2 SIF data assimilation in LDAS-Monde 25
5.2.1 Baseline and experiments 25
5.2.2 Evaluation on LAI monitoring 25
5.2.3 Discussion on the uncertainties 28
5.3 SIF data assimilation in the IFS ECLand 29
5.3.1 LAI increments 29
5.3.2 Observation operators comparison and evaluation against satellite LAI 30
5.3.3 Evaluation against CGLS LAI 30
5.3.4 Impact on 2m temperature and GPP forecasts 32
6 Conclusion 34
7 References 36

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 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 CO2MVS Research on Supplementary Observations (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), 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 presents the methodology and intermediate results from Task 4.3, which is dedicated to the SIF data assimilation developments and numerical testing to constrain land surface in the coupled IFS used for the CO2MVS.

It uses the SIF observation operators developed in Task 4.1 as described in deliverables D4.1 (First review and improvement of land surface forward operators for SIF and low frequency MW data) that was issued in December 2023, and D4.2 (final review and improvement of land forward operator for SIF and MW data) delivered in December 2024.

2.2.2 Work performed in this deliverable

In this task we used pre-processed SIF observations from Sentinel-5p/TROPOMI from Task 4.1. We also used observation operators described in Task 4.1 using neural network (NN) techniques and physically based forward models in four different surface models (ECLand, ISBA, ORCHIDEE, D&B). We used ORCHIDAS and D&B, and we developed the ECMWF ECLand and Meteo-France Land Data Assimilation Systems (LDAS) to assimilate SIF observations. We conducted numerical experiments assimilating SIF observations in these four systems.

In this document, SIF data assimilation methods and results are presented.

2.2.3 Deviations and counter measures

There was deviation to the plan.

2.3 Task 4.2 partners

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

3 Data

The IFS-based CO2MVS assimilates the same observations as used for Numerical Weather Prediction (NWP). 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.

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 the ESA funded project cover the period 2018-2021 and are available from https://s5p-troposif.noveltis.fr/data-access/. Since, 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) are being generated on a regular basis from 2018 onwards at LSCE (https://doi.org/10.14768/b391bda9-fdfb-40cb-9deb-59b121a18cfb).

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. With a 4DVar assimilation framework (ORCHIDAS), we assimilated SIF retrievals from the Copernicus Sentinel-5p TROPOMI instrument (TROPOSIF product for a set of selected pixels at 0.1°/8-day resolutions) and daily GPP data inferred from eddy-covariance flux measurements to calibrate the main parameters of ORCHIDEE related to photosynthesis and phenology. To assess the informational constraint brought by satellite SIF data on the model parameter, three assimilation experiments were conducted: one where only SIF data are assimilated, one where only GPP data are assimilated, and one where both data streams are combined. We analysed the improvements in the modelled GPP (using the parameters optimised for each of the experiments) by comparing with independent data at the site scale (*in situ* data) and at the pixel, regional, and global scales (data-driven estimates).

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 FluorMODleaf 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 minimise the misfit function (Bastrikov et al., 2018).

Data assimilation experiments are conducted on a PFT-basis, against *in situ* GPP data and TROPOMI SIF retrievals for a collection of selected homogeneous grid cells (0.1°). Although the co-assimilation of these two variables is expected to prevent parameter overfitting, three assimilation experiments are conducted, with different dataset combinations as described above. In the following, only the results for Boreal Needleleaf Evergreen Forest PFT are presented (the co-assimilation for the other PFTs is in progress).

The data used for assimilation and evaluation are presented below.

4.1.3 Data

Aside from TROPOSIF SIF data, *in situ* GPP data from FLUXNET are used for model calibration (data assimilation).

For the evaluation of the model improvement following the three data assimilation experiments, we compare the model simulations to data-driven GPP estimates (FLUXCOM-X-BASE (Nelson et al., 2024) and FluxSat (Joiner et al., 2018)). Using multiple reference datasets provides a more robust assessment of the model improvement.

in situ GPP data

Daily *in situ* GPP estimates from FLUXNET (Baldocchi et al., 2001; Pastorello et al., 2020) are assimilated. A particular effort has been dedicated to data cleaning (filtering inconsistent gap-filled data, while removing negative GPP values). Table 1 describes the main characteristics of the GPP data at the sites considered for data assimilation (11 sites) as well as for those used in evaluation (6 independent sites), and Figure 1 shows their location worldwide.

Table 1: Characteristics of the eddy-covariance sites considered for the assimilation (blue) and evaluation (orange) of GPP.

Site ID	Latitude (°)	Longitude (°)	Main PFT Fraction	Years	Total years
US-NR1	40.03	-105.55	0.9	1999-2014	16
US-GLE	41.37	-106.24	0.8	2005-2014	10
CA-Qfo	49.69	-74.34	0.9	2004-2010	7
CA-Ojp	53.92	-104.69	0.9	2000-2005	5
CA-Obs	53.99	-105.12	0.95	2000-2005	5
CA-Man	55.88	-98.48	0.85	1998-2003	6
RU-Zot	60.80	89.35	0.8	2002-2004	3
FI-Hyy	61.85	24.29	0.9	1996-2019	24
SE-Fla	64.11	19.46	0.85	1996-1998+2001-2003	6
SE-Ros	64.17	19.74	0.85	2015-2020	6
FI-Var	67.75	29.61	0.8	2016-2020	5
US-Syv	46.24	-89.35	0.45	2001-2006+2012-2014	9
CA-NS5	55.86	-98.48	0.7	2002-2005	4
CA-NS6	55.92	-98.96	0.8	2002-2005	4
SE-Svb	64.26	19.77	0.85	2014-2016+2018-2020	6
FI-Sod	67.36	26.64	0.7	2001-2014	14
FI-Ken	67.99	24.24	0.85	2018-2019	2

Gridded SIF and GPP data

For each of the 14 vegetation PFTs, we selected forty grid cells at 0.1° with the highest thematic homogeneity while ensuring a correct sampling of the global distribution for assimilation and evaluation. Figure 2 below shows the spatial distribution of these selected homogeneous grid cells: 20 grid cells being allocated for assimilation purposes and 20 for evaluation. The latitudinal profile shows a relatively homogeneous distribution across latitudes for both assimilation and evaluation grid cells.

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 smaller than 40° and cloud fraction below 0.5 were considered.

Data-driven GPP estimates at 0.1° from FLUXCOM-X-BASE (Nelson et al., 2024) and FluxSat (Joiner et al., 2018) over the period 2001-2021 are used for model evaluation.

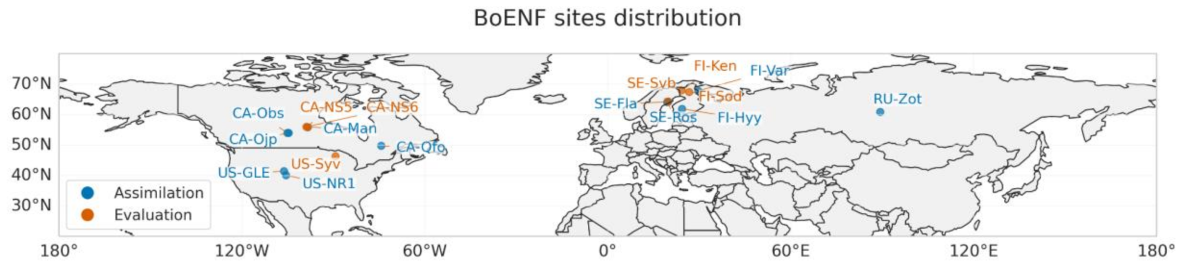


Figure 1: Location of the eddy-covariance sites considered for the assimilation of GPP data (blue) and for the evaluation (orange)

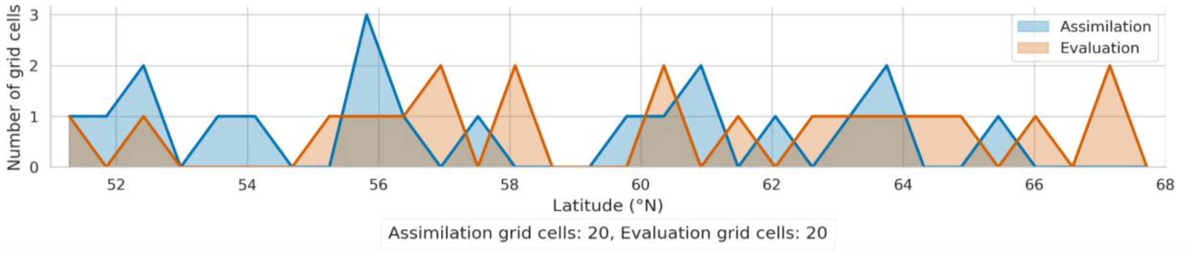


Figure 2: Top) Map of the location of the 0.1° grid cells considered for the assimilation of TROPOSIF SIF data (blue) and for evaluation (orange) against SIF and data-driven GPP estimates; and bottom), latitudinal profile of the number of considered grid-cells.

Model-data error

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. For the co-assimilation experiment, we balance the misfit functions associated with SIF and GPP with respect to their respective number of observations.

4.1.4 Optimized parameters

Prior to assimilation, we conducted a sensitivity analysis using the Morris (1991) method to identify the most influential parameters on SIF and GPP to be calibrated by data assimilation. The analysis was conducted both on the grid cells and over the eddy flux sites, with respect to both SIF and GPP simulations, considering the corresponding meteorological data and PFT fraction characterization. It is worth noting that the analysis for GPP on the two datasets provided consistent results. At the end, we chose to optimise 14 parameters for SIF and 10 for GPP, affecting different processes as seen in Table 2 below. All parameters are optimised in the co-assimilation experiment.

Table 2: Characteristics of the parameters optimised against SIF and GPP data.

Parameter	Description	Prior values	Variation range	Observational constraint	
				SIF	GPP
Photosynthesis					
VCMAX25	Maximum carboxylation rate limited by Rubisco activity at 25°C ($\mu mol \cdot m^{-2} \cdot s^{-1}$)	45	[33.75, 56.25]	✓	✓
ASJ	Entropy parameter offset for Jmax temperature dependence ($J \cdot K^{-1} \cdot mol^{-1}$)	660	[495, 825]	✓	✓
ASV	Entropy parameter offset for Vcmax temperature dependence ($J \cdot K^{-1} \cdot mol^{-1}$)	668	[501, 835]	✓	✓
ARJV	Offset for Jmax/Vcmax ratio temperature acclimation ($\mu mol \cdot e^{-1} \cdot m^{-2} \cdot s^{-1}$ / $\mu mol\ CO_2 \cdot m^{-2} \cdot s^{-1}$)	2.59	[1.94, 3.238]	✓	✓
E_JMAX	Energy of activation for Jmax ($J \cdot mol^{-1}$)	49880	[37410, 62360]	✓	✓
Phenology					
LEAFAGECRIT	Critical leaf age (days)	910	[682.5, 1138]	✓	✓
LAI_MAX	Maximum leaf area index ($m^2 \cdot m^{-2}$)	4.5	[3.375, 5.625]	✓	✓
Canopy Structure					
ALA	Average leaf angle (°)	75	[55, 85]	✓	
CLUMPING	Clumping index (-)	0.55	[0.5, 0.8]	✓	✓
Allocation					
SLA	Specific leaf area ($m^2 \cdot g^{-1}$)	0.00926	[0.006945, 0.01158]	✓	✓
SIF & GPP models					
k_F	Fluorescence relative rate constant (s^{-1})	0.1	[0.04, 0.11]	✓	

a_psII	Photosystem II absorption (-)	0.5	[0.375, 0.625]	✓	✓
p1_NPQr	NPQ reversible model parameter 1 (-)	0.94	[0.705, 1.175]	✓	
p2_NPQr	NPQ reversible model parameter 2 (-)	5.15	[3.862, 6.438]	✓	

4.1.5 SIF observations and observation operator

We use TROPOSIF weekly means to decrease the relatively high random error associated with 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.

4.1.6 Numerical experiments

In this sub-section 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 as presented in the CORSO D4.2 report (https://www.corso-project.eu/deliverables). The assessment is performed over the selected 40 homogeneous grid cells at 0.1°/weekly resolution and over the period 2001-2021.

Figure 3 presents the boxplot distributions over the 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. We observe a generally consistent (mis)match between model and data across the various PFTs for both SIF and GPP variables (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., TrEBF or C4GRA), indicating where co-assimilation should be even more relevant.

For Boreal Needleleaf Evergreen Forest (BoENF), the median SIF error is relatively low (about 0.25 $mW.m^{-2}.sr^{-1}.nm^{-1}$) compared to the median error across all PFTs (0.35 $mW.m^{-2}.sr^{-1}.nm^{-1}$); the median error for GPP is about 3 $gC.m^{-2}.d^{-1}$ (with larger errors when considering FLUXCOM-X-BASE compared to FluxSat), which is slightly higher than the median value over all PFTs (2.4 $gC.m^{-2}.d^{-1}$). The high coefficient of determination for GPP (around 0.9) suggests that ORCHIDEE captures a variability consistent with that of the data-driven products; this is however not the case for SIF (R^2 about 0.5).

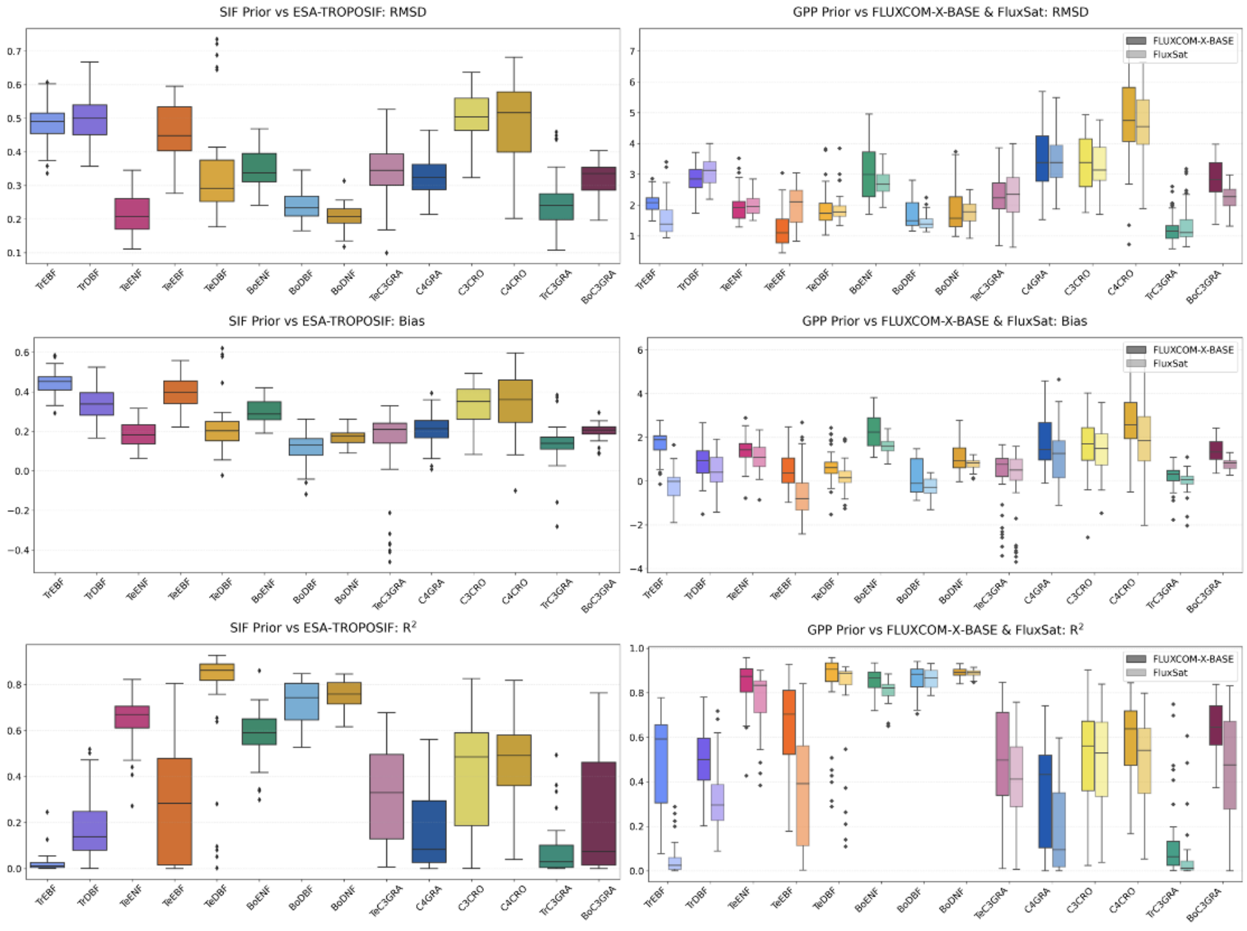


Figure 3: Boxplot of (top) Root Mean Squared Differences (RMSD), (middle) bias and (bottom) coefficient of determination (R²), for: (left) prior SIF (in mW.m^{-2}.sr^{-1}.nm^{-1}) vs TROPOSIF observations over the period 2019-2022; (right) prior GPP (gC.m^{-2}.d^{-1}) vs FLUXCOM-X-BASE / FluxSat estimations over the period 2001-2021, over an ensemble of homogeneous pixels (0.1°).

4.2 DALEC & BETHY and TCCAS

This subsection presents the DALEC & BETHY (D&B) terrestrial biosphere model (Knorr et al., 2014) and the Terrestrial Carbon Community Assimilation System (TCCAS https://tccas.inversion-lab.com) around it. Both are open-source developments of a larger team.

4.2.1 Land surface model

The D&B model (Knorr et al., 2014) 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 CO2 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, considering 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).

4.2.2 SIF observations and 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 canopy layer SIF, Sn, 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 sSIF. This scaling factor compensates for large uncertainties in some of the constants needed to calculate Sn 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 μmol $m^{-2}s^{-1}$ and then averaged.

4.2.3 TCCAS

TCCAS is a variational assimilation system that is set up for assimilation of a range of data streams linked to terrestrial carbon cycling. It assimilates all data streams in a single long assimilation window, which ensures conservation of carbon. The assimilation adjusts process parameters of the D&B model and of the observation operators. The calibrated process parameters can then be used to estimate terrestrial carbon fluxes and pools consistent with the assimilated observations.

4.2.4 Numerical experiments

Here we operate TCCAS for assimilation of only TROPOMI SIF observations into a seven-year run from 2015 to 2021 with an assimilation window from 2017 to 2021, i.e. we allow two years of spin-up. Observations are assimilated at the time of the overpass. We perform an experiment for the ICOS field site Sodankylä. For validation, we use two independent data sets, GPP derived from eddy-covariance measurements and FAPAR derived by the JRC-TIP (Pinty et al., 2007) from MODIS broadband albedos (Pinty et al., 2011).

4.3 ISBA modelling and LDAS-Monde data assimilation framework

Météo France worked on SIF data assimilation 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 developed in Task 4.2 were used. They are based on neural networks (NNs) trained with ISBA simulations and LAI observations from the PROBA-V satellite to predict the microwave signal. The globally trained NN-based observation operators (one NN for all grid cells) were implemented in LDAS-Monde, which allows the sequential assimilation of backscatter observations (Corchia et al. 2023).

4.3.1 Land surface model

The version of the model that is used for this study can represent 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.

4.3.2 SIF observations and 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 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 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.

Here the NN operator as designed and used in this study, follows a feedforward architecture with two hidden layers with 128 neurons. The inputs are the LAI, the geographical coordinates and the day of year. More inputs can be added to reach better accuracy on the NN prediction of TROPOSIF (Bacour et al., 2019), but it was for us a trade off with its integration in the assimilation scheme.

4.3.3 Simplified Extended Kalman filter (SEKF)

Data assimilation process is finding our best control vector (i.e. our surface state) feasible according to a model and minimising the discrepancies toward observations and a prior estimate. The best control vector resulting from these processes is usually called analysis. Our analysis is done using a simplified extended Kalman filter. Thereafter the control vector will be referred to as **x** and the subscript will denote the temporal step.

The analysis update equation at t=i of the Kalman filter from the background control vector at t=i-1 is as follows:

$$\mathbf{x}_i^{\mathbf{f}} = \mathbf{M}_i \, \mathbf{x}_{i-1}^{\mathbf{f}}$$

$$\mathbf{x}_i^{\mathbf{a}} = \mathbf{x}_i^{\mathbf{f}} + \mathbf{K}_i(\mathbf{y}_i^{\mathbf{o}} - \mathbf{H}_i(x_i^f))$$

where exponent “a”, “f”, and "o" stand for analysis, forecast and observation, respectively. The operator **M** and **H** are respectively the forward model (i.e. ISBA-A-gs) and the linear observation operator that maps the control vector into the observation space.

The Kalman gain $\mathbf{K}_i$ is defined at time t=i as follows:

$$\mathbf{K}_i = \mathbf{B}\mathbf{M}_i^{\mathbf{T}}\mathbf{H}_i^{\mathbf{T}}(\mathbf{R} + \mathbf{H}_i\mathbf{M}_i\mathbf{B}\mathbf{M}_i^{\mathbf{T}}\mathbf{H}_i^{\mathbf{T}})^{-1},$$

where **B** and **R** are error covariance matrices characterising the background and observation vectors. Superscript T indicates matrix transpose. These formulations assume the model and the observation operator to be linear. That is why, in our case we use an extended Kalman filter by using the tangent linear of the operator **J** of the composition of the observation operator with the forward model instead of its non-linear counterpart. The extended Kalman gain is then defined as follow:

$$\mathbf{x_i^a} = \mathbf{x_i^f} + \boldsymbol{K_i}(\mathbf{y_i^o} - \mathbf{H_i}(x_i^f))$$

$$\boldsymbol{K_i} = \mathbf{BJ_i^T}(\mathbf{R} + \mathbf{J_iBJ_i^T})^{-1},$$

where the tangent linear **J** (and his transposed $\mathbf{J^T}$), thereafter named Jacobians, are computed using finite differences. It is done by perturbing each component of the control vectors.

To simplify the extended Kalman filter, the background error covariance matrix and the observation error covariance matrix are assumed to be diagonal and the covariances values are fixed (Mahfouf at al., 2009; Albergel et al., 2010; Barbu et al., 2011; Fairbairn et al., 2017; Bonnan et al., 2020; among others) for a given assimilation window of 24h. The initial state is given by the analysis made over the previous 24h assimilation window.

4.3.4 Numerical experiments

The neural network operator was trained over Europe at 0.1 degree of resolution over a period covering June 2018 up to May 2019 and tested over June 2019 to May 2020. As shown in Figure 4 below, the neural network tends to capture the large-scale features present in TROPOSIF observation data. The prediction is usually underestimating high TROPOSIF values. Nevertheless, both the Pearson correlation and the RMSE are good and consistent between the training and the test dataset (Table 3).

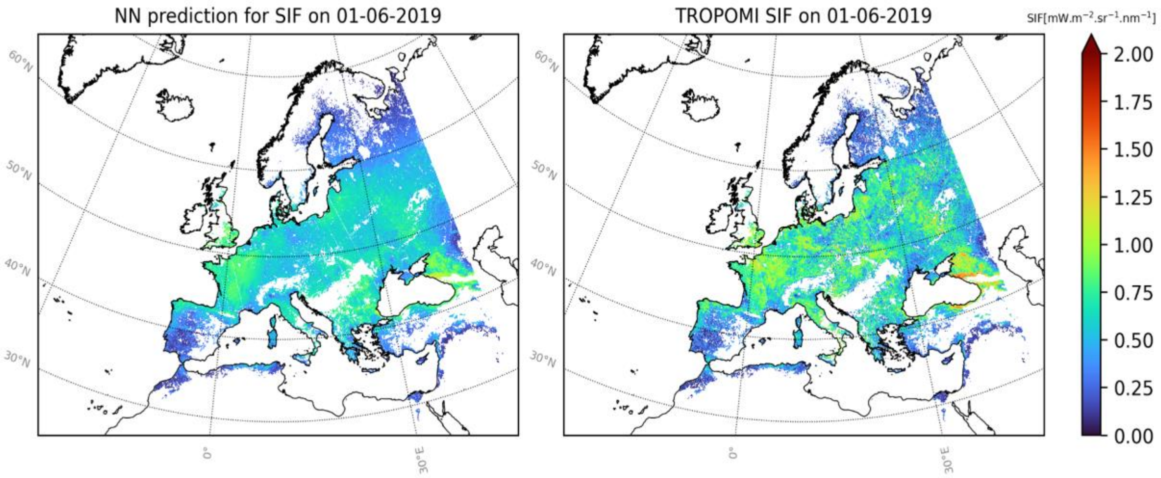


Figure 4: Examples of neural networks predictions of TROPOSIF for the 2019, 1st of June. Left the neural network prediction, right the TROPOSIF values.

Table 3: Performances of the MF neural network SIF observation operator for the training and test periods

Dataset (Start -End)	RMSE [mW.m^{-2}.sr^{-1}.nm^{-1}]	Pearson correlation
Train - (06/2018-05/2019)	0.14	0.82
Test - (06/2019-05/2020)	0.15	0.81

The LAI used from the training was linearly interpolated from the 10-days LAI -V1 product from the Copernicus Land Monitoring Service (CLMS). This way, the neural network was not sensitive to our land model. Yet, once integrated in our assimilation scheme, it will use the LAI resolved by the land model.

To reduce the computational cost of the assimilation experiment, we've done it on a sub-domain of 40 x 40 pixels. The tested area is over the Ebro basin in Spain. The simulation of the vegetation in this area is known to be complexified by uncharted irrigated croplands (see Figure 5) making it a good test for the impact of the assimilation of SIF.

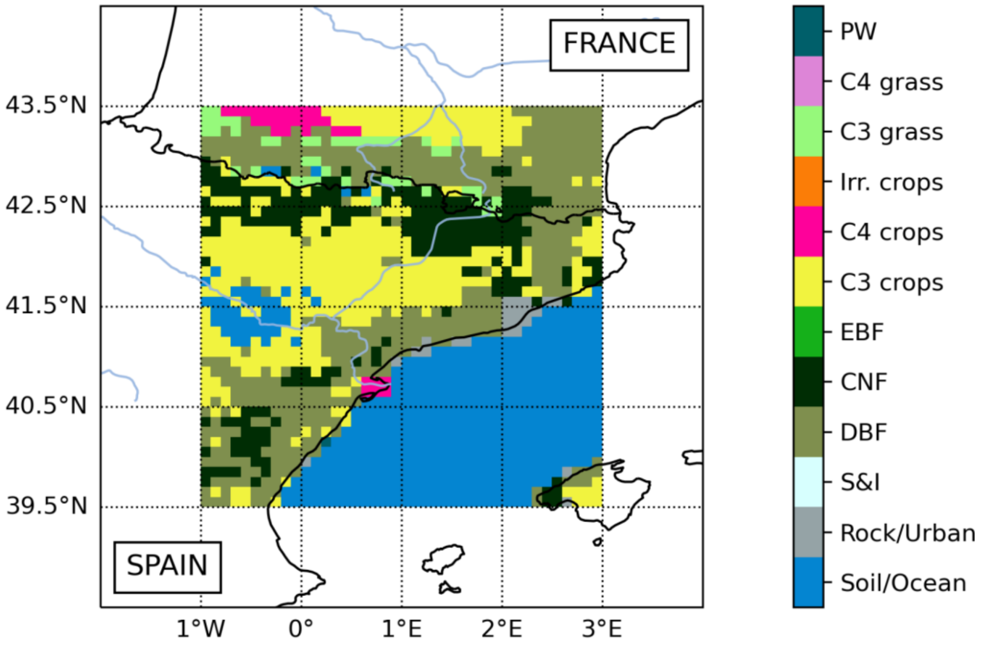


Figure 5: Most dominant PFT according to ECOCLIMAP-SG over the Ebro basin domain among the 12 PFT considered (i.e Bare soil, Bare rock, Snow and Ice, Deciduous Broadleaf Forest, Coniferous Needle Forest, Evergreen Broadleaf forest, C3 crops, C4 crops , Irrigated crops, C3 grassland , C4 grassland, Peats and Wetlands).

4.4 ECLand modelling and data assimilation framework

The work of ECMWF was dedicated to the implementation of the machine learning-based observation operators developed in Task 4.1 to assimilate SIF in the ECMWF LDAS to update the LAI climatology used to initialise coupled land-atmosphere forecasts in the IFS. The work was conducted at global scale with IFS cycle 49r1 implemented in operations in November 2024.

The methodological approach is presented by Figure 6. First SIF observations are assimilated in the LDAS to update the low and high vegetation LAI variables of ECLand, second the updated LAI variables are used in IFS coupled forecast experiments to evaluate their impacts on the prediction of carbon fluxes (GPP) and low-level meteorological variables (2m humidity and temperature, 10m winds).

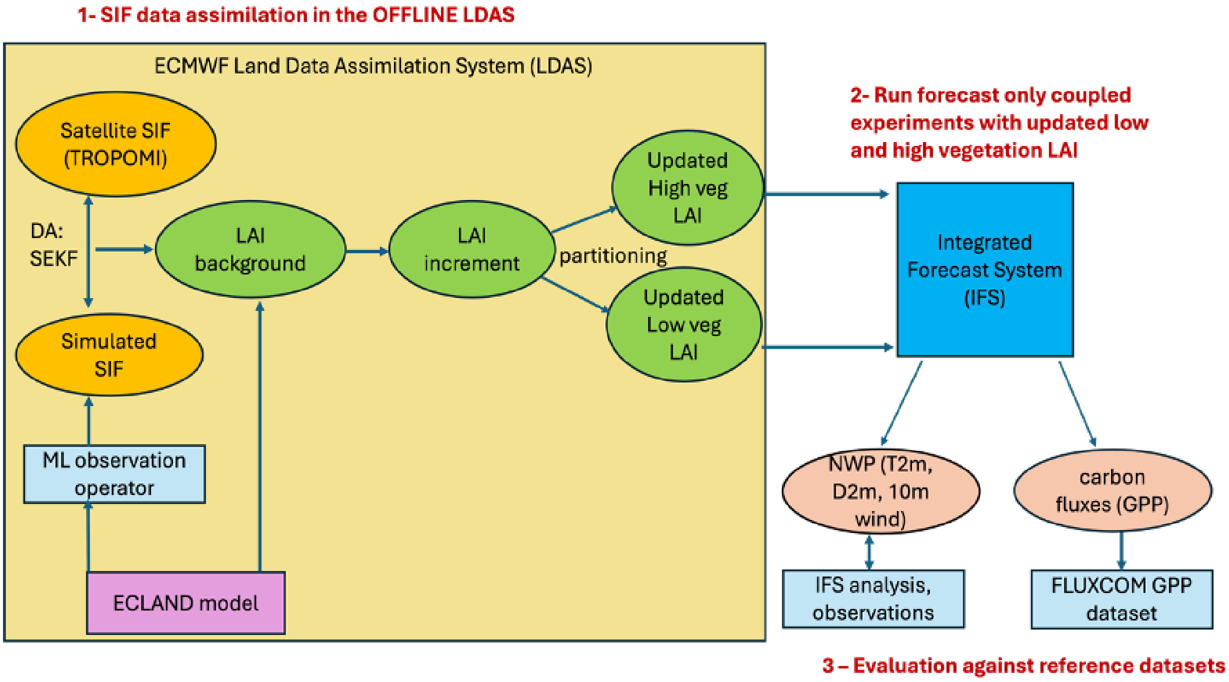

1- SIF data assimilation in the OFFLINE LDAS
ECMWF Land Data Assimilation System (LDAS)
Satellite SIF (TROPOMI)
DA: SEKF
LAI background
LAI increment
partitioning
Updated High veg LAI
Updated Low veg LAI
Simulated SIF
ML observation operator
ECLAND model
2- Run forecast only coupled experiments with updated low and high vegetation LAI
Integrated Forecast System (IFS)
NWP (T2m, D2m, 10m wind)
carbon fluxes (GPP)
IFS analysis, observations
FLUXCOM GPP dataset
3 – Evaluation against reference datasets

Figure 6: Schematic representation of the SIF data assimilation approach at ECMWF.

4.4.1 Land surface model

The ECMWF land-surface modelling system is ECLand. It is based on HTESSEL model (Tiled ECMWF Scheme for Surface Exchanges over Land incorporating land surface hydrology), that represents vertical processes and exchanges with the atmosphere (Balsamo *et al*., 2009; Boussetta *et al*., 2021). The vegetation representation in ECLand relies on a tile approach which accounts for dominant low (grassland, crop, shrubland) and high (forest) vegetation. In the current version of ECLand used in the IFS, vegetation parameters such as LAI are specified as seasonally varying climatological monthly mean maps in the ECMWF Numerical Weather Prediction (NWP) system. This climatology uses the latest Copernicus Global Land Service (CGLS) LAI dataset (over 1993-2019) from the CONFESS project (Boussetta & Balsamo, 2021) and has been shown to have a significant impact on the quality.

One of the main weaknesses of the current approach is that the inter-annual variability of the vegetation is not considered. Inter-annual differences in vegetation can be large because of meteorological events such as droughts, above average rainfall and variations in 2 metre temperature. Results from the CoCO2 projects highlighted promising impact of Passive Microwave (PMW) Vegetation Optical Depth (VOD) data assimilation in the ECMWF LDAS [Calvet et al 2023, CoCO2 D3.4 Demonstrator systems for using remote sensing data (LAI, VOD, SIF) in online global prior fluxes for the CO2MVS prototype].

Building-up on the COCO2 project, but using SIF observations instead of PMW VOD data, we use data assimilation to update the ECland LAI daily. Using this approach enables dynamic updates to the vegetation parameters to respond to inter-annual variations.

4.4.2 SIF observations and observation operator

To assimilate SIF in the ECMWF LDAS, we used the ML-based observation operators based on the XGBOOST (XGB) gradient boosted trees (Chen et al., 2016) which were developed in Task 4.1 and presented in the deliverable D4.2. The XGB models were trained over the 2019-2020 period, tuned over 2021 and evaluated over 2022. Three models based on distinct sets of predictors were tested (Table 4). M1 and M2 rely on selected ECLand/IFS physical predictors while M3 was trained from the CGLS 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, the fraction of high (CVH) and low (CVL) vegetation, were removed.

Table 4: Predictors of the ML-based observation operators evaluated in this work. SM is the soil moisture of the top soil layer (7 cm), SM-1m is the root-zone soil moisture within 1m of soil, ST is the soil temperature of the top soil layer, T2M and D2M are the 2m temperature and dewpoint temperature, 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

Results from Task 4.1 show that while all the models accurately predict SIF at global scale, M3 has the highest performance scores (see D4.2 for more details).This work further evaluates these models by testing them in the ECMWF LDAS, and it evaluates their impacts on the LAI increments produced by the data assimilation system.

4.4.3 ECMWF Simplified Extended Kalman Filter

IFS cycle 49r1 SEKF

The ECMWF LDAS is composed of several components for the screen-level parameters (2-m temperature and relative humidity), snow, and soil moisture, soil temperature, and snow temperature (de Rosnay et al., 2022). The screen level analysis and the snow analysis are conducted using a 2D-OI (2-Dimensional Optimal Interpolation) and the soil moisture analysis is conducted using a simplified Extended Kalman Filter (SEKF) approach. The soil temperature and snow temperature analyses are conducted using a 1D-OI. The ECMWF LDAS runs twice per day using 12-hour data assimilation windows.

In this study we use the LDAS in offline mode (Rodríguez-Fernández et al, 2019). The LDAS reads atmospheric forcing from ERA5 (Hersbach et al, 2020). The SEKF is run and the land surface analysis produced is passed to the land surface forecast model ECLand. This has the disadvantage that there is only a one-way coupling between the atmosphere and the land but allows for developments to be tested at low computing cost for multi-year periods before being implemented in the weakly coupled IFS LDAS.

The SEKF analyses 3 layers of soil moisture (0-7 cm, 7-28 cm, 28-100 cm) using the aforementioned pseudo-observations of 2 metre temperature and relative humidity in combination with surface soil moisture observations from the Advanced Scatterometer (ASCAT) instrument onboard the MetOp series of satellites. The analysis is calculated using the Kalman filter equation:

$$\mathbf{x_a} = \mathbf{x_b} + \mathbf{K}\,(\mathbf{y} - H(\mathbf{x_b})) \tag{4.1}$$

$$\mathbf{K} = \mathbf{BH}^{T}\,(\mathbf{R} + \mathbf{HBH}^{T})^{-1} \tag{4.2}$$

where $\mathbf{x_a}$ is the analysis, $\mathbf{x_b}$ is the background, $\mathbf{y}$ are the observations, H is the observation operator, **H** contains the Jacobians from the ensemble of data assimilations (EDA) to link the model variables to the observed variables, **B** is the background error covariance matrix and **R** is the observation error covariance matrix.

Updated SEKF with SIF assimilation

The objective in CORSO is to assimilate SIF to update the low and high vegetation LAI climatology variables of ECLand at a daily timestep (step 1 in Figure 6).

The smoothed 8-day SIF observations were resampled at a daily time step assuming constant daily value within the native 8-day period of the TROPOSIF product and transformed into GRIB files which were interfaced with the ECMWF LDAS. Quality control and removal of unfavourable surfaces (snow, frozen soil, orographic regions, water bodies) were applied to the SIF observations prior to their assimilation.

The SIF observation was added in the SEKF observation vector, and the control vector was appended to include the total LAI variable. The finite difference method was used to compute the Jacobian of SIF with respect to LAI. At this stage, the SIF assimilation can only update LAI and the Jacobian of SIF with respect to soil moisture was set to zero. The LAI increments, namely the difference between the updated LAI and the LAI climatology used as the model background, produced by the SEKF are partitioned into low and high vegetation LAI according to the fractions of low and high vegetation, respectively. The analysis is performed only for the 00z data assimilation window to ensure a single daily update of LAI. The updated LAI values for low and high vegetation are used subsequently as input LAIs, instead of the climatology, of the coupled land-atmosphere forecasts.

4.4.4 ECMWF global LDAS Numerical experiments

Table 5 presents the ECMWF LDAS experiments which were conducted with the three ML-based observation operators presented in Table 4. Two additional experiments (LDA4 and LDA5) were run to test different values of the background and observation errors. The experiments were performed over the 2022-2023 period. Evaluation results for the year 2022 are presented in this report.

An intercomparison of the LAI increments produced by each experiment is conducted to characterize the impacts of using different sets of predictors in the observation operators on the SIF assimilation. The updated LAI is evaluated against the CGLS LAI dataset based on

SENTINEL-3/OLCI satellite observations. Global maps of temporal correlation and RMSE are computed between 1) the updated LAI and the satellite LAI and 2) the LAI climatology and the satellite LAI. The difference in correlation and RMSE between 1) and 2) are used to assess the increments produced by each DA configuration.

Table 5: List of the ECMWF LDAS SIF data assimilation experiments

Experiment name	**ML model predictors**	**Background LAI error standard deviation (unit m^{2} m^{-2})**	**SIF Observation error standard deviation (unit mW m^{-2} nm^{-1}sr$^{-1)}$**
LDA1	**M1**	**0.4**	**0.1**
LDA2	**M2**	**0.4**	**0.1**
LDA3	**M3**	**0.4**	**0.1**
LDA4	**M2**	**1**	**0.1**
LDA5	**M2**	**1**	**0.05**

4.4.5 ECMWF Coupled NWP experiments

The updated LAI produced by the LDAS data assimilation experiments were used instead of the default LAI climatology in IFS forecast (fc) experiments. The fc experiments were conducted over summer 2022 (from 01/06/2022 to 31/08/20222) and over winter 2022-2023 (from 01/12/2022 to 28/02/2023). Separate IFS experiments were conducted for each of the LDA experiments listed in Table 4.5. For each simulation period a control experiment based on the default LAI climatology was performed.

The impacts on NWP forecasts (low-level meteorological variables: 2m temperature and humidity and 10 m wind) are assessed against the operational analysis (IFS cycle 49r1) considered here as a reference. The impact on the GPP forecast is assessed by comparison with the simulated GPP from the control experiment. At the time of this report, the FLUXCOM dataset was not available for the year 2022. The comparison against FLUXCOM will be conducted in the final report.

5 Results

5.1 Impact of SIF data assimilation in ORCHIDEE

We illustrate the informational constraint provided by space-borne SIF retrievals on GPP simulations with ORCHIDEE for the Boreal Needleleaf Evergreen forest PFT.

To assess the impact of the different assimilation experiments on model performance for BoENF, we compare the prior and posterior simulations against independent observations at multiple scales. The evaluation focuses on: 1) mean seasonal cycles (Figure 7), comparing results from the three assimilation experiments (SIF-only, *in situ* GPP-only, and SIF-GPP) for 20 independent grid cells (against TROPOSIF data as well as the data-driven GPP estimates from FluxSat and FLUXCOM-X-BASE), and 6 independents FLUXNET sites for GPP; 2) the boxplots of the RMSD between model simulations and the several datasets calculated for each pixel / site (Figure 8).

As mentioned previously, prior SIF and GPP simulations overestimate the corresponding observations. All three data assimilation experiments significantly improve for the two variables, as quantified by the RMSD (decreasing of the values) and R² (values increasing). One important result is that the assimilation of SIF data only leads to an improvement in GPP (RMSD reduction of 32% against *in situ* data, 65% against FLUXCOM-X-BASE and 50% against FluxSat) which is of the same magnitude as when *in situ* GPP data are assimilated (RMSD reduction of 31% against *in situ* data, 67% against FLUXCOM-X-BASE and 49% against FluxSat). Slightly greater improvements observed when SIF is assimilated compared to the GPP-only case can be explained by the higher number of optimized parameters. Although the assimilation of *in situ* GPP data positively impacts the simulated SIF, the optimized model still overestimates the TROPOSIF data. The co-assimilation of SIF and GPP results in the highest model-data agreement (RMSD reduction wrt GPP of 33% against *in situ* data, 66% against FLUXCOM-X-BASE and 53% against FluxSat), although it is only slightly higher than when SIF or GPP data are assimilated alone.

For this BoENF PFT, differences between model simulations and observations remain even after calibration. For SIF, the assimilation of SIF data primarily corrects the simulated magnitude during the growing season, while leading to an underestimation of SIF (with respect to TROPOSIF) during the winter and spring months. Data assimilations also mostly impacts the magnitude of the simulated GPP without correcting its seasonal pattern. ORCHIDEE simulates an earlier peak of GPP (early July) compared to *in situ* GPP data (mid-July) and data-driven estimates, even after the various calibrations. Different patterns in model-data agreement are seen whether data-driven data (at 0.1° resolution) or *in situ* data are considered: while ORCHIDEE still overestimates FLUXCOM-X-BASE and FluxSat over the whole seasonal cycle, an underestimation of GPP is observed at the site scale after the peak of the growing season (starting from early July) up to October.

Despite the remaining model-data discrepancies, the study has highlighted the strong constraint provided by space-borne SIF data within our modelling and data assimilation frameworks, which significantly improves the temporal dynamics of GPP for BoENF. These preliminary results highlight the potential of SIF observations for the CO2MVS system. The same assessment for the other PFTs of ORCHIDEE is ongoing.

5.1 SIF data assimilation into D&B

The assimilation of TROPOSIF in D&B improves the fit to the SIF observations (Figure 9). Figures 10 and 11 present evaluation against two independent data sets of GPP derived from eddy covariance measurements, and FAPAR derived by the JRC-TIP (Pinty et al., 2007) from MODIS broadband albedos (Pinty et al., 2011), respectively. Figure 10 shows a reduction of GPP RMSE by 7% from 2.242 10-5 gC/m^2/s for the prior to 2.095 10-5 gC/m^2/s for the posterior through improved amplitude and seasonality, earlier start and later decline in productivity. In Figure 11, results show a reduction of FAPAR RMSE by 45% from 0.479 for the prior to 0.262 for the posterior through improved amplitude, in particular in spring.

Owing to the long assimilation window and to the transfer of information through the calibrated process parameters, the assimilation also improves the fit against the independent observations in the year 2017, i.e. before the availability of the SIF product that was assimilated. RMSEs have been reduced for both independent observational data types (~7% for GPP and ~45% for FAPAR).

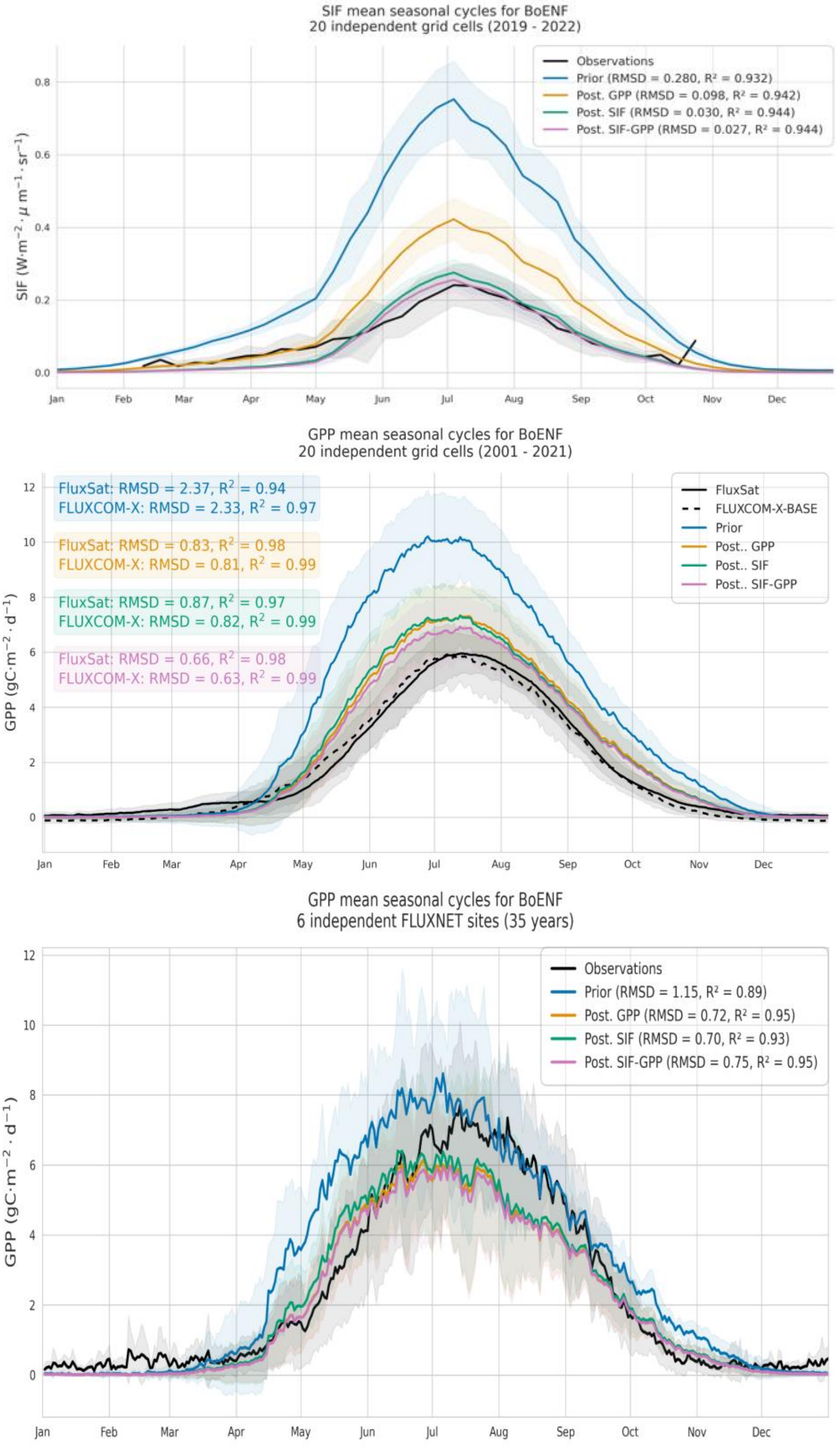


Figure 7: Compared mean seasonal cycles of SIF (top) and GPP (middle and bottom) between evaluation observation datasets and model simulations performed with the prior and optimized parameters (for the three data experiments). For SIF, the assessment is performed for 20 independent grid-cells; For GPP, the same grid cells

are considered (middle) together with in situ data (bottom). RMSD and determination coefficients are determined with respect to the mean seasonal cycles.

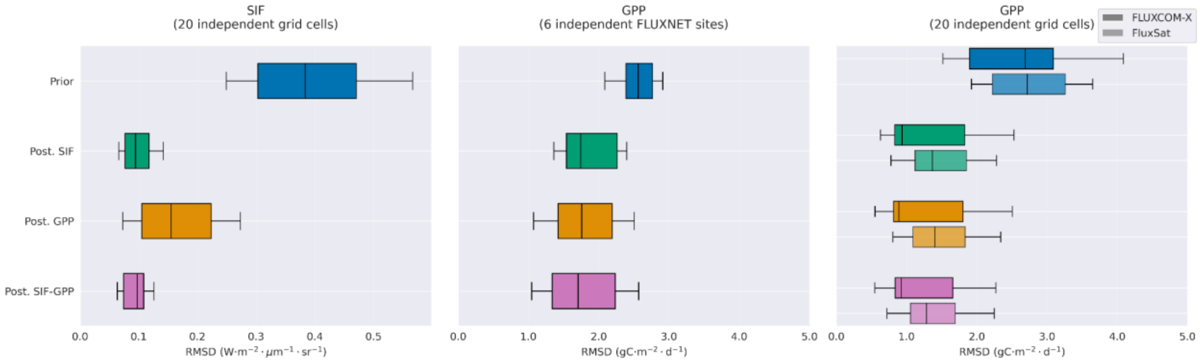


Figure 8: Boxplots of the RMSD between model simulations and the different datasets (TROPOSIF data as well as FLUXCOM-X-BASE and FluxSat GPP estimates for 20 independent grid-cells, as well as independent in situ data for GPP) for the different cases (prior model parameters and optimized values following the three data assimilation experiments).

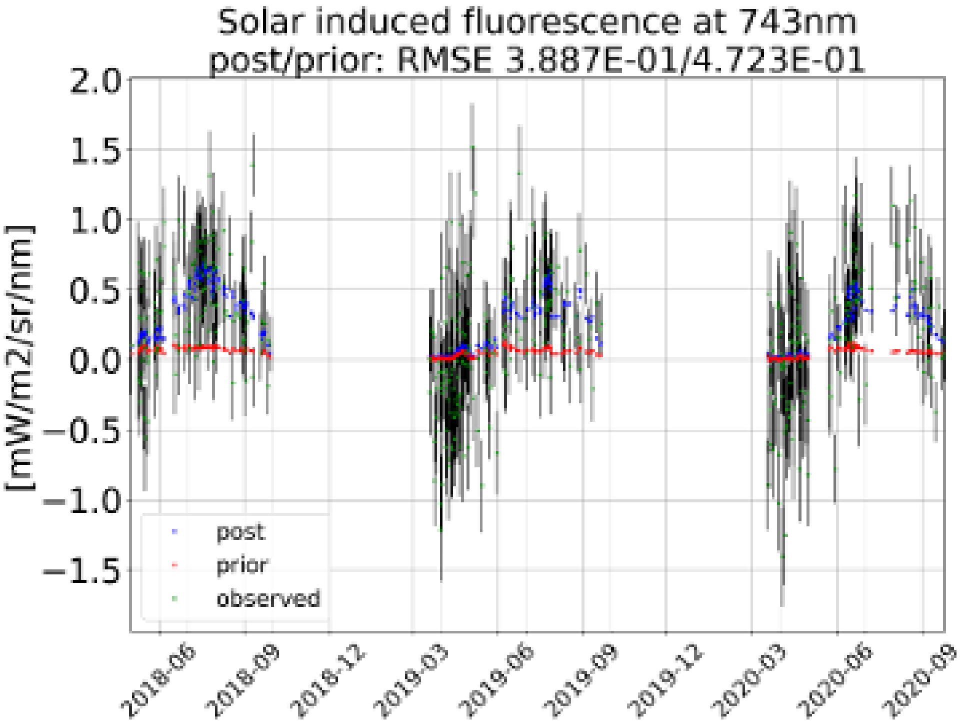


Figure 9: SIF simulated with D&B prior (red) and posterior after assimilation of SIF (blue), along with observational TROPOSIF product (green) and its uncertainty (black).

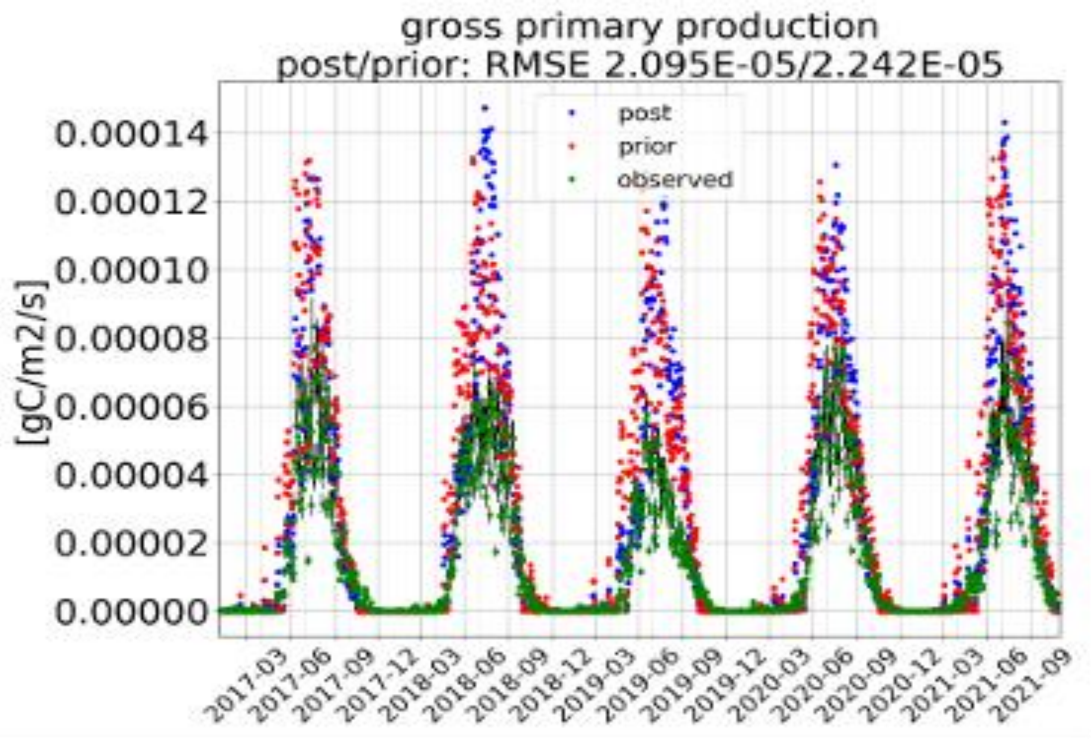


Figure 10: GPP simulated with D&B prior (red) and posterior after assimilation of SIF (blue) and GPP derived from eddy covariance measurements (green).

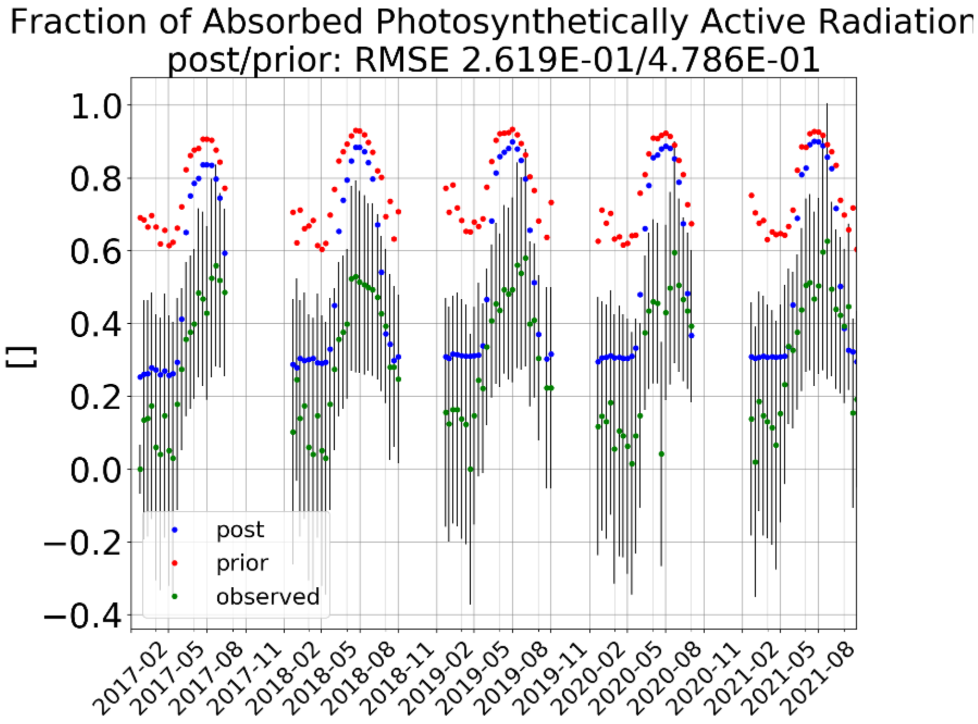


Figure 11: FAPAR simulated with D&B prior (red) and posterior after assimilation of SIF (blue) and retrieved by JRC-TIP (green).

5.2 SIF data assimilation in LDAS-Monde

5.2.1 Baseline and experiments

To evaluate the benefit from assimilating TROPOSIF, several experiments are run. One with no assimilation (i.e. Open-Loop), one with the assimilation of LAI300 (Fuster et al 2020) which will represent the “state-of-the-art” for vegetation monitoring, one that assimilate TROPOSIF and one that assimilate both TROPOSIF and LAI300. As a reminder, the NN used to assimilate SIF was trained on CLMS LAI-V1 and not LAI300 and on the same grid resolution. The assimilation experiments cover from January 2018 to December 2020.

The evaluation will be done considering two indicators, the improvement on the RMSE wrt LAI observed by PROBA-V and the improvement on the correlation.

5.2.2 Evaluation on LAI monitoring

The open-loop estimation of LAI cannot predict changes due to anthropogenic factors such as agricultural practices (i.e. harvesting season, heavy irrigation, …) like the ones in the Ebro basin around the centre of the domain. This can partially be corrected by assimilating satellite LAI products, but their temporal coverage is very low. In contrast, the TROPOSIF product, which is available daily, can constrain day-to-day LAI variations. In the TROPOSIF data assimilation experiment, LAI is directly updated as it is the only input of the observation operator, in addition to the structural parameters (DOY,lat, lon).

Figure 12 shows monthly mean LAI map from the open loop experiment (left), from the TROPISIF data assimilation experiment (middle) and their difference, from July 2018 (top) to December 2018 (bottom). It shows that, compared to the open loop experiment, the TROPOSIF data assimilation experiment increases LAI in the irrigated area around the Ebro basin, while it reduces it in other areas.

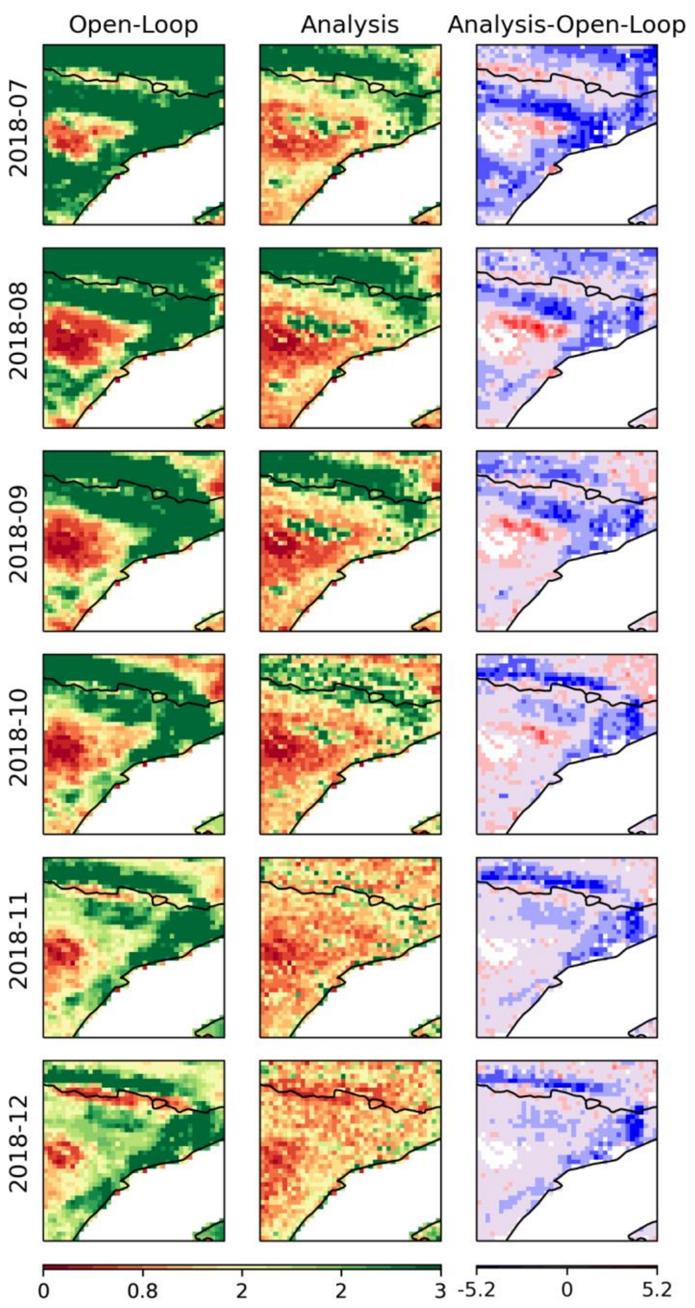


Figure 12: Monthly mean comparison of LAI for the open-loop experiment and the assimilation of TROPOSIF for the Ebro basin. From left to right, the averaged LAI for the Open-Loop, the averaged LAI for the Analysis and the difference between the two, each month from July (top) to December (bottom) 2018.

The increase of LAI in the Ebro basin is similar to what we can expect by assimilating a CLMS LAI product. Yet, as it is presented in Figures 13-14 below, the assimilation of LAI300 is decreasing, by locally up to 2 $m^2.m^{-2}$, LAI RMSE on the whole domain compared to the open-loop experiment, while the assimilation of SIF has overall less impact. In case of a co-assimilation of SIF with the LAI300, the addition of SIF decreases further the RMSE, especially on deciduous forest of the north of the Pyrenees. These results indicate that the assimilation of TROPOSIF only has a relatively neutral impact, but the co-assimilation of TROPOSIF and LAI have complementary impact to overall improve LAI over the study area.

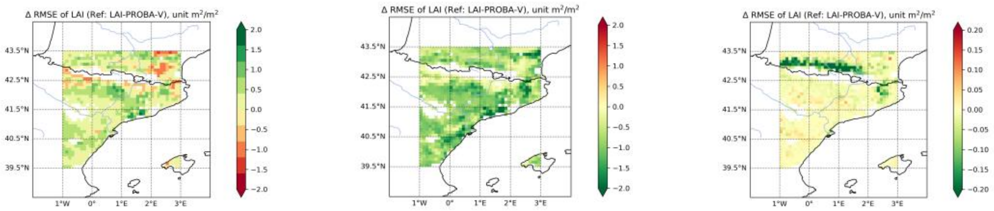

Δ RMSE of LAI (Ref: LAI-PROBA-V), unit m²/m²
43.5°N
42.5°N
41.5°N
40.5°N
39.5°N
1°W
0°
1°E
2°E
3°E
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Δ RMSE of LAI (Ref: LAI-PROBA-V), unit m²/m²
Δ RMSE of LAI (Ref: LAI-PROBA-V), unit m²/m²
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20

Figure 13: Impact of TROPOSIF data assimilation (left), combined TROPOSIF and LAI300 data assimilation (middle) and difference between TROPOSIF and combined data assimilation (right), expressed as LAI RMSE difference with the open loop experiment, against PROBA-V LAI.

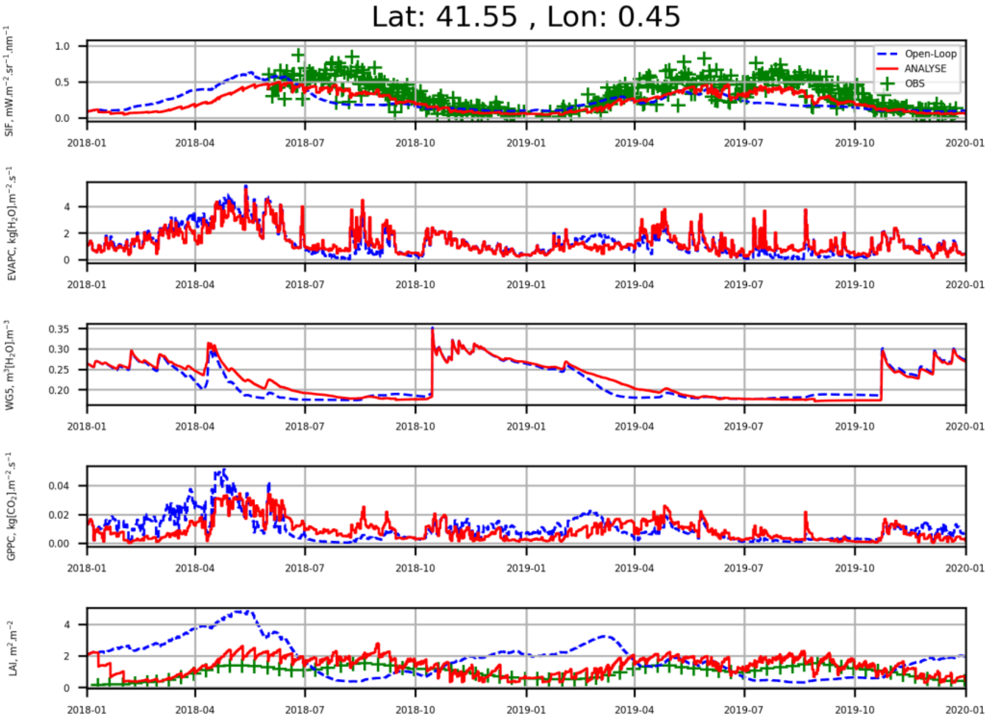


Figure 14: Times series for a given pixel in the irrigated area, showing from top to bottom, SIF, daily cumulated evapotranspiration, soil wetness at root level, daily cumulated GPP and LAI, for the open loop experiment (blue), the combined LAI300 and TROPSIF data assimilation experiment (red), and observations (green).

On this specific pixel, the period where the fields are irrigated along the late summer are clearly visible, leading to increases on the LAI and the SIF and drier soil. The behaviour of the analysis timeseries in “shark teeth” shows the 10 periods of the LAI300 synthesis assimilate while the small variation in between and partially correcting the deviation towards the open loop solution are done by TROPOSIF observations.

Overall, the assimilation of TROPOSIF in addition to a usual LAI product improves the correlation with the different observations even in areas where anthropization is high and providing significant changes in the fluxes like the GPP or the evapotranspiration.

The assimilation of TROPOSIF benefits slightly to the LAI monitoring, but less than assimilating a LAI product. The best configuration found is to co-assimilate TROPOSIF with a LAI 10 days synthesis.

5.2.3 Discussion on the uncertainties

SIF daily products are known to be noisy, with in the case of the TROPOSIF product, uncertainty values around 30%. In addition, neural network observation operators approximate TROPOSIF observation with an RMSE of about 0.1 mW.m^{-2}.sr^{-1}.nm^{-1} and a relative error of the order of 20% or above. These uncertainties need to be accounted for in the data assimilation system. The use of a flat SIF observation error of for example 0.1 mW.m^{-2}.sr^{-1}.nm^{-1} would be consistent with the product RMSE, but it does not account for observation operator uncertainties and it is expected to be not suitable for large values of SIF which are associated to larger uncertainties. For LAI data assimilation, we tested observation errors ranging from 5% to 40% (not shown) with best results obtained for LAI observation errors of 20% which is retained for this study. For SIF, we tested several configurations using either flat observation error of 20%, or gradual increase of the error relative to SIF values from 0.1 mW.m^{-2}.s^{-1}.nm^{-1} for SIF lower than 0.5 mW.m^{-2}.s^{-1}.nm-1 and of 20% for SIF values larger than this threshold. As it is illustrated in Figure 15, this configuration (purple line) provides overall more consistent improvements in LAI (except in summer 2018) than using flat SIF errors (cyan and red lines). These results demonstrate the feasibility and the relevance of SIF data assimilation. Further work will be conducted to refine the error specification in the data assimilation system following the approach of Desroziers et al., (2005).

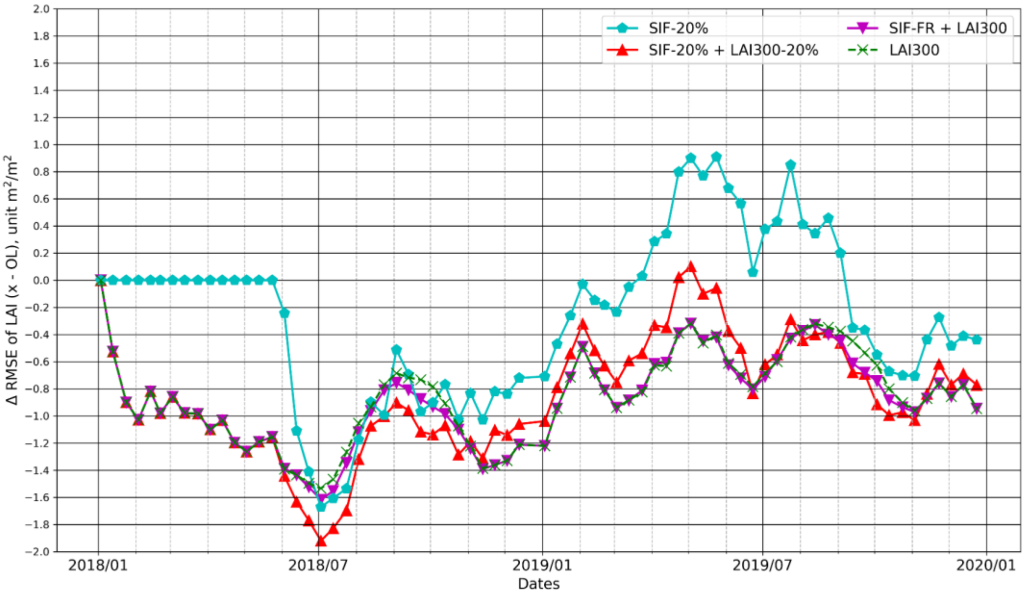


Figure 15: Differences in LAI RMSE over the whole domain between data assimilation experiments the open loop experiment, from January 20218 to December 2019. The cyan and red lines show results using constant observation error of 20% for SIF only and combined SIF and LAI data assimilation, respectively. The green and purple lines show results with relative observation error as explained in the text, for LAI only and combined SIF and LAI data assimilation, respectively.

5.3 SIF data assimilation in the IFS ECLand

5.3.1 LAI increments

All experiments produced low magnitude of LAI increments (Figures 16 and 17) within the range of -0.3 m2/m2 to 0.3 m2/m2. Similar level of magnitude was reported in the Météo France regional study. The LAI increments show consistent spatiotemporal patterns such as the greening of low vegetation in the Sahel region and in Europe in April (Figure 16). High vegetation increments are positive over the Amazon and part of the Central Africa rainforest (Figure 17). Boreal forest displays both positive and negative increments in July depending on the region. Spurious low vegetation LAI increments are obtained over central Australia in January where the SIF observation operator was associated with larger uncertainties (CORSO report 4.2).

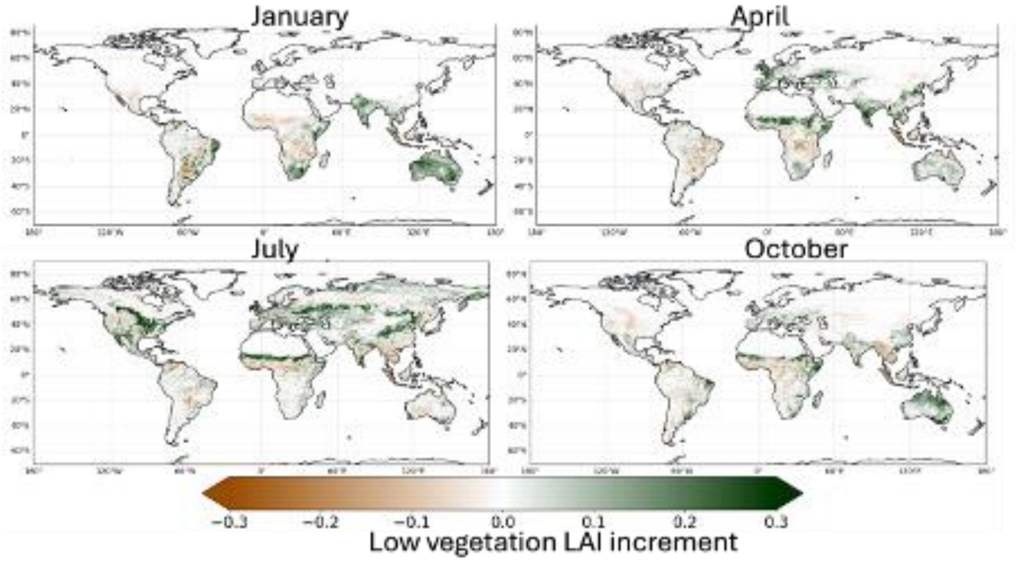


Figure 16: 2022 Monthly mean maps of low vegetation LAI increments produced from experiment LDA2.

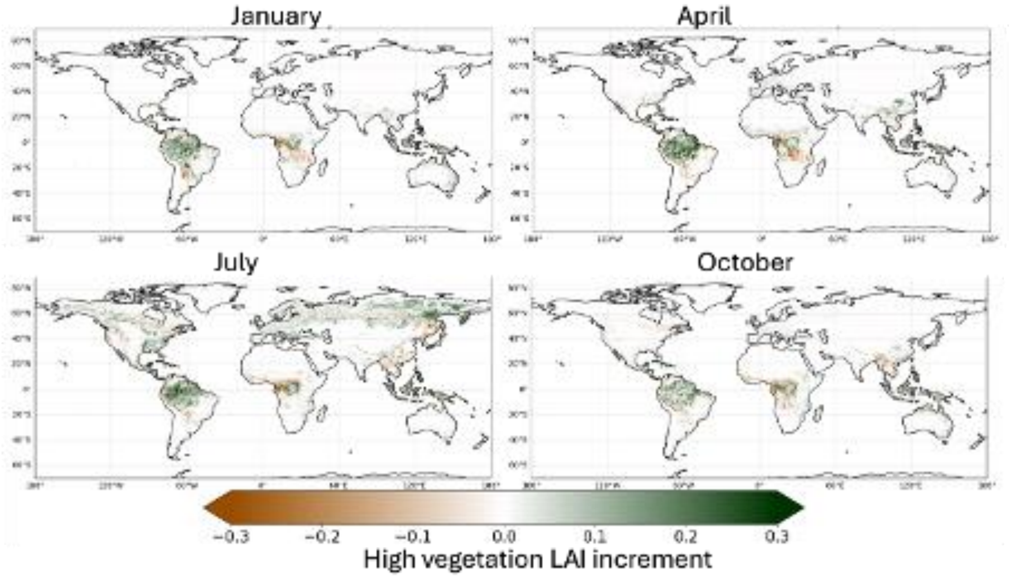


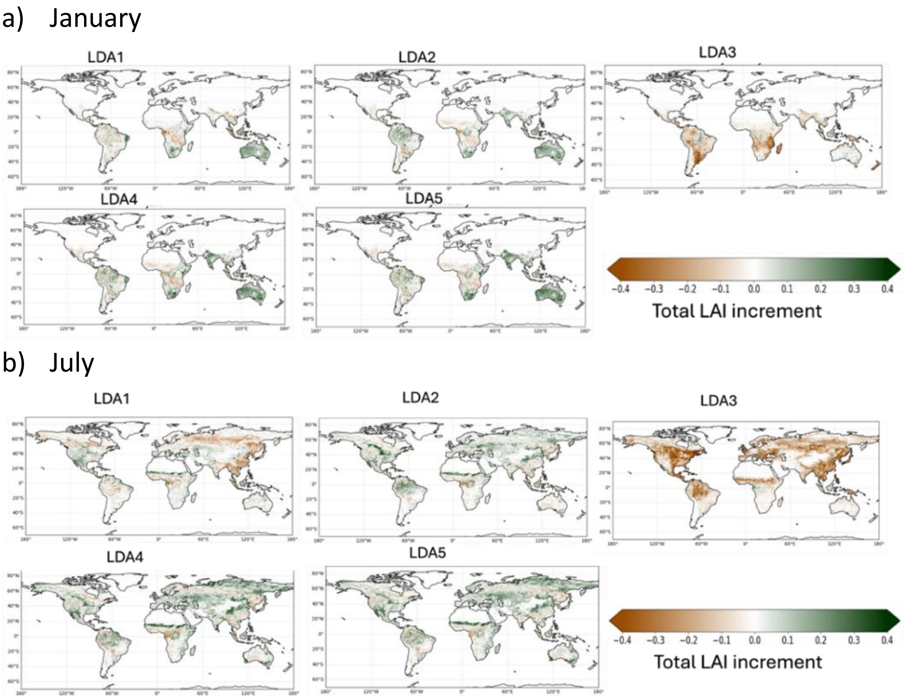
Figure 17: 2022 Monthly mean maps of high vegetation LAI increments produced from experiment LDA2.

5.3.2 Observation operators comparison and evaluation against satellite LAI

Figure 18 shows LAI increments obtained for January (a) and July (b) for all the ECMWF LDAS experiments presented in Table 5. It shows that the use of different sets of predictors in the ML-based observation operators leads to contrasted spatial patterns of LAI increments. While the LDA1 and LDA2 experiments conducted with the M1 and M2 ML models trained using the IFS model fields (see Table 4) provide consistent spatial patterns of both positive and negative increments, the LDA3 experiment based on the model M3 which uses latitude, longitude, time and the satellite LAI as predictors, leads to widespread negative LAI increments, consistent with results previously obtained in CoCO2 project using VOD data assimilation in ECLand (Calver et al., 2023). LDA4 and LDA5 show similar maps of increments as LDA2 because they are based on the same observation operator M2, however the magnitude of the increment is amplified by the higher background error in LDA4 and the lower observation error in LDA5. Although M1 and M2 rely on similar sets of physical predictors, the lack of low and high vegetation fractions in LDA2 produces larger positive increments over the Amazon, North America and Europe.

5.3.3 Evaluation against CGLS LAI

Figure 19 illustrates the impact of SIF DA expressed in terms of correlation differences (with DA minus without DA) between ECLand LAI and CGLS LAI for 2022. It shows that the impact of SIF data assimilation is generally limited to specific regions. This limited impact in most of the regions for all the experiments (green areas in Figure 19) is expected given that the LAI climatology used as background in the data assimilation experiment is highly correlated with the GCLS LAI (correlation frequently above 0.85) since it is derived from the same satellite dataset acquired over a different period. Largest impact for all the experiments is obtained in the Amazon, Western Australia, Southern Argentina and Western USA where the correlation is decreased. The Figure shows that LDA3 exhibits largest degradations in terms of correlation over tropical forests. It shows relatively large impact on correlation values over desertic areas in Australia in areas. These results need to be taken with care due to artifacts in the correlation metrics in the absence of temporal variability.


a) January

b) July

Figure 18: 2022 January (a) and July (b) mean of total LAI increment maps produced by each LDA experiment.

The impact of SIF DA on LAI RMSD is shown in Figure 20. It shows that LDA3 exhibits larger impact of SIF DA on LAI RMSD than the other experiments. It leads to larger increase in RMSD over tropical forests than the other experiments, particularly over the Amazon. LDA3 also indicates widespread improvements in non-tropical forest areas of South America and Africa, as well as part of the US and Europe. No impact is observed in terms of RMSD over Australia which is consistent with very low LAI increments over Australia for LDA3 (Figure 18). Compared to LDA3, the changes in RMSD are of less amplitude and for smaller areas for the experiments LDA1,2,4,5. Areas presenting largest degradations for these experiments mainly concern semi-arid and sparse vegetation regions (e.g. grassland in Central and Western Australia, Somalia, Sahel, Southern USA) as well as some structures in the Amazon that could be related to deforestation. LDA 2,4,5 display areas of significant improvement (reduction of RMSD) over Northern Eurasia and scattered regions in North and South America, central Europe, Eastern and Southern Australia. Besides, they show similar spatial patterns of difference in RMSD because they are based on the same ML-observation operator M2. The changes are however amplified in LDA4 and LDA5 by the use of a lower observation error and higher background error.

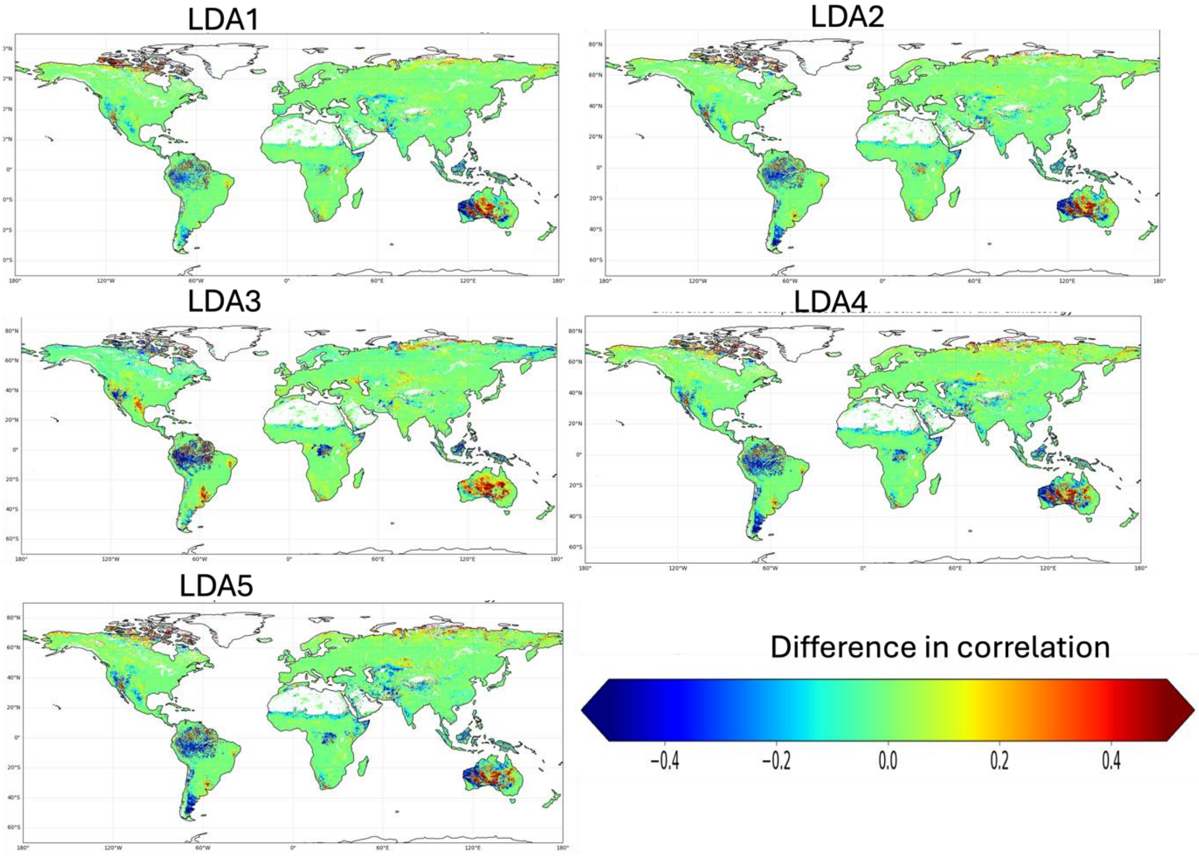


Figure 19: Impact of SIF DA (Table 5) shown as correlation differences with vs without DA, between ECLand LAI and CGLS LAI for 2022.

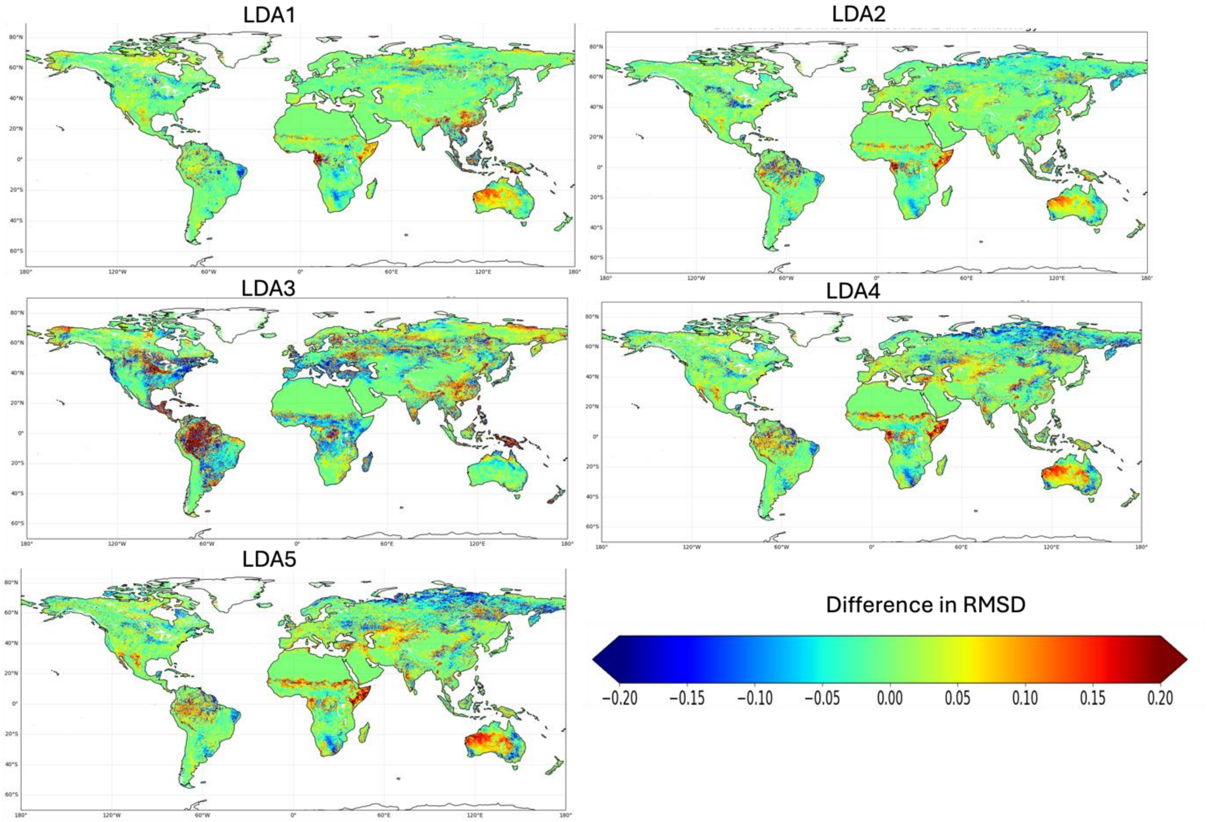


Figure 20: Impact of SIF DA (Table 5) shown as RMSD differences with vs without DA, between ECLand LAI and CGLS LAI for 2022.

5.3.4 Impact on 2m temperature and GPP forecasts

As illustrated by Figure 21, the updated LAI from the SIF DA has a neutral impact on the prediction of 2m temperature.

Figure 22 shows that the use of the updated LAI resulting from SIF data assimilation induces changes in the IFS in GPP over tropical rainforests in Africa and Amazon, North America and Europe. The magnitude of those changes is low and no clear patterns of increase or decrease in GPP are observed.

The evaluation on NWP and GPP forecasts presented here concerns a limited period for one experiment from Table 5. Further evaluations over both winter and summer periods will be conducted for all the experiments in the next report.

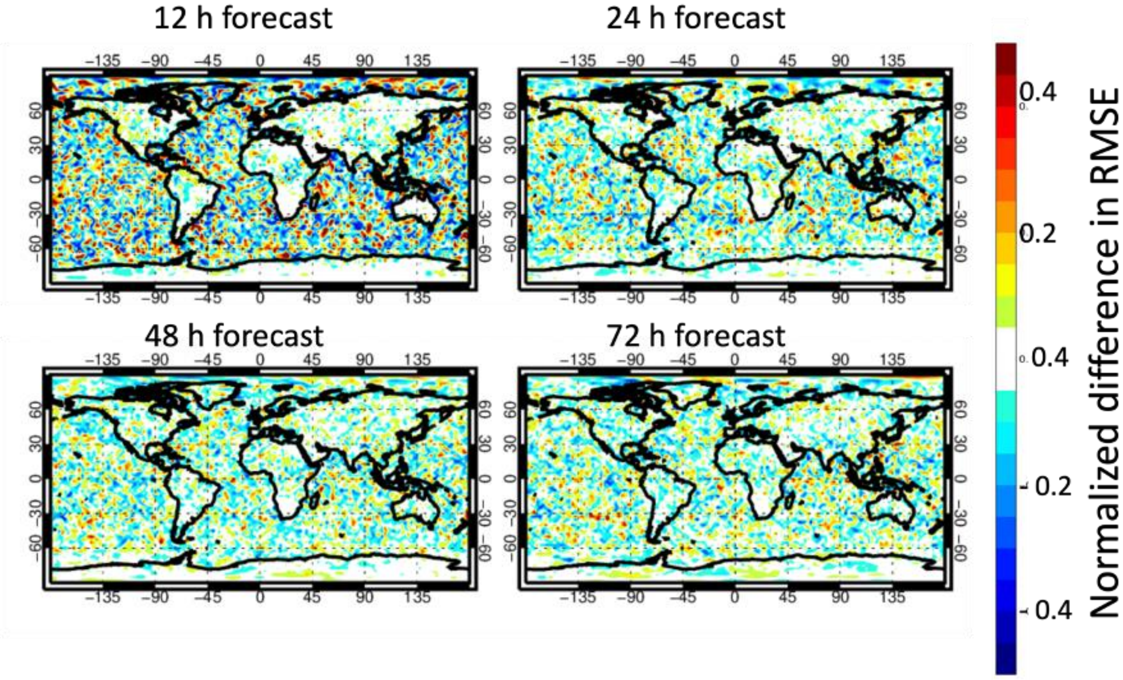


Figure 21: Difference between RMSE of 2m temperature forecast from the updated LAI (LDA2 experiment) and RMSE of 2m temperature forecast from the LAI climatology for June 2022.

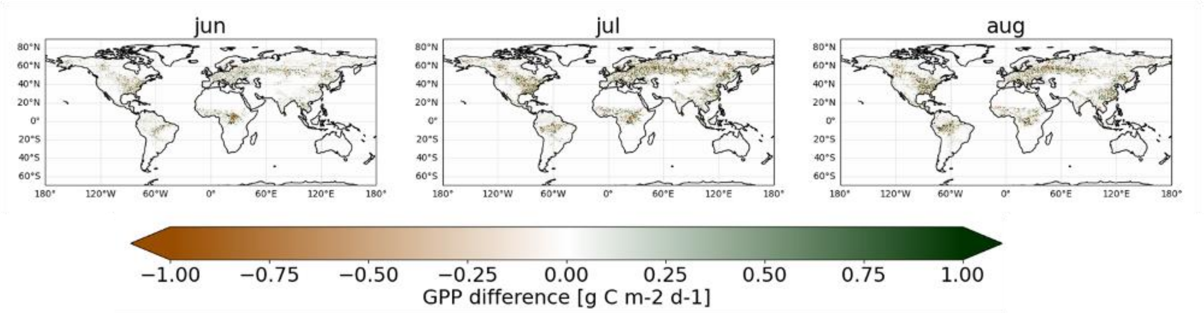


Figure 22: Difference in GPP forecast (24 h accumulated flux) between an IFS experiment with the updated LAI from LDA1 experiment and a control IFS experiment based on the LAI climatology for summer 2022.

6 Conclusion

The objective of this work was to investigate Solar Induced Fluorescence data assimilation to consistently analyse soil moisture and vegetation variables to constrain NWP and the CO2MVS carbon fluxes in the ECMWF IFS. This work presented here relies on observation operators (D4.2) and land data assimilation systems developments that were conducted in the IFS ECLand Land Data Assimilation System (ECMWF). Developments in the IFS have been supported by work conducted in other models of various degrees of complexity and using different modelling and data assimilation approaches: ORCHIDEE (CEA), and D&B (iLab/ULund), and ISBA (MF), allowing to explore different methodologies to exploit SIF observations.

Results from ORCHIDEE highlighted the strong constraint provided by space-borne SIF data within their modelling and data assimilation frameworks, which significantly improves the temporal dynamics of GPP for Boreal NeedleLeaf Evergreen Forest. The same assessment for the other PFTs of ORCHIDEE is ongoing. Results of SIF data assimilation in the D&B model improves the fit to the SIF observations and to two independent data sets of GPP and FAPAR. In comparison, results in NWP compatible models such as ISBA and ECLand-IFS showed a more limited impact of SIF data assimilation. In the ISBA model, results of experiments at the regional scale over the Ebro basin demonstrated that the assimilation of TROPOSIF, in addition to LAI, improves the correlation with the different observations. Results showed that it is the case even in areas where anthropisation is high and providing significant changes in the fluxes like the GPP or the evapotranspiration.

The work conducted in the IFS at the global scale demonstrates the capability to analyse LAI by assimilating SIF satellite observations in the ECMWF LDAS. The comparison of distinct ML-based observation operators shows the impact of the set of predictors on the assimilation results. While the use of latitude, longitude and satellite LAI as predictors leads to the most accurate SIF prediction, the resulting LAI increments show mixed results with improvements over Europe and degradation over tropical forests. These results will be further investigated in the last year of the project to understand the sources of these spurious increments and implement appropriate solutions to filter them out. The observation operators based on the IFS physical predictors provide more realistic spatiotemporal patterns of low and high vegetation increments. A better agreement with the satellite LAI is mainly observed over Northern Eurasia and scattered regions in North and South America, central Europe, Eastern and Southern Australia. Lower performances are obtained over tropical rainforest (Amazon) and semi-arid/sparse vegetation regions where the prediction of SIF by the ML observation operator is more uncertain (report CORSO D4.2. Besides, the magnitude of the produced increments is too low to have an impact on NWP and carbon flux forecast. A possible reason for this is the lack of sensitivity of the observation operator to the analysed variable (LAI). An important lesson-learned from this work is that the evaluation of the prediction performances of the observation operator is not sufficient. The objective is not to develop an emulator of SIF but an observation operator that can be used in the data assimilation system to predict the model-counterpart of the SIF satellite observation. Testing the observation operator in the DA system is paramount to verify that it provides enough sensitivity to the analysed variable (here LAI). The next steps will consist in (1) enhancing the sensitivity of the observation operator to LAI by testing other model architectures (e.g. feedforward neural network), (2) tuning the observation and background errors; (3) implementing the cross correlation between vegetation variables and soil moisture to produce coupled soil moisture and LAI increments; (4) evaluate the GPP forecast using independent reference products.

To summarise, this report shows rather good performance of SIF data assimilation in the two most advanced systems (ORCHIDEE and D&B) using physical observation operators. In contrast, relatively neutral results are obtained with the simpler models ISBA and ECLand which use NN observation operators. Compared to ISBA and ECLand, ORCHIDEE and D&B rely on complex physically based surface models with a comprehensive representation of processes related carbon cycle. They use data assimilation to update model parameters, with long (up to pluri-annual for D&B) DA windows to fully exploit the information from SIF observations. For practical reasons such models and data assimilation configurations are not applicable for global near-real time operational applications. ISBA and ECLand rely on state-of-the-art land surface models used for NWP, using relatively simple carbon cycle representation, daily or sub-daily assimilation windows, and observation operators using machine learning approaches in line with the level of complexity of these land surface models. Preliminary results presented here with ISBA and ECLand show limited impact, but they demonstrate for the first time the proof of concept of SIF data assimilation in this type of models. They show interesting features and promising results with both systems, with high complementarity with LAI assimilation demonstrated in ISBA, and LAI increments obtained with SIF for one configuration of the observation operator, consistent with those obtained in CoCO2 with VOD data assimilation. Further improvements planned in the observation operators and data assimilation settings as discussed above will consolidate the approach for global NRT application and potential usage in the CO2MVS.

7 References

Albergel, C., Calvet, J.-C., Mahfouf, J.-F., Rüdiger, C., Barbu, A., Lafont, S., Roujean, J.-L., Walker, J., Crapeau, M., and Wigneron, J.-P.: Monitoring of water and carbon fluxes using a land data assimilation system: a case study for southwestern France, Hydrology and Earth System Sciences, 14, 1109–1124, 2010.

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.

Balsamo, G., P. Viterbo, A. Beljaars, B. van den Hurk, M. Hirsch, A. Betts, K. Scipal A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the Itegrated Forecast System J. Hydrometeorol., 10 (2009), pp. 623-643, 2009.

Barbu, A., Calvet, J.-C., Mahfouf, J.-F., Albergel, C., and Lafont, S.: Assimilation of Soil Wetness Index and Leaf Area Index into the ISBA-A-gs land surface model: grassland case study, Biogeosciences, 8, 1971–1986, 2011.

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.

Bonan, B., Albergel, C., Zheng, Y., Barbu, A. L., Fairbairn, D., Munier, S., and Calvet, J.-C.: An ensemble square root filter for the joint assimilation of surface soil moisture and leaf area index within the Land Data Assimilation System LDAS-Monde: application over the Euro-Mediterranean region, Hydrology and Earth System Sciences, 24, 325–347, https://doi.org/10.5194/hess-24-325-2020, 2020.

Boussetta, S., Balsamo, G., Beljaars, A., Kral, T., & Jarlan, L. (2013). Impact of a satellite-derived leaf area index monthly climatology in a global numerical weather prediction model. International Journal of Remote Sensing, 34(9–10), 3520–3542. https://doi.org/10.1080/01431161.2012.716543

Calvet J.-C., B. Bonan, O. RojasMunoz, A. Agusti-Panareda, P. de Rosnay, P. Weston, P. Peylin, C. Bacour, V. Bastrikov, F. Maignan, T. Kaminski, W. Knorr, M. Vossbeck, M. Scholze, "Demonstrator systems for using remote sensing data (LAI, VOD, SIF) in online global prior fluxes for the CO2MVS prototype", CoCO2 H2020 project D3.4, June 2023, https://coco2-project.eu/sites/default/files/2023-11/CoCO2-D3-4-V2-1.pdf

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;. p. 785–94. Available from: http://doi.acm.org/10.1145/2939672.2939785, 2016

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

de Rosnay P., M. Drusch, D. Vasiljevic, G. Balsamo, C. Albergel and L. Isaksen: A simplified Extended Kalman Filter for the global operational soil moisture analysis at ECMWF, Q. J. R. Meteorol. Soc., 139:1199-1213, 2013. doi: 10.1002/qj.2023

de Rosnay, P., P. Browne, E. de Boisséson, D. Fairbairn, Y. Hirahara, K. Ochi, D. Schepers, P. Weston, H. Zuo, M. Alonso-Balmaseda, G.. Balsamo, M. Bonavita, N. Bormann, A. Brown, M. Chrust, M. Dahoui, G. De Chiara, S. English, A. Geer, S. Healy, H. Hersbach, P. Laloyaux, L. Magnusson, S. Massart, A.. McNally, F. Pappenberger, F. Rabier: "Coupled data assimilation at ECMWF: current status, challenges and future developments", QJRMS, 148(747), pp 2672-2702, 2022,
doi: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4330

Fairbairn, D., Barbu, A. L., Napoly, A., Albergel, C., Mahfouf, J.-F., and Calvet, J.-C.: The effect of satellite-derived surface soil moisture and leaf area index land data assimilation on streamflow simulations over France, Hydrology and Earth System Sciences, 21, 2015–2033, 2017.

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.

Fuster, Beatriz and Sánchez-Zapero, Jorge and Camacho, Fernando and García-Santos, Vicente and Verger, Aleixandre and Lacaze, Roselyne and Weiss, Marie and Baret, Frederic and Smets, Bruno: Quality Assessment of PROBA-V LAI, fAPAR and fCOVER Collection 300 m Products of Copernicus Global Land Service, Remote Sensing, 12(6) 10.3390/rs12061017, 2020

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.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horanyi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Holm, M. Janiskova, S. Keeley, P. Laloyaux, P. Lopez, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, J.-N. Thépaut: The ERA5 Global Reanalysis, QJRMS, , 146, 1999-2049,2020, https://doi.org/10.1002/qj.3803

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.

Mahfouf, J.-F. and Bilodeau, B.: A simple strategy for linearizing complex moist convective schemes, Quarterly Journal of the Royal Meteorological Society, 135, 953–962, https://doi.org/https://doi.org/10.1002/qj.427, 2009.

Nelson, J. A., Walther, S., Gans, F., Kraft, B., Weber, U., Novick, K., ... & Wang, Y. (2024). X-BASE: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, FLUXCOM-X. *EGUsphere*.

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.

Rodríguez-Fernández, N.; de Rosnay, P.; Albergel, C.; Richaume, P.; Aires, F.; Prigent, C.; Kerr, Y. SMOS Neural Network Soil Moisture Data Assimilation in a Land Surface Model and Atmospheric Impact. *Remote Sens.* **2019**, *11*, 1334. https://doi.org/10.3390/rs11111334

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.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33, 161–174. https://doi.org/10.1080/00401706.1991.10484804

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, 2019.

Knorr, W.: Annual And Interannual CO2 Exchanges Of The Terrestrial Biosphere: Process-Based Simulations And Uncertainties, Glob. Ecol. Biogeogr., 9, 225–252, 2000.

Knorr, W., M. Williams, T. Thum, T. Kaminski, M. Voßbeck, M. Scholze, T. Quaife, L. Smallmann, S. Steele-Dunne, M. Vreugdenhil, T. Green, S. Zaehle, M. Aurela, A. Bouvet, E. Bueechi, W. Dorigo, T. El-Madany, M. Migliavacca, M. Honkanen, Y. Kerr, A. Kontu, J. Lemmetyinen, H. Lindqvist, A. Mialon, T. Miinalainen, G. Pique, A. Ojasalo, S. Quegan, P. Rayner, P. Reyes-Muñoz, N. Rodríguez-Fernández, M. Schwank, J. Verrelst, S. Zhu, D. Schüttemeyer, and M. Drusch. A comprehensive land surface vegetation model for multi-stream data assimilation, D&B v1.0. *EGUsphere*, 2024:1–40, 2024. (doi:10.5194/egusphere-2024-1534)

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, Journal of Geophysical Research: Biogeosciences, 124, 1491–1507, https://doi.org/https://doi.org/10.1029/2019JG005029, 2019.

Pinty, B., Lavergne, T., Voßbeck, M., Kaminski, T., Aussedat, O., Giering, R., Gobron, N., Taberner, M., Verstraete, M. M., and Widlowski, J.-L.: Retrieving Surface Parameters for Climate Models from MODIS-MISR Albedo Products, J. Geophys. Res.-Atmos., 112, D10116, doi:10.1029/2006JD008105, 2007.

Pinty, B., Clerici, M., Andredakis, I., Kaminski, T., Taberner, M., Verstraete, M. M., Gobron, N., Plummer, S., and Widlowski, J.-L.: Exploiting the MODIS albedos with the Two-stream Inversion Package (JRC-TIP): 2. Fractions of transmitted and absorbed fluxes in the vegetation and soil layers, J. Geophys. Res.-Atmos., 116, D09106, doi:10.1029/2010JD015373, 2011.

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.

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, 2005.

CORSO D4.2 “Final review and improvement of land surface forward operators for SIF and low frequency MW data”. https://www.corso-project.eu/deliverables

Document History

Version	Author(s)	Date	Changes
1.0	Patricia de Rosnay	23.10.2024	Initial version
1.1	Patricia de Rosnay, Cédric Bacour, Bertrand Bonan, Jean-Christophe Calvet, 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.	22.11.2024	Input from WP4 partners
1.2	Patricia de Rosnay	03.12.2024	Consolidated draft version 1.2
1.3	Patricia de Rosnay and Sébastien Garrigues	10.12.2024	Consolidated version including M3 operator results
1.4	Patricia de Rosnay	12.12.2024	Revised after internal reviews
1.5	As above	17.12.2024	Revised after internal reviews

Internal Review History

Internal Reviewers	Date	Comments
Paul Palmer (UEDIN), Richard Engelen (ECMWF)	December 2024	